Making Deep Learning Models Clinically Useful - Improving Diagnostic Confidence in Inherited Retinal Disease with Conformal Prediction

Biraja Ghoshal[1,2], William Woof[1,2], Bernardo Mendes[1,2], Saoud Al-Khuzaei[3], Thales Antonio Cabral De Guimaraes[1,2], Malena Daich Varela[1,2], Yichen Liu[1], Sagnik Sen[1,2], Siying Lin[1,2], Mital Shah[1,2], Yu Fujinami-Yokokawa[4], Andrew R. Webster[1,2], Omar A. Mahroo[1,2], Kaoru Fujinami[4], Frank Holz[7], Philipp Herrmann[7], Juliana Sallum[6], Konstantinos Balaskas[1,2], Savita Madhusudhan[5], Susan M Downes[3], Michel Michaelides[1,2], and Nikolas Pontikos[1,2]

[1] University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom
[2] Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom
[3] Oxford Eye Hospital, Oxford, United Kingdom
[4] Laboratory of Visual Physiology, Division of Vision Research, National Institute of Sensory Organs, National Hospital Organization Tokyo Medical Center, Japan
[5] St Paul's Eye Unit, Liverpool University Hospitals NHS Foundation Trust, Liverpool, United Kingdom
[6] Department of Ophthalmology and Visual Sciences, Escola Paulista de Medicina, Federal University of Sao Paulo, São Paulo, SP, Brazil
[7] Department of Ophthalmology, University Hospital Bonn, Rheinische-Friedrich-Wilhelms Universität Bonn, Germany b.ghoshal@ucl.ac.uk

Abstract. Deep Learning (DL), which involves powerful "black box" predictors, has achieved state-of-the-art performance in medical image analysis. However, these methods lack transparency and interpretability of point predictions without assessing the quality of their outputs. Knowing how much confidence there is in a prediction is essential for gaining clinicians' trust in the technology and its use in medical decision-making. In this paper, we explore the use of Conformal Prediction (CP) methods to recommend statistically rigorous reliable prediction sets to a clinician, using multi-modal imaging for the genetic diagnosis of the 36 most common molecular causes of inherited retinal diseases (IRDs). These are monogenic conditions that represent a leading cause of blindness in children and working-age adults and require a costly and time-consuming genetic test for diagnosis. Three methods of CP were assessed: Least Ambiguous Adaptive Prediction Sets (LAPS), Adaptative Prediction Sets (APS), and Regularized Adaptive Prediction Sets (RAPS). Our IRD classifier (Eye2Gene), in combination with the three conformal predictors, was evaluated on an internal holdout subset and datasets from four external clinical centres. RAPS proved to be the best-performing method with single-digit set sizes and coverage above 90% at a confidence level of 80%. Implementing adaptive CP methods has the potential to reduce waiting time and costs of genetic diagnosis of IRDs by improving upon the current gene prioritisation systems, while simultaneously enabling safety-critical clinical environments by flagging clinicians for a second opinion.

Keywords: Deep Learning · Uncertainty · Conformal Prediction · IRD.

1 Introduction

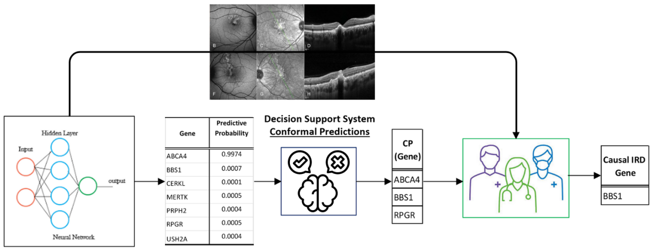
While the application of Deep Learning (DL) to medical imaging has achieved impressive performance standards, its focus on obtaining continuously higher point-prediction accuracy is not typically balanced with an evaluation of these predictions in safety critical automated medical decision-making. DL models generally lack uncertainty estimation or are not well calibrated which can lead to overconfident prediction [9]. Instead of overwhelming the accuracy of point prediction of the model, providing ordered estimates of outcomes that cover the true value with a statistical confidence guarantee, called uncertainty-aware conformal prediction set, is fundamental for the adoption of Artificial Intelligence (AI) into the clinical diagnosis pipeline [6].

