


Next-generation phenotyping of inherited retinal diseases from multimodal imaging with Eye2Gene

Received: 9 September 2024

Accepted: 11 April 2025

Published online: 18 June 2025

 Check for updates

A list of authors and their affiliations appears at the end of the paper

Rare eye diseases such as inherited retinal diseases (IRDs) are challenging to diagnose genetically. IRDs are typically monogenic disorders and represent a leading cause of blindness in children and working-age adults worldwide. A growing number are now being targeted in clinical trials, with approved treatments increasingly available. However, access requires a genetic diagnosis to be established sufficiently early. Critically, the timely identification of a genetic cause remains challenging. We demonstrate that a deep learning algorithm, Eye2Gene, trained on a large multimodal imaging dataset of individuals with IRDs ($n = 2,451$) and externally validated on data provided by five different clinical centres, provides better-than-expert-level top-five accuracy of 83.9% for supporting genetic diagnosis for the 63 most common genetic causes. We demonstrate that Eye2Gene's next-generation phenotyping can increase diagnostic yield by improving screening for IRDs, phenotype-driven variant prioritization and automatic similarity matching in phenotypic space to identify new genes. Eye2Gene is accessible online (app.eye2gene.com) for research purposes.

Inherited retinal diseases (IRDs) are a group of rare monogenic conditions affecting 1 in 3,000 people, with more than 270 different IRD-associated genes identified so far^{1–3,4}. IRDs cause degeneration of the retina, the light-sensitive tissue at the back of the eye responsible for vision. Some individuals with IRDs may be profoundly visually impaired from birth, while others experience progressive peripheral and/or central vision deterioration over time. Cumulatively, IRDs are a leading cause of blindness in children and the working-age population, with a substantial psychological and socioeconomic impact⁵.

Revealing the genetic cause of an IRD is a prerequisite to optimally determining prognosis, providing genetic counselling and inclusion in gene-directed clinical trials. However, this genetic diagnosis remains elusive in more than 40% of cases on average according to studies conducted in the UK^{6–8}, and the rate of diagnosis is likely to be much lower in parts of the world where genetic testing is less widely available^{9,10}. This is mostly due to limited access, lack of resources and infrastructure to support genetic testing services, inefficiencies in their provision and a shortage of specialists who can interpret the findings from molecular tests¹¹.

IRDs often have distinct phenotypic features that clinicians learn to recognize aided by modern high-resolution retinal imaging technology that rapidly and non-invasively acquires images of the retina with minimal inconvenience to the patient. These scans can be performed by a variety of imaging modalities such as fundus autofluorescence (FAF), infrared (IR) reflectance imaging and spectral-domain optical coherence tomography (SD-OCT), each of which convey different information about retinal health and architecture (Supplementary Fig. 1). FAF images can yield data relating to outer retinal and retinal pigment epithelium (RPE) health. Hyperautofluorescence in FAF images can result from either accumulation of fluorescent material (including lipofuscin) or from loss of photoreceptor outer segments or macular luteal pigment (which usually absorb the incoming short wavelengths), while loss of autofluorescence can be associated with loss of the RPE. These patterns are associated with particular IRDs¹². IR images are usually acquired together with SD-OCT scans. Brightness in IR images can be associated with levels of melanin and some early lesions associated with certain IRDs, such as pattern dystrophies, can be more apparent on IR than on FAF¹³. SD-OCT gives a high-resolution

✉ e-mail: n.pontikos@ucl.ac.uk

cross-sectional image of the retinal layers (including photoreceptor outer segments, external limiting membrane, outer and inner plexiform and nuclear layers, ganglion cell and nerve fibre layers) and the RPE. Photoreceptors are the primary cell type affected in many IRDs, and the reflectivity and width of the hyperreflective ellipsoid zone (the mitochondria-rich portion of the inner segment of the photoreceptors) is used to assess photoreceptor integrity and as a marker for disease progression¹⁴.

This high-resolution in-depth multimodal imaging information enables ophthalmologists to identify gene-specific patterns of disease, allowing prediction of the disease-associated gene in some cases. However, given the sparsity of these diseases individually, the experience required to make accurate clinical diagnoses is not widely available and limited to a handful of specialists and clinics who have developed this expertise over several decades. Wider access to this expertise could be deployed via an AI system trained to detect gene-specific patterns from multimodal retinal imaging scans.

Due to transformational improvements in imaging technology and a comprehensive genetic testing framework for IRDs embedded in specialist healthcare services over the past decade in the UK¹⁵, there are now a sufficient number of genetically characterized patients with detailed retinal phenotyping to build representative datasets for deep learning. Moorfields Eye Hospital (MEH) in the UK currently provides one of the most extensive datasets in the world¹.

We have leveraged this resource to develop a deep learning model, Eye2Gene, able to predict the causative IRD gene from the retinal scans of a patient with an IRD, acquired using the three aforementioned imaging modalities of FAF, IR and SD-OCT. Eye2Gene was trained on retinal scans acquired in individuals with IRDs seen at MEH who had undergone genetic testing and where a confirmed genetic cause had been identified by an accredited diagnostic laboratory. Eye2Gene was internally validated on a held-out set of retinal scans from MEH, and externally validated on retinal scans from individuals with IRDs from five different external clinical centres. Eye2Gene performance was also compared to that of expert clinicians.

Results

Eye2Gene is an ensemble of a total of 15 constituent CoAtNet deep convolutional neural networks, which takes one or more retinal scans of three different imaging modalities from a given patient and outputs a gene-level prediction score for 63 distinct IRD genes¹⁶. Collectively, these 63 genes cover over 90% of genetically characterized IRD cases in the European population^{14,17–19}. Given approximately 60–70% of IRD cases are molecularly diagnosed followed genetic testing, this may represent 54–63% of the total IRD population that includes both diagnosed and undiagnosed patients¹⁰. The Eye2Gene model was trained on a total of 58,030 scans from 2,451 patients (4,801 eyes, 9,291 appointments) from MEH, split into three different modalities: FAF ($n = 16,708$), IR ($n = 20,659$) and SD-OCT volumes ($n = 20,663$). For each of the three modalities, five distinct CoAtNet deep convolutional neural networks were trained, resulting in a total of 15 neural networks with identical architecture but different network weights (Supplementary Fig. 2). For each modality, these five networks were then combined into three modality-specific models by ensembling. The combination of these three models constitutes Eye2Gene. Given a single input scan of one of the three supported modalities, Eye2Gene applies the ensemble model corresponding to the modality of the scan to obtain a single scan-level gene prediction. Given multiple scans from a single patient over one or more appointments, Eye2Gene is applied to each scan in turn and the resulting predictions are combined to produce a single prediction for the patient, by taking the average over individual (postsoftmax) scan-level predictions per modality and then averaging across modalities (Fig. 1 and Supplementary Fig. 2).

Eye2Gene generalizes across IRD clinics

To evaluate Eye2Gene, we simulated the scenario of applying Eye2Gene to retinal scans acquired over one or more appointments per patient in our internal MEH test set of 28,174 retinal scans from 524 patients from a held-out internal test dataset, as well as on a further external test dataset of 39,596 retinal scans from 836 patients from five different external IRD clinics, which included the Oxford Eye Hospital (UK), Liverpool University Hospital (UK), University Hospital Bonn (Germany), Tokyo Medical Center (Japan) and the Federal University of São Paulo (Brazil). For each patient we ran Eye2Gene on all scans to get an overall prediction per patient and compared the prediction of Eye2Gene to the underlying gene diagnosis.

Eye2Gene attained an overall top-five accuracy (the proportion of cases the correct gene appeared in the top-five ranked choices of the model) of 83.9% (81.7–86.0%) across patients in all test datasets. A breakdown of results per centre is shown in Table 1.

Both ensembling across multiple networks and ensembling across images and/or modalities is crucial to Eye2Gene performance. Mean per-network top-five accuracy percentages were 68.9, 70.8 and 74.9% for FAF, IR and OCT, respectively, which improved to 71.0, 72.7 and 77.2% after ensembling across networks. In both the ensembled (83.9%) and unsembled (81.5%) case, combining predictions across multiple images and modalities led to higher accuracy than the best performing single modality.

Along with top-five accuracy, a commonly used metric in the literature for large multiclass problems^{20–23}, we also applied more flexible conformal prediction sets that dynamically selects a number of predictions according to a desired accuracy threshold²⁴. Selecting the 90% threshold we found a mean prediction set size of 8.1 genes with an empirical coverage (accuracy) of 90.5% (Extended Data Table 1). We also tested whether Eye2Gene's accuracy was biased based on ethnicity or demographic parameters such as age and sex. A slightly lower performance was found in the Asian ethnic group, however, none of the differences were statistically significant in our test set (Supplementary Fig. 3). No statistically significant differences in accuracy on the basis of age and sex were found either.

Eye2Gene predictions outperforms human expert-level accuracy

To contextualize Eye2Gene's performance at image interpretation relative to clinical specialists, we asked eight ophthalmologists specializing in IRDs, with 5–15 years of experience, to predict the causative gene based on a single FAF image per patient across 50 different patients from the internal held-out test set. In this task, ophthalmologists were asked to identify the correct diagnostic gene from 36 provided (instead of Eye2Gene's 63) when shown an FAF scan of a patient with an IRD. Historically, FAF has been a widely used imaging modality for the characterization of IRDs and hence a modality that many specialist ophthalmologists are likely to be the most familiar with using when diagnosing IRDs^{25,26}. On this task the ophthalmologists achieved an average top-five accuracy of 29.5%, compared to 76% for the Eye2Gene module trained on FAF images only when applied to the same images (restricted to single-image predictions and only using the five-network FAF ensemble for a fair comparison). The results of the human benchmarking by ophthalmologists are presented in Table 2. As expected, human performance tended to improve with the level of experience but the performance of Eye2Gene was considerably better than any single human expert.