The safe implementation of black-box models in medical diagnosis demands prediction of plausible diagnoses to be associated with uncertainty quantification in order to prevent consequential model failures. Epistemic uncertainty in deep learning is also referred to as model uncertainty, and is due to limited training data. There are multiple approaches to measure model uncertainty, which encompasses different methods to approximating full Bayesian neural networks [5, 7, 8]. A limitation of this approach is lack of any standard or intuitive meaning to aid clinicians in their decision making.

Conformal Prediction (CP) is a distribution-free uncertainty measurement framework for producing predictive sets with finite-sample, guaranteed predictive coverage without any model assumptions [2, 4, 16, 19–21]. Vovk (2015) and Barber et al. (2021) further improved statistical efficiency by reusing data for both training and calibration. The advantages of conformal prediction for uncertainty measurement are that it has a low computational cost, is compatible with any deep learning model, and provides a clinically-intuitive coverage guarantee [12].

This study uses three Conformal Prediction (CP) methods - Least Ambiguous Adaptive Prediction Sets (LAPS), Adaptive Prediction Sets (APS), and Regularized Adaptive Prediction Sets (RAPS) - to quantify the uncertainty for a large-scale, real-world inherited retinal diseases (IRDs) detection problem. IRDs are challenging to diagnose genetically and represent a leading cause of blindness in children and working-age adults worldwide. Clinicians learn to recognize phenotypic features of IRDs using high resolution retinal imaging technology such as fundus autofluorescence (FAF), infrared reflectance (IR) imaging, and spectral-domain optical coherence tomography (SD-OCT). Consequently, we trained our IRD classifier, Eye2Gene, on multimodal scans acquired from patients with IRDs seen at Moorfields Eye Hospital (MEH) who had undergone genetic testing and where a confirmed genetic cause had been identified by an accredited diagnostic laboratory.

In this study, we demonstrate how we applied Conformal Prediction to our underlying model, Eye2Gene, to select a set of potential causative genes from multimodal scans of patients with IRD with a confidence guarantee. We expect this approach to reduce costs and waiting time of genetic tests for IRDs by utilising personalised gene panels or by prioritising candidate genetic variants from whole genome sequencing based on the integrated output from the Eye2Gene AI model and the interpretation of the multidisciplinary diagnostic team.



Gene	Predictive Probability
ABCA4	0.9974
BBS1	0.0007
CERKL	0.0001
MERTK	0.0005
PRPH2	0.0004
RPGR	0.0005
USH2A	0.0004

Fig. 1: **Overview of clinical workflow.** Conformal Prediction (CP) selects a set of potential predictions from the underlying model. The clinician receives the recommended subset, together with the multi-modal images, and determines the most likely causative gene according to a significance level.

2 Methods

We investigate the effectiveness of multi-modal imaging and selection of a subset of causal genes in the diagnosis of Inherited Retinal Diseases (IRD) from multiple clinical centres by using three methods of Conformal Prediction: Least Ambiguous Adaptive Prediction Sets (LAPS) [18, 3], Adaptive Prediction Sets (APS) [17], and Regularized Adaptive Prediction Sets (RAPS) [2].

2.1 Dataset and Underlying Model

The MEH IRD dataset comprises the most extensive collection of retinal scans from patients with a confirmed molecular diagnosis of IRD. Information on the genetic diagnosis, age of presentation, and mode of inheritance for each patient was collated from the MEH electronic health record (OpenEyes), while their scans were obtained from the MEH Heidelberg Imaging (Heyex) database. As this project only utilised data previously collected and anonymised, informed consent was not sought from participants.

The Eye2Gene model was trained on a total of 44,817 scans from 1,907 patients (3749 eyes, 6397 appointments), from Moorfields Eye Hospital, split into three different modalities: FAF (N=13,509), IR (N=20,098) and OCT volumes

(N=11,344) and a calibration set of 13,012 scans. Currently, the development set contains 36 causative genes, which collectively, cover over 80% of molecularly characterized IRD cases in the European population [10, 14, 15, 22]. Test sets were assembled from a holdout set of retinal scans from MEH, and retinal scans from patients with IRDs from four external clinical centers: Oxford Eye Hospital (UK), Liverpool University Hospital (UK), University Hospital Bonn (Germany), and the Federal University of São Paulo (Brazil).

For each of the three modalities, five distinct Inception V3 deep deep convolutional neural networks (CNN) were trained, resulting in a total of fifteen neural networks with identical architecture but different network weights. For each modality these five networks are then combined into three modality-specific models via ensembling. The combination of these three models constitutes Eye2Gene. Each CNN was initialised with pre-trained ImageNet weights and the final layer was replaced by a liner layer with 36 outputs, followed by Softmax. These were trained for 100 epochs, with a batch size of 128, learning rate of 0.0001, and with the default Keras parameters for the Adam optimiser ($\beta_1 = 0.9, \beta_2 = 0.999$).

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 prediction. Given multiple scans from a single patient, 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 (post-softmax) scan-level predictions [11, 13].

	Training Sets		**Test Sets**				
Imaging Modality	**Deep Neural Network Training Set (n = 44,817)**	**Conformal Prediction Calibration Set (n = 13,012)**	**Moorfields (n = 1900)**	**Oxford (n = 346)**	**Liverpool (n=210)**	**Bonn (n = 473)**	**São Paulo (n = 104)**
FAF	13,509	4,124	733	106	70	241	16
IR	20,098	5,898	692	120	70	0	36
OCT	11,344	2,990	475	120	70	232	52

Table 1: Description of IRD dataset used for training of the underlying model and CP, as well as the test sets used for multi-site validation.

2.2 Conformal Prediction Methods

Given a test image x_i and a user-specified confidence level $\epsilon \in (0, 1)$, a conformal predictor outputs a prediction set $\mathit{\Gamma}_i^\epsilon \subseteq Y$ that contains the true class label $y_i \in Y$ with probability $1 - \epsilon$. In deep learning classification, the Softmax function is used to generate predictive probabilities. Conformal Predictors are implemented as an additional layer to deep learning algorithms and use a separate calibration step after training to predict confidence sets of classes. This framework as shown in Fig. 2 was developed in Python (v 3.10.12) with TensorFlow (v 2.12.0) and the Keras API, using a T4 GPU.

Least Ambiguous Adaptive Prediction Sets (LAPS): By sorting the class scores for an example and selecting the classes in the prediction set that exceed a predetermined confidence threshold, we employ a naive method of generating prediction sets [18, 3].

1. Train a deep learning model $f(x) : R^{W \times H \times C} \to (0, 1)^{36}$ on D_{train}

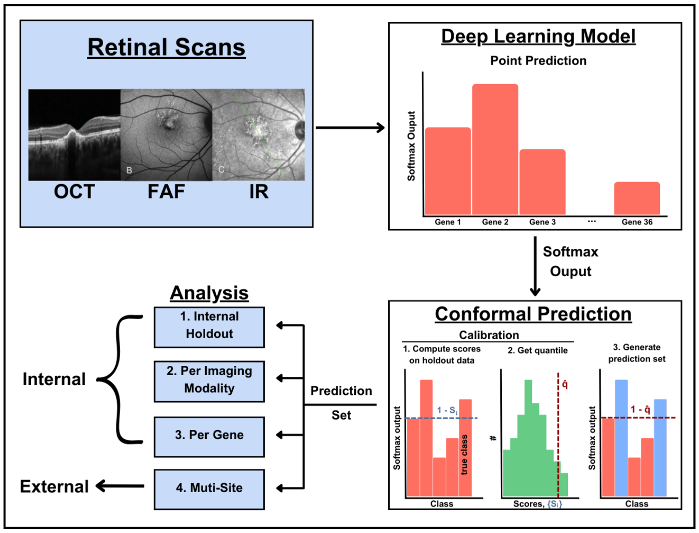


Fig. 2: **Study Design.** The study was divided into four experiments of three imaging modality and 36 gene class. Model generalization was evaluated in external test sets from four clinical centres: Oxford, Liverpool, Bonn, and São Paulo.