Eye2Gene for phenotype-driven genetic variant prioritization

Genetic variant prioritization is an important task to diagnose single-gene conditions such as IRDs especially given the large number of genetic variants reported by whole genome sequencing^{27,28}. We evaluated the use of Eye2Gene in aiding genetic variant prioritization. Clinical notes and retinal scans of 130 individuals with IRD with

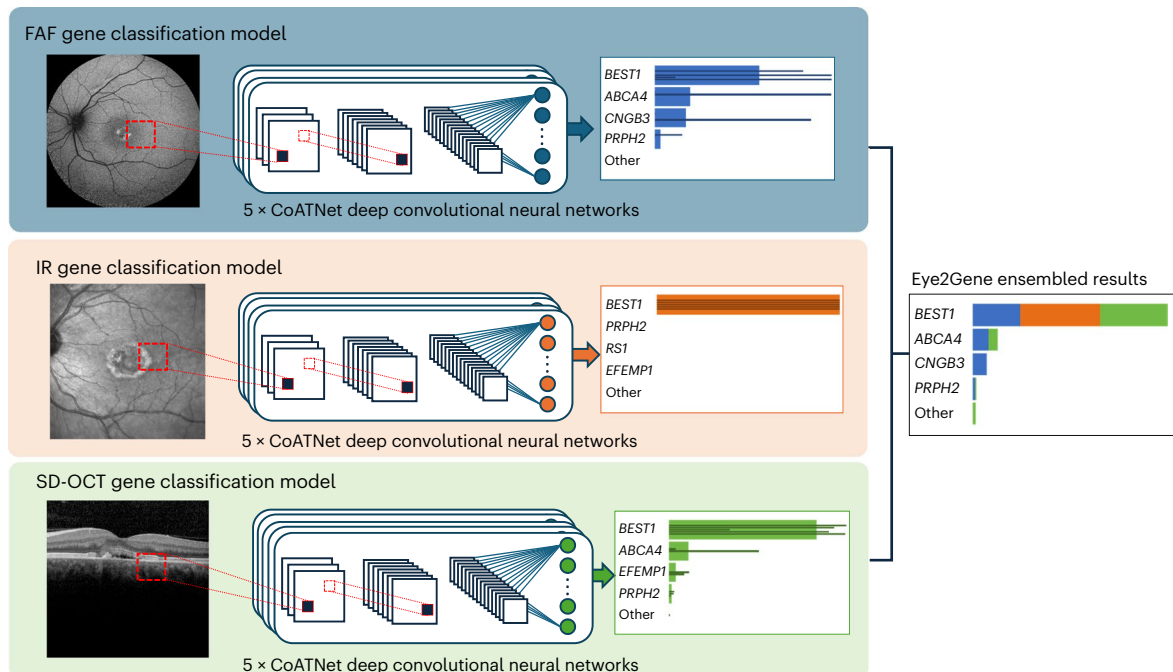


Fig. 1 | Eye2Gene model. Eye2Gene provides IRD-gene prediction given a retinal scan of one of three imaging modalities (FAF, IR or SD-OCT) for up to 63 gene classes. Images are initially resized to 256 by 256 pixels and rescaled to the range [0,1]. Each image modality-specific predictor block consists of an ensemble

of five CoATNet neural networks. The outputs are averaged to produce a final prediction output. The performance of Eye2Gene is evaluated on a held-out internal test dataset from MEH consisting of 28,174 images acquired from 524 patients over 9,291 patient visits since 2006.

confirmed gene diagnosis, who were part of the Eye2Gene test set, were reviewed and 21 distinct human phenotype ontology (HPO) terms were identified (4.7 on average per patient). Each patient's HPO terms were used as input for the Exomiser-hiPHIVE phenotype-gene scoring algorithm²⁹. Each patient's retinal scans were used as input for the Eye2Gene phenotype-gene score. We found that Eye2Gene provided a rank for the correct gene higher or equal to the Exomiser-hiPHIVE score in more than 75% of the patients (Wilcoxon rank sum $P < 5.0 \times 10^{-10}$), showing that image-based gene predictions can outperform HPO-only predictions for IRDs even when those include non-retinal specific HPO terms such as 'Sensorineural hearing impairment' and 'Mild hearing impairment' (Fig. 2).

Eye2Gene identifies new gene–phenotype groupings

We evaluated the use of Eye2Gene as a next-generation phenotyping tool to identify new gene–phenotype groupings. We extracted the activations from the penultimate layer of the FAF component of Eye2Gene to obtain high-dimensional latent embeddings of each FAF scan. The uniform manifold approximation and projection (UMAP)³⁰ dimensionality reduction algorithm was applied to the extracted activations to obtain two-dimensional representations for visualization purposes (Fig. 3a). Patients with the same genetic diagnosis frequently cluster in embedding space, even in the case of genes never encountered by Eye2Gene such as *ARHGEF18* and *CDH3* (Fig. 3b). As well as where patients cluster together, outliers can also be informative. For example, the outlying *MFRP* patient in Fig. 3b has hyper-autofluorescent optic disc drusen not present in the other individuals with *MFRP*. These observations could be useful for highlighting specific phenotypic subgroups associated with certain genes, linked to specific inheritance patterns (for example, *BEST1* associated with autosomal dominant Best disease and autosomal recessive bestrophinopathy³¹). Additionally, this could assist in identifying potential misdiagnoses or individuals with dual pathology. These embeddings also matched known phenotypic associations, for example genes associated with Retinitis Pigmentosa (for example *USH2A*, *RPGR*, *EYS*) cluster top right,

while cone-rod associated *ABCA4* appears overwhelmingly on the left, with phenotypically similar *PRPH2* and *PROM1* occupying the space in between.

We also found that individuals within the same family group tended to cluster closer together (Supplementary Fig. 4), with a mean inter-patient Euclidean distance of 41.3 in embedding space for individuals within the same family group, compared to 58.9 for a non-parametric bootstrap that resamples random pairs of patients ($P < 10^{-10}$).

To see what genes tended to appear close to each other we applied hierarchical clustering to the raw embeddings (pre-UMAP) to produce a hierarchy of phenotypic similarities according to Eye2Gene embedding space. The resulting dendrogram (Supplementary Fig. 5) captures some of the known phenotypic similarities in IRDs such as Stargardt phenotype genes (*ABCA4*, *PRPH2* and *PROM1*), retinitis pigmentosa genes (*RPGR* and *USH2A*) and achromatopsia genes (*CNGA3* and *CNGB3*). This shows that Eye2Gene's data-driven approach has the potential to identify phenotypically similar genes, even on new gene classes not included in the training data (Supplementary Fig. 6).

Eye2Gene distinguishes genetic from non-genetic disease

As not every patient that presents to an IRD clinic necessarily has a genetic condition, an additional Eye2Gene module was developed as a screening tool to detect presence or absence of IRDs, to identify individuals with other non-IRD conditions. Given there are many retinal conditions that cause atrophy, it is often challenging to distinguish IRD from other non-monogenic disease aetiologies, such as age-related macular degeneration (which can mimic macular dystrophies and vice versa), inflammatory changes such as autoimmune retinopathies, posterior uveitis and acute zonal outer occult retinopathy, as well as certain drug toxicities (including retinopathies associated with hydroxychloroquine, pentosan polysulfate and antiretroviral medications). Eye2Gene can serve as a screening tool based on conditions with a similar presentation (differential diagnosis), providing an area under the receiver operating curve (AUROC) of 0.98 (Fig. 4).

Table 1 | Overview of Eye2Gene results on test data across different IRD clinics plus demographic characteristics

IRD clinics	Number of patients (number of images)	Number of unique genes	Median age	Percentage female	Ethnicity distribution				Anticipated top-five accuracy	Top-five accuracy
					Percentage white	Percentage Asian	Percentage Black	Percentage admixed		
Oxford	390 (29,145)	33	44.5	38	90	10			89.6%	90.1%
Liverpool	156 (6,174)	27	30	55	89	10		1	87.9%	88.2%
Bonn	129 (2,838)	12	43.7	48	85			15	90.4%	87.6%
Tokyo	60 (1,493)	24	29	54	2	97		1	71.0%	70.4%
São Paulo	40 (1,494)	10	35	55	45			55	89.1%	93.9%
All external	775 (39,854)	42	39.7	47	66	30		4	87.9%	87.9%
Moorfields	524 (28,174)	63	39	44	50	18	3	29	–	77.8%
All test data	1,299 (68,028)	63	39.4	46	59	25	1	14	–	83.9%

Top-five accuracy represents the proportion of case for which the correct gene appeared in the top-five ranked choices of the model. Anticipated top-five accuracy refers to extrapolated accuracy based on per-gene accuracy at MEH extrapolated to target dataset gene distribution. Note that unspecified ethnicity is not accounted in the ethnicity distribution reported here. Bold indicates summary data across sites.

Table 2 | Benchmarking Eye2Gene against humans

Ophthalmologist	Years of experience specializing in IRDs	Correct top-five guesses
1	5	13 (26%)
2	5	14 (28%)
3	5	14 (28%)
4	6	15 (30%)
5	7	13 (26%)
6	10	15 (30%)
7	15	16 (32%)
8	15	18 (36%)
Ophthalmologists' average		14.75 (29.5%)
Eye2Gene		38 (76%)

For comparison, eight ophthalmologists and Eye2Gene classified 50 FAF retinal scans of 50 patients results from the MEH internal held-out test dataset. Bold indicates summary data across sites.

Eye2Gene predictions outperform other AI approaches

As well as comparing favourably against gene prioritization approaches based on HPO terms^{21,32}, Eye2Gene outperforms previously published image-based AI approaches trained on smaller less diverse imaging datasets and for more limited classification tasks that distinguish up to four genes (Supplementary Table 1). For example, Miere et al.³³ achieved an accuracy of 88% in distinguishing *ABCA4* from *PRPH2* based on FAF scans, whereas on the same task Eye2Gene performs 90.8% (Supplementary Table 1). Miere et al.³⁴ also obtained an accuracy of 94.6% for three broad IRD phenotypes (retinitis pigmentosa, Best disease and Stargardt disease) from FAF images, compared to 95.6% for Eye2Gene on a similar task. Similarly Fujinami-Yokokawa et al achieved an accuracy of 89.3% on OCT³⁵ and 94.6% on FAF³⁶ in distinguishing *ABCA4*, *RP11* and *EYS*, whereas Eye2Gene, on the same task, achieves an accuracy of 96.3% on OCT and 96.3% on FAF (Supplementary Table 1). Shah et al.³⁷ trained a classifier on SD-OCT to distinguish only one type of IRD (Stargardt) from normals and obtained an AUROC of 0.99 whereas Eye2Gene is able to distinguish any IRD out of 189 from non-genetic conditions with a similar phenotype with an AUROC of 0.98 (Supplementary Table 1). Although Eye2Gene, by definition, is not able nor designed to outperform genetic testing, as it trained on gene labels derived from genetic testing, we have demonstrated its utility in guiding the prescription and interpretation of genetic testing by AI-powered phenotyping of retinal scans.

Eye2Gene is accessible as an online web application

To demonstrate how Eye2Gene could be used to assist in diagnosing individuals with IRD, Eye2Gene is accessible as a web application online at <https://app.eyegene.com> to be used as a research tool. Users upload a series of scans of the supported modalities (SD-OCT, 55-degree FAF and 30-degree IR) from a single patient. These images are then passed to the Eye2Gene model, which outputs a set of prediction scores for each of the 63 genes on each of the input scans. This information is aggregated into an overall Eye2Gene prediction score for that case and presented to the user (Extended Data Fig. 1). This score can be embedded in genetic variant prioritization frameworks.

Discussion

We present Eye2Gene, a deep learning model for classifying 63 causative genes in individuals with IRDs, using retinal scans acquired using three different imaging modalities. These scans can be obtained via non-invasive eye scans using widely available technology. We have comprehensively evaluated Eye2Gene on internal datasets, against human experts and validated it on external datasets. We have shown that Eye2Gene can greatly improve the phenotyping capabilities required in the genetic diagnosis of IRDs.

So far, only four previous studies have tried to apply AI to IRDs, all in much smaller datasets of fewer than 150 patients and across substantially fewer genes (Supplementary Table 1). By comparison, Eye2Gene, due to the benefit of having been trained on one of the world's largest datasets of genotyped individuals with IRD ($n = 2,451$), has potential for much broader utility, as it supports as many as 63 gene diagnoses, multiple imaging modalities and has conducted external validation.

In clinical settings, where decisions directly affect patient care, interpretability of AI decision support systems in terms of both the input images as interpreted by expert users and the output probabilities as interpreted by all users is essential for validating model outputs and building trust with patients and healthcare providers. Eye2Gene currently provides interpretability of input images (1) by attention maps (Supplementary Fig. 7) and (2) by visualization of cases in embedding space to identify phenotypically similar cases (Supplementary Fig. 8), and of output probabilities (3) by uncertainty estimates through conformal predictions (Supplementary Table 2 and Supplementary Fig. 9). In most applications, we anticipate Eye2Gene will be used with clinician oversight and backed up with genetic testing (where available), reducing the associated risks.

Of course, Eye2Gene predictions will not replace the need for genetic testing or counselling at specialized IRD centres, especially when a treatment such as a gene therapy is to be administered on the basis of a confirmed genetic diagnosis. Furthermore, in some

Patients	Gene	MOI	Phenotypes																			Exomiser rank	Eye2Gene rank		
			Rod-cone dystrophy	Abnormal electroretinogram	Cataract	Nyctalopia	Constriction of peripheral visual field	Visual impairment	Juvenile onset	Macular dystrophy	Sensorineural hearing impairment	Visual loss	Mild hearing impairment	Retinal dystrophy	Undetectable electroretinogram	Nystagmus	Cone/cone-rod dystrophy	Abnormal rod dystrophy	Abnormal light-adapted electroretinogram	Hypoplasia of the fovea	Achromatopsia			Abnormal central response of multifocal electroretinogram	Occult macular dystrophy
P1	USH2A	AR	■	■	■	■	■																2	1	
P2	ABCA4	AR						■	■	■														2	1
P3	USH2A	AR	■								■	■	■											2	1
P4	EYS	AR						■						■										5	1
P5	USH2A	AR	■	■	■	■	■																	3	1
P6	USH2A	AR	■	■	■	■	■																	3	1
P7	CRB1	N/S	■	■	■	■	■																	8	1
P8	ABCA4	AR										■			■	■	■							2	1
P9	ABCA4	AR						■									■	■	■					3	1
P10	CRB1	N/S										■			■	■	■							1	1
P12	RPGR	AR						■			■													3	1
P13	ABCA4	AR						■			■													1	1
P14	USH2A	AR	■									■	■	■										2	1
P15	CNGB3	AR						■							■			■	■					1	1
P16	CNGB3	AR						■						■			■				■			14	4
P17	USH2A	AR	■									■	■	■										2	2
P18	PROM1	AR						■								■	■	■						1	1
P19	BBS1	AR	■	■	■	■	■																	8	3
P20	EYS	AR	■	■	■	■	■																	1	5
P21	RPGR	XL	■	■	■	■	■																	5	1
P22	PROM1	AR	■	■	■	■	■																	1	1
P23	RP1L1	AR																			■	■		1	1
P24	GUCY2D	AR/AD										■			■	■	■							3	2
P25	CRB1	AR	■	■	■	■	■																	9	5
P26	CERKL	AR	■	■	■	■	■																	3	5
P27	RP1L1	AR	■	■	■	■	■																	2	3

Fig. 2 | Eye2Gene for gene prioritization. Phenotype grid for sample of 27 from 130 individuals with IRD for which the Exomiser-hiPHIVE gene rank based on HPO-only was compared to the Eye2Gene gene rank based on retinal scans. Eye2Gene outranks the HPO-only approach in 75% of the cases. Each row represents a patient. Each column represents an HPO term.

Dark blue cells represent presence of HPO term and light blue represent absence. MOI indicates mode of inheritance, which can be autosomal recessive (AR), autosomal dominant (AD) or X-linked (XL). Note that some HPO terms are not retinal specific such as sensorineural hearing impairment and mild hearing impairment. N/S, not specified.

contexts using more transparent approaches may be required, for example when conducting longitudinal analysis that Eye2Gene is not designed for. In these instances, quantitative analysis of biomarkers with conventional statistical approaches, for example using AI segmented imaging features via AIRDetect³⁸, may be preferable to ‘black box’ deep learning-based approaches such as Eye2Gene. Therefore, currently, the main anticipated application of Eye2Gene is in facilitating more efficient cost-effective genetic diagnosis by indicating when molecular testing would be worthy of consideration, and guiding specific genetic testing and/or its interpretation (Supplementary Fig. 10). It could also open the door to point-of-care identification of potential causative genes especially in geographic regions where genetic testing is not available³⁹.

Human expert benchmarking showed that the task of identifying likely genes from retinal phenotypes is very challenging even for IRD clinical experts with many years of experience. However, retinal

phenotyping is a task that is often required of clinician experts, as they either need to provide input as to whether a patient with a retinal condition should undergo genetic testing, or whether the retinal phenotype is in keeping with a specific genetic condition when interpreting the results of a genetic test as part of multidisciplinary team meetings⁴⁰. Additionally, the process of gene discovery involves identifying other individuals with similar retinal phenotypes. We have shown that the Eye2Gene AI is very adept at these tasks.

Eye2Gene gene predictions aid the interpretation of the results of a multigene genetic test by scoring and thereby prioritize genetic variants where the gene matches the phenotype²¹. In disorders with a highly distinct phenotype, the matching phenotype PP4 criterion in the American College of Medical Genetics guidelines provides a higher level of evidence in variant classification when the phenotype is consistent with the genetic aetiology, which might be crucial for the transition from a variant of unknown significance to a likely pathogenic variant.

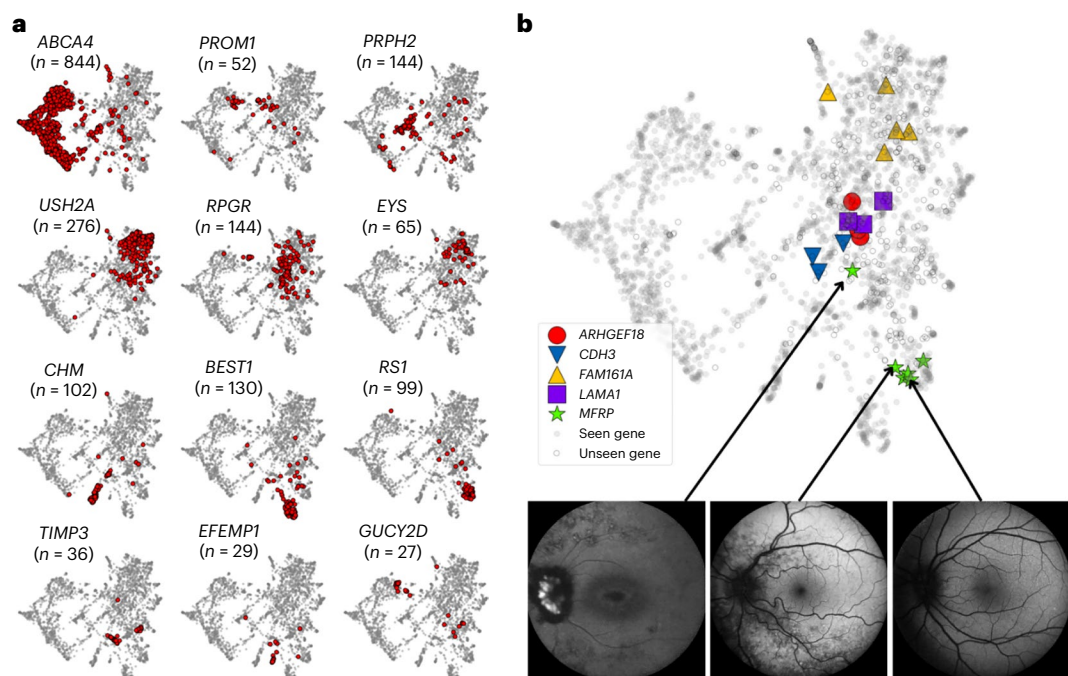


Fig. 3 | Visualization of Eye2Gene embeddings. a, Two-dimensional visualization of the embeddings obtained from Eye2Gene for select genes. Each point corresponds to an individual FAF scan. Each point in red represents a scan from a patient with the corresponding gene. A total of 170 unique genes are represented, a full list is included in the supplementary materials (Supplementary Fig. 6). These 2D embeddings are obtained by applying the

UMAP dimensionality reduction algorithm to the penultimate layer of Eye2Gene, a 768-dimensional vector. **b**, Embeddings for five unseen genes. Solid circles represent individuals with a gene diagnosis for one of the 63 genes that Eye2Gene was trained on, hollow circles represent individuals from other 'unseen' genes. Five exemplar genes, not included in the Eye2Gene training dataset of 63 genes, are highlighted by the different symbols.

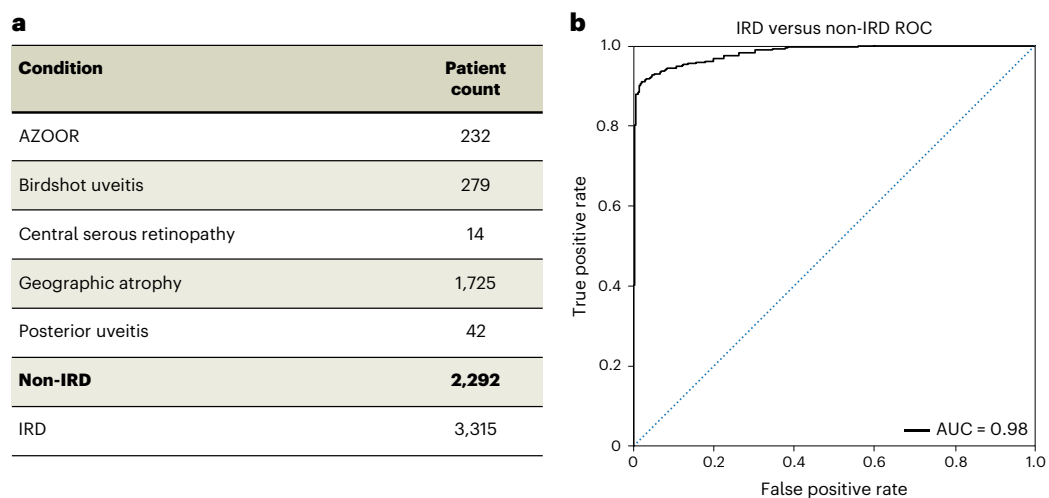


Fig. 4 | Eye2Gene as a screening model. a, Dataset including individuals with IRD and non-IRD. Individuals with a non-IRD were selected on the basis of having a condition with a similar retinal presentation in FAF. Bold indicates the total count of non-IRD individuals. AZOO, acute zonal outer occult retinopathy.

b, ROC classifier of a binary CoAtNet deep learning classifier trained on dataset **a**. The accuracy is 93.8% on the test set of 20% of the images, with an AUROC of 0.98. AUC, area under the curve.

This approach has been shown to improve diagnostic yield for other highly heterogeneous disorders such as syndromic intellectual disability and can improve the diagnostic yield in other pipelines where over 10% of the variants are of unknown significance⁸. Using next-generation phenotyping approaches such as Eye2Gene, GestaltMatcher²³ and DeepPlasia⁴¹ for monogenic conditions it is now possible to incorporate PP4 into variant interpretation in an objective way, which could increase the diagnostic yield for these conditions without relying on phenotypic labels such as HPO terms that can be ambiguous and also increase the risk of unwanted incidental findings in genomic pipelines.