2. Define $s(x,y) = 1 - f(x)_{y_{true}}$ where $f(x)_{y_{true}}$ is the softmax output of the true class
3. Compute $s_1, s_2, ..., s_{n_{cal}}$ on the calibration set D_{cal}
4. Compute $\hat{q}$ as the $\frac{\lceil (n_{cal}+1)(1-\alpha) \rceil}{n_{cal}}$ quantile of the calibration scores.
5. Prediction sets with $1 - \alpha$ confidence as:

$$C(x_{test}) = \{y : f(x_{test})_y \geq 1 - \hat{q}\} \tag{1}$$

.

Adaptive Prediction Sets (APS): The prediction sets produced by the above naive method have a smaller average set size, but often undercover harder gene classes and overcover easier ones. We apply the Adaptive prediction sets (APS) method to compensate for this [17].

1. Suppose $\pi(x)$ a permutation of $f(x)$ that orders the softmax output in descending order, i.e., from the most likely class to the less likely.
2. Select $s(x,y) = \sum_{i=1}^{k} \pi(x)_y$ where k is the minimum number of class labels to go through until reach the true class.
3. $s(x,y) = \sum_{i=1}^{k} \hat{\pi}(x)_{y_i}$, where k is the minimum number of classes used until to find the true class.
4. Compute $\hat{q}$ as the $\frac{\lceil (n_{cal}+1)(1-\alpha) \rceil}{n_{cal}}$ quantile of the calibration scores $s_1 = s(x_1, y_1), ..., s_n = s(x_{n_{cal}}, y_{n_{cal}})$ on the calibration dataset.

5. Prediction sets with $1-\alpha$ confidence as:

$$C(x_{test}) = \{y : \sum_{i=1}^{k} \pi(x)_y \geq \hat{q}\}. \tag{2}$$

Regularized Adaptive Prediction Sets (RAPS): The APS method yields predictive sets with exact coverage, but these tend to have the largest average size. However, Regularized Adaptive Prediction Sets (RAPS) often delivers much smaller sets in practice [2].

1. Suppose $\pi(x)$ a permutation of $f(x)$ that orders the softmax output in descending order, i.e., from the most likely class to the less likely.
2. Choose $s(x,y) = \sum_{i=1}^{k} \pi(x)_y$ where k is the minimum number of classes we have to go through until reach the true class.
3. $s(x,y) = \sum_{i=1}^{k} (\hat{\pi_{\hat{i}}}(x_j) + \lambda[\hat{i} \geq k_{reg}])$, where where k is the model's ranking of the true class y_i and $\hat{\pi_{\hat{i}}}(x_j)$ is $\hat{i}$ the largest score for the $\hat{j}^{th}$ image.
4. Compute $\hat{q}$ as the $\frac{\lceil (n_{cal}+1)(1-\alpha) \rceil}{n_{cal}}$ quantile of the calibration scores $s_1 = s(x_1, y_1), ..., s_n = s(x_{n_{cal}}, y_{n_{cal}})$ on the calibration dataset.
5. Prediction sets $\hat{k}$ highest-score classes with $1-\alpha$ confidence as:

$$C(x_{test}) = \{y : \sum_{i=1}^{\hat{k}} (\hat{\pi_i}(x_{(n+1)} + \lambda[j \geq k_{reg}]) \geq \hat{q}\}. \tag{3}$$

Every class beyond the k_{reg} most likely classes is subject to a constant regularization penalty factor λ which deselects in the predictive set. APS is a special case of RAPS with $\lambda = 0$.