Eye2Gene is an AI algorithm that demonstrates strong potential for clinical use in the challenging area of IRDs. Despite these promising results, we recognize that there are still several limitations to Eye2Gene and it remains in active development to address and alleviate these issues.

First, Eye2Gene accuracy is affected by the gene distribution of the target dataset (Supplementary Fig. 11), which we show in Table 1 accounts for most of the variability seen in Eye2Gene accuracy on the basis of the anticipated top-five accuracy across external sites. While the Eye2Gene development and test datasets are likely to closely

match the underlying gene distribution of the underlying patient population of London, UK, where MEH is based, gene distributions can vary between different patient populations and across patients of different ethnic backgrounds (Supplementary Fig. 12). For example, Eye2Gene appears to perform slightly worse on patients from Asian and South Asian backgrounds (although not statistically significant) (Supplementary Fig. 3). This highlights the need to include diverse training datasets that are sufficiently representative of the target patient population(s), and avoid widening existing health disparities⁴². In the future we plan to include data collected from more global sites into our training data, particularly those with large patient populations from non-European ethnic backgrounds such as from Asia and Africa. To this end strategies such as federated learning may be critical in reducing barriers surrounding sharing of data⁴³. We are also considering other methodological approaches for mitigating dataset bias such as data augmentation with generative AI and nearest matches based approaches for rare classes^{22,44}. Nonetheless in spite of dataset limitations, as it stands, Eye2Gene performance is within the anticipated accuracy (<5%) as extrapolated from the MEH training dataset to the gene distribution of the different centres. Furthermore, Eye2Gene already generalizes well as the top-five accuracy is consistently above 70% even in the ethnically distinct Tokyo site where the ethnicity of patients is predominantly Asian (Table 1). We also did not find any statistically significant evidence of sex or age bias in Eye2Gene's accuracy, which suggests the bias is unlikely to be large. Eye2Gene's ability to generalize is in part due to the MEH training dataset being primarily based in London, which represents an ethnically diverse population that receives patients from all over the UK (Supplementary Table 2 and Supplementary Fig. 13). Although we did not identify any large systematic difference in terms of image quality between the sites (Extended Data Table 3), accuracy is known to be linked to image quality⁴⁵ (Supplementary Fig. 14). In future work it may also prove useful to do a more in-depth analysis of how image acquisition parameters may influence Eye2Gene results.

Second, Eye2Gene is currently limited to predicting 63 gene classes out of a potential of 281 genes that are currently known to be associated with IRDs, not all of which are present in the Moorfields dataset (Supplementary Fig. 15), and hence cannot be used to predict IRDs that are not in those 63. However, we found that Eye2Gene embeddings were effective in identifying phenotypically similar cases, even for genetic conditions that Eye2Gene was not trained to identify (Fig. 3). In the future, by leveraging these embeddings, Eye2Gene could be extended to identify additional or similar IRDs, for example through a matching approach similar to GestaltMatcher²², which could also have positive effects for accuracy on rare genes.

Third, the comparison with experts in the present study was based on images only. We acknowledge that although this task does not capture the full scope of standard clinical practice as clinical decisions include other clinical parameters, it demonstrates the strength of Eye2Gene and its potential in assisting even IRD specialists in tasks such as genetic results interpretation that could include the secondary analysis exploration of the genetic data for less obvious genetic candidates such as non-coding or structural variants.

Future comparisons could also incorporate additional clinical information, however, our experiment demonstrates that Eye2Gene already has a role as an adjunct tool that experts could use in multidisciplinary team meetings to assist image interpretation²³.

In conclusion, Eye2Gene shows that next-generation phenotyping using AI is a promising approach to aid in the genetic diagnosis for individuals with IRDs, something that is not only important for improving patient experience and reducing associated overheads, but is likely to become especially important due to the growing number of potentially treatable IRDs where a rapid genetic diagnosis can lead to an improved outcome for the patient⁴⁶.

Methods

Dataset quality control and preparation

The MEH IRD cohort was previously described by Pontikos et al.¹ and encompasses 4,501 individuals with IRDs caused by variants in 189 distinct genes, of which 324 individuals (with variants in 72 genes) were younger than 18 years of age as of 2 August 2019. Individuals with an IRD and a confirmed genetic diagnosis by an accredited genetic diagnosis laboratory were identified and information about the genetic diagnosis was exported from the MEH electronic health record (OpenEyes) using a SQL query on the Microsoft SQL Server hospital data warehouse database.

Images were exported from the MEH Heidelberg Imaging (Heyex) database (Heidelberg Engineering) for all individuals with an IRD, on the basis of their hospital number, for records between 25 March 2004 and 22 October 2019. We selected the Heidelberg Spectralis as it is one of the most widely used medical imaging devices in IRD clinics worldwide and has previously been applied to AI-based approaches on IRDs. This resulted in a dataset of 2,103,692 images from 264,299 scans in 4,510 patients. For the quality control and data preparation process, images were divided by modality, with 87,534 short-wavelength FAF images in 4,000 patients, 35,608 IR images in 3,731 patients and 1,647,349 SD-OCT images in 3,731 patients. Since SD-OCT produces several B-scans, for each SD-OCT volume we selected only the median four B-scans corresponding to the four scans closest to the scan that traverses the fovea, as they were likely to be the most informative. Following this, 141,895 SD-OCT B-scans remained in 3,728 patients.

For all three modalities, we applied the filtering as shown in Supplementary Fig. 16. Any corrupted (unreadable) images were discarded. FAF scans feature two different imaging magnification levels, 30 degrees and 55 degrees. We kept 55-degree images and all other images were discarded, using data from the scan metadata to distinguish the two cases.

To remove low-quality and defective images, we used Retinograd-AI model to filter out poor-quality scans⁴⁷. These models were applied to the all FAF and SD-OCT images (using only the median B-scan for SD-OCT) within our dataset to obtain a prediction for each, and then all scans with a gradeability score of partially and un-gradeable were rejected. Since IR and SD-OCT scans are captured simultaneously, for IR scans we took the gradeability score of the corresponding SD-OCT volume as the label and filtered similarly. After this process, 27,433 FAF, 33,706 IR and 134,293 (33,712 volumes) SD-OCT images remained in 3,315, 3,715 and 3,715 patients, respectively.

The number of images rejected at each stage of the process is provided in Supplementary Fig. 16. To ensure sufficient data for training and testing, we restricted our datasets to only genes with at least ten patients remaining after filtering, leaving 63 individual genes. The distribution of all 63 genes is presented in Supplementary Fig. 10 and the full breakdown by gene is given in Extended Data Table 2. Following the quality control and gene selection process, 25,233 FAF, 31,357 IR and 124,975 (31,363 volumes) SD-OCT images remained across 3,652 patients in 63 distinct genes. The phenotype distribution of a subset of 2,103 of these patients is provided in Supplementary Table 4 per gene and per phenotype in Supplementary Table 5. The visual acuity distribution across genes is provided in Supplementary Fig. 17.

Postquality control, these patients were split into a 'development' set of 3,128 patients, and a held-out internal test set of 524 patients (28,174 images). Stratified sampling was used to ensure at least three representative patients for each gene were present in the test set, and to ensure no families were present in both test and development sets. The development set was further split into train and validation sets according to an approximate 80/20 split, with 2,451 patients (119,755 images) in the training set and 677 patients (31,605 images) in the validation set. The training set was used to train our 15 constituent Eye2Gene networks, while the internal test set was kept separate to enable testing of the final Eye2Gene model.

In addition to the MEH data described above we also obtained images from a further five centres to enable external validation of Eye2Gene. Oxford Eye Hospital (UK) provided a sample of 29,145 scans from 390 patients with distinct gene diagnoses in 33 different genes. The University Eye Hospital of Liverpool (UK) provided a sample of 6,174 scans from 156 patients with distinct gene diagnoses in 27 different genes. The University Eye Hospital Bonn (Germany) provided a sample of 2,838 scans from 129 patients with distinct gene diagnoses in 12 different genes. The Tokyo Medical Center (Japan) provided a sample of 1,493 scans from 60 patients with distinct gene diagnoses in 24 different genes. The Federal University of Sao Paulo (Brazil) provided a sample of 1,494 scans from 40 patients with distinct gene diagnoses in ten different genes. The MEH (UK) internal test dataset consisted of 28,174 scans from 524 patients across 63 gene diagnoses. Further breakdown by dataset is available in Extended Data Table 3 and with further breakdown by gene in Supplementary Table 6. Retrospective images from patients were selected by the clinical team at each of the five external centres according to the requirements: (1) that the patients had a confirmed genetic diagnosis within one of the 63 genes that are currently recognized by Eye2Gene; (2) the patient had retrospective retinal imaging available that was 55-degree FAF images or 30-degree OCT images obtained from the Heidelberg Spectralis as part of routine care and (3) the retinal images were considered of good quality. No further requirements were given regarding the ethnicity or the sex of the cases. For each patient a set of scans was selected, typically consisting of one scan per-patient per-eye per modality, along with their genetic diagnosis. These data were shared with us through our secure online portal, with the exception of the data from Bonn where the Eye2Gene models were run locally. No preprocessing or standardization of the retinal images took place but the image quality computed using Retinograd-AI⁴⁷ confirmed that the images were of comparable quality between centres and of slightly higher quality overall than those in the MEH test set given those ones were not explicitly selected by a clinician (Extended Data Table 3).

Model training

On each modality, five 63-class CoAtNets were trained for 100 epochs (passes over the entire dataset) for FAF and IR, and 25 epochs for OCT, which was found to be sufficient for the validation accuracy to converge in most settings (Supplementary Fig. 18). Random initialization with different random seeds was used for each individual of the 15 networks to ensure ensemble diversity. A CoAtNet0 architecture from the keras-cv-attention-models pypi library was used, where the final output layer was replaced by a dropout layer, followed by a linear layer with 63 outputs and softmax normalization. The CoAtNet architecture was chosen on the basis of an initial comparison of a number of different architectures evaluated on the FAF dataset. Cross entropy loss was used for the loss function, using additional class-weighting inversely proportional to gene frequency in the dataset, where the labels were given by the gene diagnosis of the underlying patients. This was to help address dataset imbalance due to the non-uniform gene distribution. For training, the Adam optimizer was used with the default parameters used in the Keras library ($\beta_1 = 0.9$, $\beta_2 = 0.999$). Learning rate was set to 0.0001 as it was found to work well across a range of architectures, and the batch size was set to 16, as this was the largest we could fit in graphical processing unit (GPU) memory. Dropout probability was fixed at 50%. Training was completed within 8 h for each neural network training on a single 3090 24 GB GPU.

To avoid overfitting to the training data, data augmentation techniques were applied. A number of plausible image transformations were applied automatically to the input data during training (Supplementary Fig. 19).

Evaluation

The output prediction of Eye2Gene is obtained by combining the predictions of its 15 constituent networks using an ensemble approach.

For each single retinal scan of a specific modality, a five-model ensemble is applied by taking the simple arithmetic mean of each of the output probabilities per gene across the five constituent networks. Given a collection of retinal scans from a single patient, the appropriate ensemble model corresponding to each scan's modality is applied, then the average predictions across all scans within each modality in the collection is taken in the same manner for a per-modality prediction per patient. Finally, the average across the three modalities is calculated and used as the final prediction for the patient. This approach was applied across all available scans per patient. Although there may be individual cases where it is better to down-weight or exclude certain images, or modalities overall we find that including more images per patient improves the overall accuracy (Supplementary Fig. 20). Additionally, we experimented with different class weightings, performing a grid search over modality weightings in 0.1 increments from 0 to 1.0 on the development validation set (the weights do not need to sum to 1 as only the relative class predictions affect the prediction). This improved validation top-five accuracy from 82.6 to 84.0%, with weights of 0.8, 0.1 and 0.5 for FAF, IR and OCT; however, applying this same weighting to the test data led to decreased top-five accuracy from 83.9 to 83.5%. With sufficient calibration data it may be possible to more accurately determine the optimal modality weighting; however, in the absence of other evidence, equal weighting provided a sufficiently good heuristic.

The model predictions were then compared against the underlying gene diagnoses for each patient to compute the overall accuracy of the model on the test data, the top- k accuracy (the proportion of images where the correct gene was within the highest k predictions of the network) for $k = 2, 3, 5, 10$, and the average per-class F1, weighted F1, mean average precision (MAP) and AUROC (Supplementary Table 7). Accuracy was calculated by counting the number of times Eye2Gene's top prediction matched the gene of the underlying patient. Per-gene precision-recall curves (for MAP) and ROC were produced for each gene in a one-versus-rest setup, using the Eye2Gene predictions for each output gene, and areas under the respective curves were calculated using trapezoid estimation (Supplementary Fig. 21). Confidence intervals were obtained by bootstrapping over 10,000 resamplings and taking the 2.5th and 97.5th percentiles. For convenience, all predictions were compiled into a single .CSV file along with additional data about each image (such as patient study ID, gene, appointment date) and the ID of the model used to generate the prediction. Eye2Gene combines predictions across multiple models (ensembling) and across multiple images acquired during one or more patient visits.

Taking each network individually without ensembling, the mean per-network top-five accuracies per image were 68.9, 70.8 and 74.9% for FAF, IR and OCT on our full validation dataset (which includes external sites). Applying ensembling of the five models per modality to the individual images we observed accuracies of 71.0, 72.7 and 77.2% for FAF, IR and OCT. Combining individual model predictions across multiple images (without ensembling the five models per modality) at the per-patient level, on the held-out test set, we observed an overall mean top-five accuracy across models of 81.5%. In general, we found that combining all images across all three modalities typically outperformed the best performing single modality (that is restricting to images of that modality only) on most genes, demonstrating the advantage of the multi-modality approach. In both cases these were superior to the single-network results, but inferior to the overall Eye2Gene model (83.9%), suggesting that both using ensembling across networks at a per-image level, and ensembling predictions across multiple images, was advantageous.

Conformal prediction

Conformal prediction sets construct a set of candidate classes instead of single class outputs. Crucially, this set is not fixed in size but is dynamically sized to reach some user-defined confidence threshold.

This means that for ‘easy’ examples, the prediction set may be very small (or even just one class) but allows for larger prediction sets for more ambiguous examples. By adjusting the confidence threshold, we can trade off between the proportion of example instances in which the correct class was in the prediction set, which is defined as coverage, and the set size. Conformal prediction sets can be constructed for any classifier with probability outputs and are a useful tool for interpretability of model output probabilities. The basic ‘naive’ conformal set construction algorithm is fairly simple and just increases the predicted classes included in the conformal prediction set until the desired confidence threshold is met; however, model outputs are often poorly calibrated in practice. Hence, various algorithms exist to calibrate conformal prediction sets (Supplementary Table 2).

We apply the least ambiguous adaptive prediction sets conformal prediction method, taking per-class probability outputs from Eye2Gene and adding the classes to the prediction set until the pre-determined confidence threshold was exceeded. We selected least ambiguous adaptive prediction sets from among three methods, as it produced smaller average prediction sets for a given coverage value, which we report in Supplementary Fig. 9. We calibrate our conformal prediction confidence levels on the MEH model validation set ($n = 677$) (not seen by the model during training) taking the compiled predictions for each patient.

Evaluating phenotype-driven genetic variant prioritization

Clinical notes and retinal scans of 130 patients with IRD from MEH with a confirmed gene diagnosis, who were part of the Eye2Gene test set, were manually reviewed and HPO terms were identified by three ophthalmologists with expertise in IRDs as described in Cipriani et al.²¹ These HPO terms were used as input for the latest version of the Exomiser-hiPHIVE algorithm (v.14.0.0) (<https://github.com/exomiser/Exomiser>) to obtain a gene ranking for the most probably predicted gene. The Exomiser-hiPHIVE algorithm uses a gene-specific phenotype score based on the PhenoDigm algorithm between the patient’s phenotype encoded as a set of HPO terms and the phenotypic annotation of any known gene-associated phenotypes reported in disease databases that include human and model organisms such as mouse and zebra fish. The retinal scans for these 130 patients with IRD were also analysed using Eye2Gene to obtain gene predictions that were also ranked accordingly for direct comparison with the Exomiser-hiPHIVE ranking. The non-parametric Wilcoxon rank sum test was used to compare the Exomiser-hiPHIVE and Eye2Gene gene rankings. Note that the Exomiser also takes as input the genetic variants file (VCF file), which also gives it an advantage over Eye2Gene in terms of reducing the set of possible genes to consider only genes that contain genetic variants that may be considered pathogenic.

Visualization and clustering of model embeddings

Visualizing and clustering model embeddings are valuable data-driven approaches for evaluating class diversity and identifying similarity between classes, even for new classes that the model has not been trained on. We applied one of the Eye2Gene FAF networks to all FAF scans in our test dataset, and the activations of the penultimate hidden layer were extracted to give a 768-dimensional vector for each scan. The UMAP dimensionality reduction algorithm was applied to the extracted activations, to obtain two-dimensional embeddings. The embeddings in UMAP space were then clustered using hierarchical clustering with Ward linkage, to produce a gene groupings dendrogram. By visualizing the UMAP-projected embeddings of the retinal images obtained from the different centres, we were able to show that no centre clustered separately and hence the images are unlikely to be systematically different (Supplementary Fig. 22). Using the embeddings, we were also able to derive a prototype-based methods inspired approach that matches the most similar images in embedding space according to the cosine similarity (Supplementary Fig. 8).

Eye2Gene screening component

For the screening component of Eye2Gene, a neural network was trained to distinguish FAF images of patients with IRD from patients with non-IRD. Hyperparameters, network architecture and training settings were the same as for the main Eye2Gene FAF module, except no ensembling across models or images was used. For the patients with non-IRD, a number of conditions were selected for presentation similar to IRDs in FAF imaging: acute zonal outer occult retinopathy, birdshot uveitis, central serous retinopathy, geographic atrophy and posterior uveitis. Patients with these conditions were extracted from the MEH hospital database and processed in the same manner as the IRD data ($n = 2,292$). Patients with IRD, before filtering for genes with more than one case, were selected ($n = 3,315$) (Supplementary Fig. 15). For evaluation, a held-out test set of 20% of patients was kept, and used as an evaluation set. For plotting of the ROC curve and area under the curve calculation, the outputs of the network were treated as a binary classification by taking the output probability of the IRD class as the predictive probability.

Interpretability of Eye2Gene image classifications

A well-known limitation of deep learning models in general is that their interpretability currently remains challenging and that existing approaches such as gradient-based saliency maps are not, as of now, sufficiently reliable for medical decisions. We leveraged the fact that the CoAtNet architecture uses self-attention to extract attention maps from one of the constituent Eye2Gene networks on a selection of FAF images. These attention maps show the areas of the image with the highest attention weights under the network’s self-attention mechanism, and hence the areas that are most strongly incorporated into the network’s final prediction (Supplementary Fig. 7). These maps were promising, consistently attending to areas of pathology, which are likely to be indicative of a particular condition according to our evaluation by human experts (Supplementary Fig. 22).

Human benchmarking

To contextualize the performance of Eye2Gene compared to human experts, a challenge dataset of 50 FAF images from patients sampled from the MEH held-out set was created, with 36 unique genes, and no more than two of any given gene. FAF was selected since it is one of the imaging modalities most commonly used in IRD clinics and hence the one for which ophthalmologists should overall have the most experience in the assessment of IRDs. We asked eight ophthalmologists from Moorfields (M.M., A.R.W., O.A.M.), Bonn (F.G.H., P.H., B.L.), Oxford (S.R.D.S.) and Liverpool (S.M.), specializing in IRDs, with 5–15 years of experience, to predict the causative gene based only on the images provided. These eight ophthalmologists (M.M., A.R.W., O.A.M., F.G.H., P.H., B.L., S.R.D.S., S.M.) were selected on the basis of the criteria of being board certified specialists in ophthalmic genetics who run dedicated IRD clinics at their respective hospitals and would have reviewed on average retinal images from hundreds of patients per year. For each image, each ophthalmologist was asked to name the five genes they thought was most likely out of a list of 36 genes. Eye2Gene was then run on the same images, taking the top-five predictions from the full 63 genes that Eye2Gene was trained on and then compared to the clinicians’ predictions.

Ethics

This research was approved by the Institutional Review Board and the UK Health Research Authority Research (HRA) Ethics Committee (REC) reference (22/WA/0049) ‘Eye2Gene: accelerating the diagnosis of inherited retinal diseases’ Integrated Research Application System (project ID 242050). The study sponsor was the University College London Joint Research Office (UCL JRO). The UCL JRO Data Protection reference number is Z6364106/2021/11/67. A summary of the

research study can be found on the HRA website (<https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/eye2gene-10/>). The REC that approved this study is Wales REC 5 (Wales.REC5@Wales.nhs.uk). All research adhered to the tenets of the Declaration of Helsinki.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are divided into two groups, published data and restricted data. Published data constitutes synthetic data derived from the Eye2Gene training dataset and are available from University College London at <https://doi.org/10.5522/04/28604234.v1> (ref. 48). In combination with the code at <https://github.com/Eye2Gene/Classification> (ref. 49), this can be used to train a smaller local version of Eye2Gene. Restricted data are curated for Eye2Gene users under a licence and cannot be published, to protect patient privacy and intellectual property. Access request to Eye2Gene datasets for the purpose of collaboration can be made via a contact form on the Eye2Gene website (www.eye2gene.com).

Code availability

The source code for model architecture training and inference is available at <https://github.com/Eye2Gene/Classification> (ref. 49). The code can also be run online via the CodeOcean capsule (<https://codeocean.com/capsule/0706698/>). The model weights of Eye2Gene are intellectual proprietary of UCLB so cannot be shared publicly. However, they may be shared via a licensing agreement with UCLB. A running version of the Eye2Gene web app is accessible at <https://app.eye2gene.com> and via the Heidelberg Appway on invitation. Access request to Eye2Gene can be made via a contact form on the Eye2Gene website (www.eye2gene.com). This is to limit risk of misuse of Eye2Gene.