3 Results

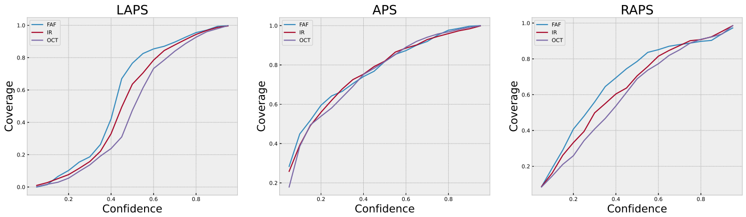
We applied the CP framework for the prediction of causal inherited retinal disease genes from multi-modal imaging using deep learning. We assessed the performance of CP in diagnosing IRDs for each imaging modality and gene class by evaluating the coverage and average prediction set size at different confidence levels.

3.1 Analysis of Holdout dataset

We performed a baseline evaluation to investigate the performance of the AI model and Conformal Predictor on scans by using a holdout subset of the MEH dataset. Eye2Gene was trained on the three most common imaging modalities for IRD diagnosis: FAF, IR, and OCT. Analysis of Table 2 and Figure 3 shows that LAPS and RAPS return predictive sets with higher coverage for FAF scans across confidence intervals, however, the coverage across modalities converges at the higher confidence levels for each CP method. OCT sets are significantly smaller than other imaging modalities for both LAPS and RAPS, while APS has similar sizes across modalities at all confidence levels. Classification with FAF scan shows the best improvement with CP, as it constitutes the modality with the lowest accuracy with point prediction.

Confidence		**85%**						**80%**						**75%**					
		Class Coverage			**Test Size**			**Class Coverage**			**Test Size**			**Class Coverage**			**Test Size**		
Imaging Modality	**Accuracy**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**
FAF	0.443	0.970	0.986	0.903	14	24	11	0.955	0.977	0.898	11	21	8	0.926	0.952	0.888	8	18	7
IR	0.458	0.968	0.974	0.923	15	24	12	0.944	0.96	0.908	12	21	9	0.912	0.945	0.902	9	18	8
OCT	0.516	0.960	0.981	0.922	13	24	9	0.926	0.968	0.909	8	21	7	0.886	0.958	0.888	6	19	6

Table 2: Coverage and set size of LAPS, APS, and RAPS conformal methods at different levels of confidence for three imaging modalities. Prediction set sizes from FAF scans have the highest average coverage for both LAPS and APS, however with RAPS, it is the lowest at these confidence levels. FAF scans show the best improvement in performance with CP when compared to its point prediction accuracy of 44.3%. Set sizes from OCT are the lowest for the three confidence levels for both LAPS and RAPS, while APS have similar sizes for all modalities.



(a) Coverage

Fig. 3: Performance across imaging modalities. LAPS and RAPS yield higher coverage for FAF scans. APS has similar coverage for all imaging modalities across confidence levels.

3.2 Analysis per Gene Class

Due to the large variety of IRD genes and their rarity, it is essential that the model meets sufficient performance for all classes. In Table 3, we report the coverage and set size of each CP method for the five most prevalent IRD genes in our cohort: *ABCA4*, *BEST1*, *PRPH2*, *RPGR* and *USH2A*. Table 4 represents the same metrics for three of the rarest genes in the MEH cohort: *CDH23*, *KCNV2* and *MTLL1*. From the prevalent gene group, we notice that the model accuracy for *PRPH2* and *RPGR* is substantially lower when compared to other genes in the group. This can be partially explained by phenotypic similarities to other IRD gene variants. *PRPH2* is the most common phenocopy of *ABCA4* (Stargardt's disease), as it mimics the phenotype of disease caused by the latter [1]. Both *RPGR* and *USH2A* are associated with Retinitis Pigmentosa, leading to similar clinical presentations. The classifier, therefore, has difficulty differentiating these two sets of genes based on their similar retinal scan features, leading to lower accuracy. From the rare group, we note that the classifier was unable to predict *KCNV2* scans correctly in the holdout test set. Despite LAPS and RAPS providing lower coverage for the under-performing genes when compared to others in their group, CP improves classification over their point prediction accuracy while maintaining a similar set size.