References

- Pontikos, N. et al. Genetic basis of inherited retinal disease in a molecularly characterized cohort of more than 3000 families from the United Kingdom. *Ophthalmology* **127**, 1384–1394 (2020).
- Georgiou, M. et al. Phenotyping and genotyping inherited retinal diseases: molecular genetics, clinical and imaging features, and therapeutics of macular dystrophies, cone and cone-rod dystrophies, rod-cone dystrophies, Leber congenital amaurosis, and cone dysfunction syndromes. *Prog. Retin. Eye Res.* **100**, 101244 (2024).
- Lee, K. E. et al. A comprehensive report of intrinsically disordered regions in inherited retinal diseases. *Genes* **14**, 1601 (2023).
- Hanany, M., Rivolta, C. & Sharon, D. Worldwide carrier frequency and genetic prevalence of autosomal recessive inherited retinal diseases. *Proc. Natl Acad. Sci. USA* **117**, 2710–2716 (2020).
- Galvin, O. et al. The impact of inherited retinal diseases in the Republic of Ireland (ROI) and the United Kingdom (UK) from a cost-of-illness perspective. *Clin. Ophthalmol.* **14**, 707–719 (2020).
- Jimán, O. A. et al. Diagnostic yield of panel-based genetic testing in syndromic inherited retinal disease. *Eur. J. Hum. Genet.* **28**, 576–586 (2020).
- Sheck, L. H. N. et al. Panel-based genetic testing for inherited retinal disease screening 176 genes. *Mol. Genet. Genomic Med.* **9**, e1663 (2021).
- 100,000 Genomes Project Pilot Investigators et al. 100,000 Genomes Pilot on rare-disease diagnosis in health care – preliminary report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).
- Yohe, S. et al. Prevalence of mutations in inherited retinal diseases: a comparison between the United States and India. *Mol. Genet. Genomic Med.* **8**, e1081 (2020).
- Britten-Jones, A. C. et al. The diagnostic yield of next generation sequencing in inherited retinal diseases: a systematic review and meta-analysis. *Am. J. Ophthalmol.* <https://doi.org/10.1016/j.ajo.2022.12.027> (2022).
- Wong, W. Inherited retinal disease pathway in the UK: a patient perspective and the potential of AI. *Bri. J. Ophthalmol.* <https://doi.org/10.1136/bjo-2024-327074> (2025).
- Yung, M., Klufas, M. A. & Sarraf, D. Clinical applications of fundus autofluorescence in retinal disease. *Int. J. Retina Vitreous* **2**, 12 (2016).
- Tanner, A. et al. Clinical and genetic findings in CTNNA1-associated macular pattern dystrophy. *Ophthalmology* **128**, 952–955 (2021).
- Tanna, P. et al. Cross-sectional and longitudinal assessment of the ellipsoid zone in childhood-onset Stargardt disease. *Transl. Vis. Sci. Technol.* **8**, 1 (2019).
- Méjécase, C. et al. Practical guide to genetic screening for inherited eye diseases. *Ther. Adv. Ophthalmol.* **12**, 2515841420954592 (2020).
- Dai, Z., Liu, H., Le, Q. V. & Tan, M. CoAtNet: marrying convolution and attention for all data sizes. Preprint at <https://arxiv.org/abs/2106.04803> (2021).
- Karali, M. et al. Genetic epidemiology of inherited retinal diseases in a large patient cohort followed at a single center in Italy. *Sci. Rep.* **12**, 20815 (2022).
- Weisschuh, N. et al. Genetic architecture of inherited retinal degeneration in Germany: a large cohort study from a single diagnostic center over a 9-year period. *Hum. Mutat.* **41**, 1514–1527 (2020).
- Perea-Romero, I. et al. Genetic landscape of 6089 inherited retinal dystrophies affected cases in Spain and their therapeutic and extended epidemiological implications. *Sci. Rep.* **11**, 1526 (2021).
- Gurovich, Y. et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat. Med.* **25**, 60–64 (2019).
- Cipriani, V. et al. An improved phenotype-driven tool for rare Mendelian variant prioritization: benchmarking exomiser on real patient whole-exome data. *Genes* **11**, 460 (2020).
- Hsieh, T.-C. et al. GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. *Nat. Genet.* **54**, 349–357 (2022).
- Schmidt, A. et al. Next-generation phenotyping integrated in a national framework for patients with ultrarare disorders improves genetic diagnostics and yields new molecular findings. *Nat. Genet.* **56**, 1644–1653 (2024).
- Angelopoulos, A. N. & Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Preprint at <https://arxiv.org/abs/2107.07511> (2021).
- Georgiou, M., Fujinami, K. & Michaelides, M. Retinal imaging in inherited retinal diseases. *Ann. Eye Sci.* **5**, 25–25 (2020).
- Pichi, F., Abboud, E. B., Ghazi, N. G. & Khan, A. O. Fundus autofluorescence imaging in hereditary retinal diseases. *Acta Ophthalmol.* **96**, e549–e561 (2018).
- Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18**, 599–612 (2017).
- Kelly, C. et al. Phenotype-aware prioritisation of rare Mendelian disease variants. *Trends Genet.* <https://doi.org/10.1016/j.tig.2022.07.002> (2022).
- Jacobsen, J. O. B., Kelly, C., Cipriani, V., Robinson, P. N. & Smedley, D. Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases. *Brief. Bioinform.* **23**, bbac019 (2022).
- Lawrence, N. D. A unifying probabilistic perspective for spectral dimensionality reduction: insights and new models. *J. Mach. Learn. Res.* **13**, 1609–1638 (2012).

31. Casalino, G. et al. Autosomal recessive bestrophinopathy: clinical features, natural history, and genetic findings in preparation for clinical trials. *Ophthalmology* **128**, 706–718 (2021).
 32. Pontikos, N. et al. Phenogenon: gene to phenotype associations for rare genetic diseases. *PLoS ONE* **15**, e0230587 (2020).
 33. Miere, A. et al. Deep learning to distinguish ABCA4-related Stargardt disease from PRPH2-related pseudo-Stargardt pattern dystrophy. *J. Clin. Med. Res.* **10**, 5742 (2021).
 34. Miere, A. et al. Deep learning-based classification of inherited retinal diseases using fundus autofluorescence. *J. Clin. Med. Res.* **9**, 3303 (2020).
 35. Fujinami-Yokokawa, Y. et al. Prediction of causative genes in inherited retinal disorders from spectral-domain optical coherence tomography utilizing deep learning techniques. *J. Ophthalmol.* **2019**, 1691064 (2019).
 36. Fujinami-Yokokawa, Y. et al. Prediction of causative genes in inherited retinal disorder from fundus photography and autofluorescence imaging using deep learning techniques. *Br. J. Ophthalmol.* <https://doi.org/10.1136/bjophthalmol-2020-318544> (2021).
 37. Shah, M., Roomans Ledo, A. & Rittscher, J. Automated classification of normal and Stargardt disease optical coherence tomography images using deep learning. *Acta Ophthalmol.* **98**, e715–e721 (2020).
 38. Woof, W. A. et al. Quantification of fundus autofluorescence features in a molecularly characterized cohort of more than 3500 inherited retinal disease patients from the United Kingdom. *Ophthalmol. Sci.* **0**, 100652 (2024).
 39. Wong, W. M. et al. Practice patterns and challenges in managing inherited retinal diseases across Asia-Pacific: a survey from the APIED network. *Asia Pac. J. Ophthalmol.* **13**, 100098 (2024).
 40. Conway, M. P. et al. The role of the ophthalmic genetics multidisciplinary team in the management of inherited retinal degenerations—a case-based review. *Life* **14**, 107 (2024).
 41. Rassmann, S. et al. Deeplasia: deep learning for bone age assessment validated on skeletal dysplasias. *Pediatr. Radiol.* **54**, 82–95 (2023).
 42. Abuzaitoun, R. O. et al. Racial disparities in genetic detection rates for inherited retinal diseases. *JAMA Ophthalmol.* <https://doi.org/10.1001/jamaophthalmol.2024.4696> (2024).
 43. Yan, B. et al. FedEYE: a scalable and flexible end-to-end federated learning platform for ophthalmology. *Patterns* **5**, 100928 (2024).
 44. Veturi, Y. A. et al. SynthEye: investigating the impact of synthetic data on artificial intelligence-assisted gene diagnosis of inherited retinal disease. *Ophthalmol. Sci.* **3**, 100258 (2023).
 45. Petzold, A. et al. Artificial intelligence extension of the OSCAR-IB criteria. *Ann. Clin. Transl. Neurol.* **8**, 1528–1542 (2021).
 46. Botto, C. et al. Early and late stage gene therapy interventions for inherited retinal degenerations. *Prog. Retin. Eye Res.* **86**, 100975 (2022).
 47. Naik, G. et al. Retinograd-AI: an open-source automated fundus autofluorescence retinal image gradability assessment for inherited retinal dystrophies. *Ophthalmol. Sci.* <https://doi.org/10.1016/j.xops.2025.100845> (2025).
 48. Pontikos, N. & Woof, W. Synthetic dataset of 100 fundus autofluorescence of inherited retinal disease. *University College London* <https://doi.org/10.5522/04/28604234.v1> (2025).
 49. Silva, A., Woof, W., Moghul, I. & Pontikos, N. Eye2Gene/ classification: *Nature Machine Intelligence* publication. *Zenodo* <https://doi.org/10.5281/ZENODO.15039304> (2025).
- MEH NHS Foundation Trust and UCL Institute of Ophthalmology (grant no. NIHR203322). N.P., W.A.W., M.M., M.S., S.M.D., K.B., S.M., K.F. and J.F. are funded by an Artificial Intelligence in Health and Care Award (NIHR AI Award grant no. AI_AWARD02488). The Artificial Intelligence in Health and Care Award is part of the NHS AI Laboratory, which has made funding available to accelerate the testing and evaluation of artificial intelligence technologies that meet the aims set out in the NHS Long Term Plan. The NHS AI Laboratory is a joint unit of teams from the Department of Health and Social Care and NHS England, driving forward the digital transformation of health and social care. <https://transform.england.nhs.uk/ai-lab/>. D.S., N.P., W.A.W. and M.M. are also funded by Sight Research UK (grant no. TRN004) to pilot Eye2Gene into the NHS. N.P., W.A.W., M.M., B.L. and I.M. are also funded by Medical Research Foundation and Moorfields Eye Charity (grant no. MRF-JF-EH-23-122) to extend Eye2Gene to multiple European sites. N.P. was also previously funded by Retina UK as part of the UK IRD Consortium, Moorfields Eye Charity Career Development Award (grant no. R190031A), HDRUK (grant no. MC_PC_18036) and by a Translational Innovation grant awarded by the UCL Translational Research Office, which has seed funded this work. The UCL Centre for Digital Innovation partnership with Amazon Web Services has also supported the online deployment of the Eye2Gene software at www.eye2gene.com. B.J. was partially funded by grant no. IIR-DE-002818 from Shire/Takeda and by the European Reference Network for Rare Malformation Syndromes, Intellectual and Other Neurodevelopmental Disorders (ERN-ITHACA). O.A.M. is supported by the Wellcome Trust (grant no. 206619/Z/17/Z). A.Y.L. is supported by an unrestricted and career development award from RPB, Latham Vision Science Awards, grant nos. NIH OT2OD032644, NEI/NIH K23EY029246 and NIA/NIH U19AG066567. This project was also supported by a generous donation by Stephen and Elizabeth Archer in memory of Marion Woods. The hardware used for analysis was supported by the BRC Challenge Fund (grant no. BRC3_027). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan Xp GPU used for this research. The views expressed are those of the authors and not the funding organizations.

Author contributions

N.P., W.A.W. and A.V. devised the experiment, analysed the data and wrote the paper. G.N., B.J., M.A.I.-A., A.J.H., I.M., K.P., M.G., M.P., M.S., J.Y., S.K.W., M.D.V., T.A.C.G., N.K., J.F., Q.N., B.L., C.M., F.M., G.A., P.H., K.F., J.S., S.M., S.R.D.S., F.G.H., K.B., A.R.W., O.A.M. and P.M.K. analysed the data. All co-authors critically reviewed the paper.

Competing interests

N.P., W.W., M.M. and A.R.W. are patent holders of PCT/EP2023/076614 filed by UCL Business. I.M. is a cofounder, share-holder and director of Phenopolis Ltd, the software company that developed www.eye2gene.com. N.P. is a cofounder and former share-holder and director of Phenopolis Ltd. A.Y.L. reports grants from Santen, personal fees from Genentech, personal fees from US FDA, personal fees from Johnson and Johnson, grants from Carl Zeiss Meditec, personal fees from Gyroscop, non-financial support from Microsoft and grants from Regeneron, outside the submitted work. M.M. has received consultancy or advisory board fees from MeiraGTx, Janssen Pharmaceuticals, Saliogen and Octant; travel grants from MeiraGTx and Janssen Pharmaceuticals and stock options from MeiraGTx. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s42256-025-01040-8>.

Acknowledgements

The research was supported by a grant from the National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) at

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42256-025-01040-8>.

Correspondence and requests for materials should be addressed to Nikolas Pontikos.

Peer review information *Nature Machine Intelligence* thanks Qiong Wu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

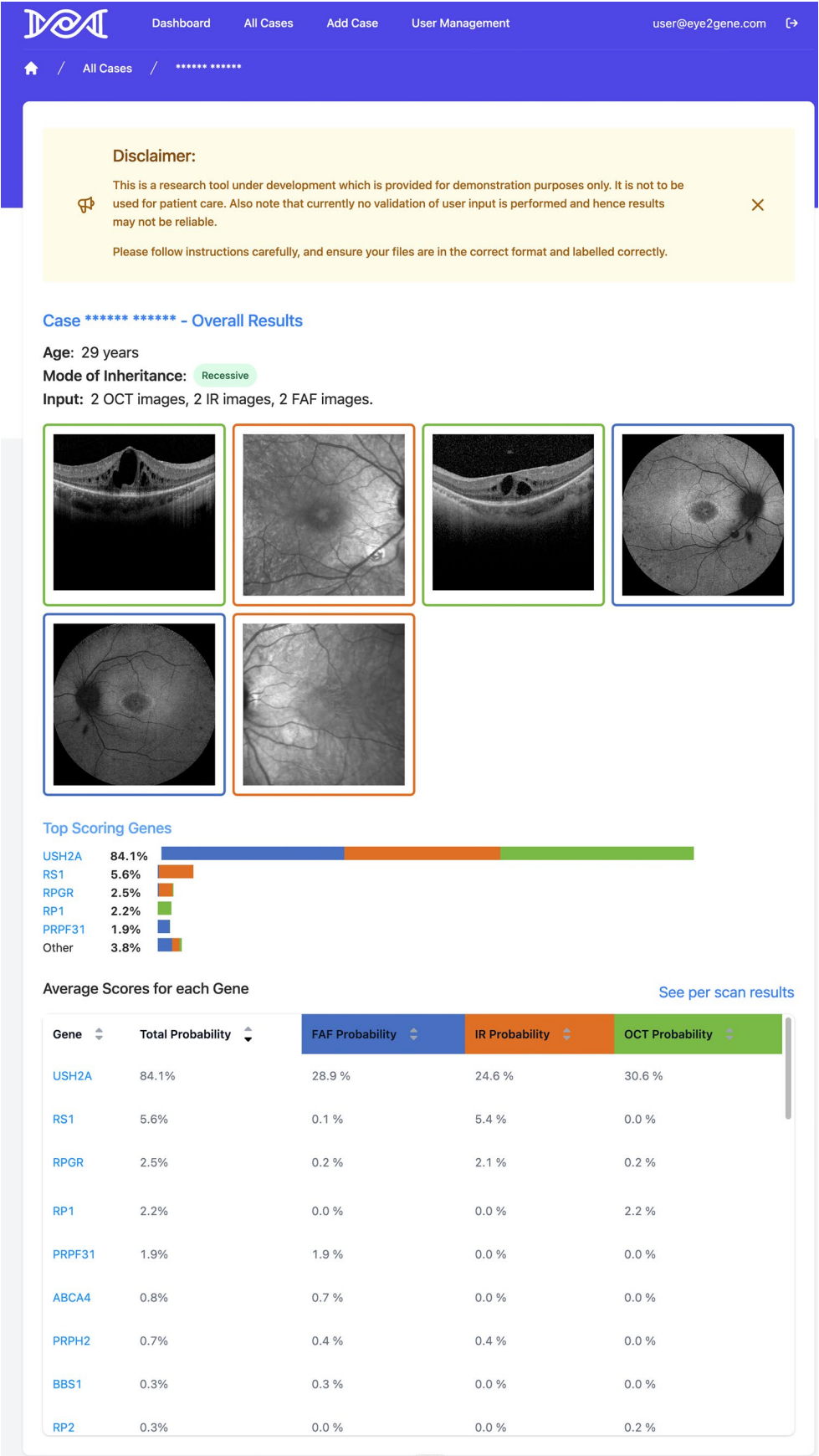
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

Nikolas Pontikos^{1,2}✉, William A. Woof^{1,2}, Siying Lin^{1,2}, Biraja Ghoshal^{1,2}, Bernardo S. Mendes^{1,2}, Advait Veturi¹, Quang Nguyen¹, Behnam Javanmardi³, Michalis Georgiou^{1,2}, Alexander Hustinx³, Miguel A. Ibarra-Arellano³, Ismail Moghul^{1,2}, Yichen Liu^{1,2}, Kristina Pfau^{4,5}, Maximilian Pfau^{4,5}, Mital Shah^{1,2,6}, Jing Yu⁷, Saoud Al-Khuzaei^{1,2,6,7}, Siegfried K. Wagner^{1,2}, Malena Daich Varela^{1,2}, Thales Antonio Cabral de Guimarães^{1,2,8}, Sagnik Sen^{1,2}, Gunjan Naik^{1,2}, Dayyanah Sumodhee^{1,2}, Dun Jack Fu^{1,2}, Nathaniel Kabiri¹, Jennifer Furman⁹, Bart Liefers², Aaron Y. Lee^{10,11}, Samantha R. De Silva^{1,2}, Caio Marques^{1,2}, Fabiana Motta¹², Yu Fujinami-Yokokawa¹³, Alison J. Hardcastle^{1,2}, Gavin Arno^{1,2}, Birgit Lorenz⁴, Philipp Herrmann⁴, Kaoru Fujinami¹³, Juliana Sallum¹², Savita Madhusudhan¹⁴, Susan M. Downes^{1,2}, Frank G. Holz⁴, Konstantinos Balaskas^{1,2}, Andrew R. Webster^{1,2}, Omar A. Mahroo^{1,2}, Peter M. Krawitz³ & Michel Michaelides^{1,2}

¹University College London Institute of Ophthalmology, University College London, London, UK. ²Moorfields Eye Hospital, London, UK. ³Institute for Genomic Statistic and Bioinformatics, University Hospital Bonn, Rheinische-Friedrich-Wilhelms University, Bonn, Germany. ⁴Department of Ophthalmology, University Hospital Bonn, Rheinische-Friedrich-Wilhelms Universität Bonn, Bonn, Germany. ⁵Department of Ophthalmology, University Hospital Basel, Basel, Switzerland. ⁶Oxford Eye Hospital, John Radcliffe Hospital, Oxford, UK. ⁷Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, Oxford, UK. ⁸Department of Ophthalmology, Faculdade São Leopoldo Mandic, Campinas, São Paulo, Brazil. ⁹UCL Translational Research Office, UCL Maple House, London, UK. ¹⁰Department of Ophthalmology, University of Washington School of Medicine, Seattle, WA, USA. ¹¹Roger and Angie Karalis Johnson Retina Center, University of Washington, Seattle, WA, USA. ¹²Department of Ophthalmology and Visual Sciences, Escola Paulista de Medicina, Federal University of São Paulo, São Paulo, Brazil. ¹³Laboratory of Visual Physiology, Division of Vision Research, National Institute of Sensory Organs, National Hospital Organization Tokyo Medical Center, Tokyo, Japan. ¹⁴St Paul's Eye Unit, The Royal Liverpool and Broadgreen University Hospitals, Liverpool, UK. ✉e-mail: n.pontikos@ucl.ac.uk



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Output of Eye2Gene app on an example case of a patient with disease-causing variants in the *USH2A* gene. The user is presented with a bar chart of the top 5 genes as predicted by the Eye2Gene model, along with the model probability score for each gene. These predictions are broken down into the contributions of the three different modalities, which are displayed as different colors on the bar-graph (blue for fundus autofluorescence, orange for

infrared and green for optical coherence tomography). The input images, color-coded by modality, and patient information are included at the top of the display. A full breakdown, with predicted probabilities for all 63 genes, is presented in the table below the bar graph. A link in the top right of the table takes the user to a breakdown of Eye2Gene's predictions for each of the uploaded images.

Extended Data Table 1 | Overview of Eye2Gene conformal prediction set sizes at different confidence levels

Confidence Level	Empirical Coverage	Average Set Size
0.95	94.4%	14.3
0.90	90.5%	8.1
0.85	86.2%	4.3
0.80	81.9%	2.7
0.75	74.7%	1.6
0.70	70.2%	1.3
0.65	66.5%	1.0

Overview of Eye2Gene conformal prediction set sizes at different confidence levels.