Confidence		**90%**						**80%**						**67%**					
		Class Coverage			**Set Size**			**Class Coverage**			**Set Size**			**Class Coverage**			**Set Size**		
Gene	**Accuracy**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**
ABCA4	0.950	1.000	1.000	0.997	13	21	15	0.993	1.000	0.988	7	13	7	0.983	0.993	0.980	3	8	3
BEST1	0.846	0.989	1.000	0.989	13	21	15	0.989	1.000	0.967	7	13	7	0.978	0.989	0.957	3	8	3
PRPH2	0.412	0.991	1.000	0.948	17	25	15	0.914	0.991	0.905	10	17	8	0.776	0.948	0.759	5	11	4
RPGR	0.353	0.994	1.000	0.975	24	32	15	0.969	0.994	0.869	15	24	9	0.769	0.988	0.744	8	17	6
USH2A	0.800	1.000	1.000	0.995	22	31	15	0.990	1.000	0.970	12	22	8	0.959	0.995	0.939	6	14	5

Table 3: Coverage and set size of conformal methods at different levels of confidence values for the five most prevalent gene classes.

Confidence		**90%**						**80%**						**67%**					
		Coverage			**Set Size**			**Coverage**			**Set Size**			**Coverage**			**Set Size**		
Gene	**Accuracy**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**
CDH23	0.500	0.964	1.000	0.857	25	34	15	0.857	0.964	0.786	14	25	9	0.786	0.893	0.786	7	16	6
KCNV2	0.000	1.000	1.000	1.000	23	31	15	0.846	1.000	0.769	13	23	8	0.615	0.923	0.615	8	14	6
MTTL1	0.667	1.000	1.000	0.875	23	33	14	0.813	1.000	0.813	11	23	8	0.750	0.875	0.750	5	13	5

Table 4: Coverage and set size of conformal methods at different levels of confidence values for the three rarest gene classes.

3.3 Multi-Site Analysis

Our model was externally validated on test sets obtained from four clinical centres. Large variation in the performance of the underlying classifier across external test sets is seen in Table 5. This can be partially attributed to significant differences in the distribution of genes between datasets, with the number of unique genes per set ranging from 3 (São Paulo) to 29 (Oxford). Note that in Figure 4, coverage for the IR modality is absent from the Bonn dataset and, thus, is not reported here.

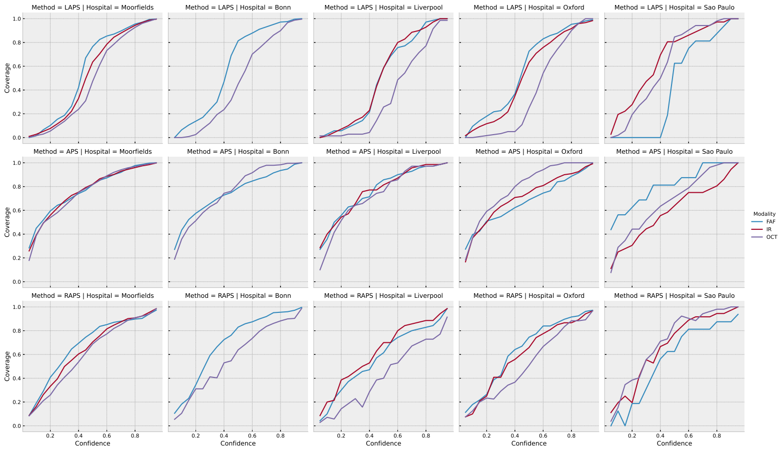


Fig. 4: Illustration of coverage of conformal prediction sets in test sets divided per imaging modality. This figure demonstrates that conformal prediction allows the successful generalisation of our classifier to external datasets for all three imaging modalities.

Confidence			**85%**						**80%**						**75%**					
			Class Coverage			**Set Size**			**Class Coverage**			**Set Size**			**Class Coverage**			**Set Size**		
Hospital	**Accuracy**	**Num of Unique Genes**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**	**LAPS**	**APS**	**RAPS**
Moorfields	0.667	32	0.966	0.980	0.916	14	24	11	0.942	0.968	0.905	10	21	8	0.908	0.952	0.893	8	18	7
Bonn	0.701	11	0.971	0.972	0.929	12	23	10	0.935	0.959	0.919	9	20	8	0.906	0.947	0.906	6	17	6
Liverpool	0.571	15	0.957	0.976	0.819	17	27	11	0.890	0.976	0.814	12	24	9	0.833	0.967	0.795	9	21	8
Oxford	0.542	29	0.960	0.947	0.900	15	25	11	0.923	0.932	0.888	11	22	9	0.874	0.916	0.865	8	19	7
Sao Paulo	0.867	3	0.970	0.954	0.933	13	22	10	0.943	0.929	0.933	9	19	7	0.900	0.913	0.897	7	17	6

Table 5: Coverage and set size at different levels of confidence values for one internal and four external test sets. A substantial variation in accuracy can be observed between the clinical centres, especially between Oxford and Sao Paulo where the disparity in accuracy is 32.5%. The Oxford dataset has the highest number of unique genes from the external centres, whereas Sao Paulo has the lowest. On the other hand, class coverage from these two datasets has a much smaller difference, e.g. for RAPS at 80% the difference is 4.5%.

4 Discussion

We demonstrated that RAPS provides suitable coverage and produces the smallest prediction sets out of the three CP methods, therefore being the most clinically informative method. Furthermore, we report that despite coverage varying for each imaging modality, it converges at confidence levels of around 80% and, therefore, CP can be implemented with any of the three modalities. However, we note that predictive set sizes for OCT are smaller with both LAPS and RAPS.

CP has the potential to reduce the disparity in performance between gene classes, as it can produce substantial improvements in the classification of genes, such as *PRPH2*, which have similar phenotypes to more prevalent classes. We further explored this feature by validating our approach on external test sets that possess vastly different gene distributions. For the test sets in which the underlying model showed the greatest variation in point prediction accuracy (Sao Paulo and Oxford), RAPS was shown to provide prediction sets with similar coverage and size. Therefore, we can conclude that Conformal Prediction can improve Eye2Gene's capacity to generalize across real-world datasets. To further improve our classifier, we plan to measure dissimilarity between patients' scans and classes in the predictive set, as well as introduce class balancing to the calibration set.

References

1. Al-Khuzaei, S., Broadgate, S., Foster, C.R., Shah, M., Yu, J., Downes, S.M., Halford, S.: An overview of the genetics of abca4 retinopathies, an evolving story. Genes **12**(8), 1241 (2021). https://doi.org/10.3390/genes12081241
2. Angelopoulos, A., Bates, S., Malik, J., Jordan, M.I.: Uncertainty sets for image classifiers using conformal prediction. CoRR **abs/2009.14193** (2020), https://arxiv.org/abs/2009.14193
3. Angelopoulos, A.N., Bates, S., Zrnic, T., Jordan, M.I.: Private Prediction Sets. Harvard Data Science Review **4**(2) (apr 28 2022), https://hdsr.mitpress.mit.edu/pub/deziirvg
4. Cauchois, M., Gupta, S., Duchi, J.: Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction (2020)
5. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (20–22 Jun 2016), https://proceedings.mlr.press/v48/gal16.html
6. Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Chen, I.Y., Ranganath, R.: Practical guidance on artificial intelligence for health-care data. The Lancet Digital Health **1**(4) (2019). https://doi.org/10.1016/s2589-7500(19)30084-6
7. Ghoshal, B., Ghoshal, B., Tucker, A.: Leveraging uncertainty in deep learning for pancreatic adenocarcinoma grading. In: Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings. pp. 565–577. Springer (2022)
8. Ghoshal, B., Tucker, A., Sanghera, B., Lup Wong, W.: Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. Computational Intelligence **37**(2), 701–734 (2021). https://doi.org/https://doi.org/10.1111/coin.12411, https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12411
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. CoRR **abs/1706.04599** (2017), http://arxiv.org/abs/1706.04599
10. Karali, M., Testa, F., Di Iorio, V., Torella, A., Zeuli, R., Scarpato, M., Romano, F., Onore, M.E., Pizzo, M., Melillo, P., Brunetti-Pierri, R., Passerini, I., Pelo, E., Cremers, F.P.M., Esposito, G., Nigro, V., Simonelli, F., Banfi, S.: Genetic epidemiology of inherited retinal diseases in a large patient cohort followed at a single center in italy. Sci. Rep. **12**(1), 20815 (Dec 2022)