Extended Data Table 2 | Number of patients and images per gene and modality (FAF, IR and SD-OCT) for the 63 genes included in the full Moorfields Eye2Gene development dataset post-QC (n=3,652)

Gene	FAF		IR		SD-OCT		Patients with images from all 3 modalities	Age at first presentation range	Total (Female:Male)
	Patients	Images	Patients	Images	Patients	Images			
ABCA4	851	6851	877	8753	878	34893	777	10-91	478:476 (50.1% f)
ABCC6	17	280	18	398	18	1592	16	27-90	8:11 (42.1% f)
ADGRV1	10	81	12	98	12	392	10	26-77	4:8 (33.3% f)
BBS1	31	253	29	214	29	847	27	20-78	17:16 (51.5% f)
BEST1	131	1180	142	1152	142	4618	121	8-86	66:86 (43.4% f)
C1QTNF5	10	66	11	80	11	320	10	50-78	6:5 (54.5% f)
CACNA1F	28	211	29	156	29	624	21	9-79	2:34 (5.6% f)
CDH23	16	194	20	148	20	592	15	15-70	13:8 (61.9% f)
CDHR1	18	106	18	102	18	408	15	26-82	10:11 (47.6% f)
CEP290	15	151	12	56	12	224	10	11-57	6:11 (35.3% f)
CERKL	22	245	21	154	21	610	19	20-80	11:13 (45.8% f)
CHM	103	1594	115	1176	115	4704	98	11-86	18:102 (15.0% f)
CLN3	10	95	9	68	9	269	9	13-62	5:5 (50.0% f)
CLRN1	7	62	10	96	10	381	7	22-75	3:7 (30.0% f)
CNGA3	32	177	49	766	49	3016	28	11-67	29:24 (54.7% f)
CNGB1	25	131	25	165	25	660	23	26-88	15:12 (55.6% f)
CNGB3	38	229	65	1083	65	4245	31	9-80	39:34 (53.4% f)
CRB1	51	391	47	365	47	1454	41	8-80	16:41 (28.1% f)
CRX	26	207	24	176	24	704	22	20-85	10:18 (35.7% f)
CYP4V2	16	68	19	72	19	288	15	21-86	9:11 (45.0% f)
EFEMP1	29	237	33	237	33	948	27	30-79	23:12 (65.7% f)
EYS	65	444	70	482	70	1922	60	10-88	26:49 (34.7% f)
GPR143	6	28	9	39	9	144	5	14-69	4:6 (40.0% f)
GUC1A1A	11	85	10	64	10	256	10	29-73	4:7 (36.4% f)
GUCY2D	27	215	29	190	29	757	24	13-74	17:15 (53.1% f)
HGSNAT	11	98	10	150	10	600	10	53-86	4:7 (36.4% f)
IFT140	6	57	9	49	9	193	5	0-95	7:3 (70.0% f)
IMPDH1	7	52	9	35	9	140	6	18-85	6:4 (60.0% f)
IMPG2	12	88	11	85	11	340	10	26-75	5:8 (38.5% f)
KCNV2	24	150	26	181	26	724	23	9-77	14:13 (51.9% f)
LHON	3	11	11	54	11	213	3	10-77	4:7 (36.4% f)
MERTK	15	141	16	137	16	542	14	17-62	9:8 (52.9% f)
MFS3D8	12	142	10	104	10	410	10	35-86	4:8 (33.3% f)
MT-TL1	12	64	12	50	12	200	12	34-82	8:4 (66.7% f)
MYO7A	44	382	53	556	53	2224	41	12-85	26:30 (46.4% f)
NR2E3	26	333	27	286	27	1135	25	12-70	15:13 (53.6% f)
OPA1	26	127	70	321	70	1284	26	11-79	33:37 (47.1% f)
PAX6	9	27	9	30	9	120	5	16-48	5:8 (38.5% f)
PCARE	8	70	11	73	11	292	8	31-69	1:10 (9.1% f)
PDE6A	12	69	13	96	13	384	12	22-85	8:5 (61.5% f)
PDE6B	23	216	29	227	29	905	23	10-87	18:11 (62.1% f)
PDE6C	6	36	9	171	9	684	5	9-58	8:2 (80.0% f)
PROM1	47	374	40	287	40	1139	38	12-86	26:23 (53.1% f)
PRPF31	50	479	63	451	63	1801	45	12-78	43:25 (63.2% f)
PRPF8	14	118	24	206	24	824	14	20-76	15:9 (62.5% f)
PRPH2	145	1149	157	1150	157	4604	137	9-91	78:88 (47.0% f)
RDH12	28	176	28	189	28	744	21	9-67	18:17 (51.4% f)
RDH5	8	96	10	56	10	221	7	13-84	1:10 (9.1% f)
RHO	91	730	113	776	113	3104	87	9-95	66:51 (56.4% f)
RP1	107	660	114	792	114	3168	99	18-93	60:62 (49.2% f)
RP1L1	18	146	18	179	18	713	17	13-90	9:10 (47.4% f)
RP2	27	313	27	146	27	578	23	11-81	2:29 (6.5% f)
RP9	8	78	12	125	12	497	8	21-74	4:9 (30.8% f)
RPE65	12	80	30	297	30	1179	8	10-82	13:21 (38.2% f)
RPGR	144	1224	171	3031	171	12088	130	10-87	34:151 (18.4% f)
RS1	100	1145	110	963	110	3831	93	9-78	3:114 (2.6% f)
SNRNP200	10	60	13	102	13	408	9	32-78	7:7 (50.0% f)
TIMP3	36	324	37	366	37	1464	33	26-78	24:16 (60.0% f)
TULP1	15	151	15	121	15	481	13	15-69	8:9 (47.1% f)
TYR	11	78	5	11	5	35	4	10-68	7:5 (58.3% f)
USH1C	12	102	12	72	12	285	8	22-81	8:8 (50.0% f)
USH2A	276	2085	323	3070	323	12265	258	14-91	162:179 (47.5% f)
WFS1	4	21	13	72	13	288	4	18-79	8:5 (61.5% f)
Total	3014	25233	3373	31357	3374	124975	2735	0-95	1606:2046 (44.0% f)

Number of patients and images per gene and modality (FAF, IR and SD-OCT) for the 63 genes included in the full Moorfields Eye2Gene development dataset post-QC (n=3,652).

Extended Data Table 3 | Number of patients, images and genes in internal (Moorfields) and external (Oxford, Liverpool, Bonn, Tokyo, and São Paulo) test datasets. Last column describes the image quality as calculated by Retinograd-AI⁴⁷

Centre	Number of Patients	Number of Images				Num of Genes	Image Quality (mean / sd)
		FAF	IR	SD-OCT	Total		
Moorfields	524	3986	4862	19326	28174	63	0.858 +/- 0.0094
Oxford	390	633	756	27756	29145	33	0.938 +/- 0.0185
Liverpool	156	268	272	5634	6174	27	0.965 +/- 0.0190
Bonn	129	258	1290	1290	2838	12	0.981 +/- 0.1172
Tokyo	60	166	204	1123	1493	24	0.959 +/- 0.0839
São Paulo	40	111	141	1242	1494	10	0.938 +/- 0.0470
Total	1299	5422	6235	55313	68,028	63	-

Bold indicates summary across sites

Number of patients, images and genes in internal (Moorfields) and external (Oxford, Liverpool, Bonn, Tokyo, and São Paulo) test datasets. Last column describes the image quality as calculated by Retinograd-AI⁴⁷.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
<input type="checkbox"/>	<input checked="" type="checkbox"/> The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
<input type="checkbox"/>	<input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
<input type="checkbox"/>	<input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
<input type="checkbox"/>	<input checked="" type="checkbox"/> A description of all covariates tested
<input type="checkbox"/>	<input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
<input type="checkbox"/>	<input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
<input type="checkbox"/>	<input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
<input type="checkbox"/>	<input checked="" type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
<input type="checkbox"/>	<input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Retrospective deidentified imaging data and population characteristics were collected via the VisionSense eCRF (https://grading.readingcentre.org) as per the study protocol: https://liveuclac-my.sharepoint.com/:w:/g/personal/rmhanno_ucl_ac_uk/EZmNh0tHYHtBgWfHr-Tj758BW0R7sb1t7HvOI19kwT8T4A?e=EhzFaE All the software relating to this project is under the Github organisation: https://github.com/Eye2Gene/ . The code for training/testing the classification model is in the private Github repository: https://github.com/Eye2Gene/Classification . The running version of the web app is accessible at https://app.eye2gene.com and users are able to login with the following credentials: user: reviewer@nature.com password: naturemedicine1995
Data analysis	The Eye2Gene model was trained and evaluated in Python 3.10.2 (www.python.org) and Keras/Tensorflow 2.15.0 (https://keras.io/) with the keras-cv-attention-models v1.3.19 library from PyPi (https://github.com/0723sjp/keras_cv_attention_models). All training and prediction code is available at https://github.com/Eye2Gene/Classification . The code can also be ran online via the CodeOcean capsule (https://codeocean.com/capsule/0706698/). Results were analysed in JupyterLab 3.2.8 (https://jupyter.org/) with Python 3.9.7 (https://www.python.org/), NumPy 1.25.0 (https://numpy.org/), Pandas 1.4.0 (https://pandas.pydata.org/), SciPy 1.12.0 (https://scipy.org/), Scikit-Learn 1.0.2 (https://scikit-learn.org/), and Matplotlib 3.5.1 (https://matplotlib.org/).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The data that support the findings of this study are divided into two groups, published data and restricted data. Published data constitutes synthetic data derived from the Eye2Gene training dataset and are available from Figshare at <https://doi.org/10.5522/04/28604234.v1.50>. In combination with the code at <https://github.com/Eye2Gene/Classification> 49, this can be used to train a smaller local version of Eye2Gene. Restricted data are curated for Eye2Gene users under a license and cannot be published, to protect patient privacy and intellectual property. Access request to Eye2Gene datasets for the purpose of collaboration can be made via a contact form on the Eye2Gene website (www.eye2gene.com).

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	Gender information has been collected for the participants as recorded in their clinical record.
Population characteristics	Age at presentation, self-reported ethnicity, gender, mode of inheritance of condition and diagnostic gene associated with the condition.
Recruitment	Retrospective deidentified data collected during standard clinical care.
Ethics oversight	This research was approved by the IRB and the UK Health Research Authority Research (HRA) Ethics Committee (REC) reference (22/WA/0049) "Eye2Gene: accelerating the diagnosis of inherited retinal diseases" Integrated Research Application System (IRAS) (project ID: 242050). The study sponsor was the University College London Joint Research Office (UCL JRO). The UCL JRO Data Protection reference number is Z6364106/2021/11/67. A summary of the research study can be found on the HRA website (https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/eye2gene-10/). The REC that approved this study is Wales REC 5 (Wales.REC5@Wales.nhs.uk). All research adhered to the tenets of the Declaration of Helsinki.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Since we are dealing with a rare condition, the entirety of the IRD data available at Moorfields Eye Hospital was use for the training.
Data exclusions	Participants were excluded based on availability/quality of imaging or missing genetic diagnosis information. The participant list was further filtered to only include participants with a gene diagnosis in one of 63 genes for training the AI model.
Replication	External validation was conducted in 4 external datasets.
Randomization	During the Eye2Gene AI training process participants were randomized to 5 different folds.
Blinding	Human benchmarking and evaluation of the Eye2Gene AI algorithm involved blinding the expert ophthalmologists and the AI algorithm to the known diagnostic gene when presented with a retinal scan.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	Z6364106/2021/11/67, UCL Data Protection reference number
Study protocol	The study protocol is available for download here: https://liveuclac-my.sharepoint.com/:w:/g/personal/rmhanpo_ucl_ac_uk/EZmNh0tHYHtBgWfHr-Tj758BW0R7sb1t7HvOI19kwT8T4A?e=EhzFaE
Data collection	Retrospective de-identified data collection over 2006-06-05 to 2018-04-05.
Outcomes	The measured outcomes are the Eye2Gene AI model prediction top-5 accuracy for identifying the correct diagnostic gene based on retinal scans. We also compare the ranking of the correct gene between two methods.