11. Mendes, B.S., Woof, W., Ghoshal, B., Nguyen, Q., Naik, G., Bagga, P., Moghul, I., Fu, D.J., Shah, M., Al-Khuzaei, S., et al.: Extending eye2gene to quantify the phenotypic diversity and similarity of 63 inherited retinal diseases using an embedding approach. Investigative Ophthalmology & Visual Science **65**(7), 4653–4653 (2024)
12. Messoudi, S., Destercke, S., Rousseau, S.: Conformal multi-target regression using neural networks. In: Gammerman, A., Vovk, V., Luo, Z., Smirnov, E., Cherubin, G. (eds.) Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications. Proceedings of Machine Learning Research, vol. 128, pp. 65–83. PMLR (09–11 Sep 2020), https://proceedings.mlr.press/v128/messoudi20a.html
13. Nguyen, Q., Woof, W., Kabiri, N., Sen, S., Varela, M.D., Guimaraes, T.A.C.D., Shah, M., Sumodhee, D., Moghul, I., Al-Khuzaei, S., Liu, Y., Hollyhead, C., Tailor, B., Lobo, L., Veal, C., Archer, S., Furman, J., Arno, G., Gomes, M., Fujinami, K., Madhusudhan, S., Mahroo, O.A., Webster, A.R., Balaskas, K., Downes, S.M., Michaelides, M., Pontikos, N.: Can artificial intelligence accelerate the diagnosis of inherited retinal diseases? protocol for a data-only retrospective cohort study (eye2gene). BMJ Open **13**(3) (2023). https://doi.org/10.1136/bmjopen-2022-071043, https://bmjopen.bmj.com/content/13/3/e071043
14. Perea-Romero, I., Gordo, G., Iancu, I.F., Del Pozo-Valero, M., Almoguera, B., Blanco-Kelly, F., Carreño, E., Jimenez-Rolando, B., Lopez-Rodriguez, R., Lorda-Sanchez, I., Martin-Merida, I., Pérez de Ayala, L., Riveiro-Alvarez, R., Rodriguez-Pinilla, E., Tahsin-Swafiri, S., Trujillo-Tiebas, M.J., ESRETNET Study Group, ERDC Study Group, Associated Clinical Study Group, Garcia-Sandoval, B., Minguez, P., Avila-Fernandez, A., Corton, M., Ayuso, C.: Genetic landscape of 6089 inherited retinal dystrophies affected cases in spain and their therapeutic and extended epidemiological implications. Sci. Rep. **11**(1), 1526 (Jan 2021)
15. Pontikos, N., Arno, G., Jurkute, N., Schiff, E., Ba-Abbad, R., Malka, S., Gimenez, A., Georgiou, M., Wright, G., Armengol, M., Knight, H., Katz, M., Moosajee, M., Yu-Wai-Man, P., Moore, A.T., Michaelides, M., Webster, A.R., Mahroo, O.A.: Genetic basis of inherited retinal disease in a molecularly characterized cohort of more than 3000 families from the united kingdom. Ophthalmology **127**(10), 1384–1394 (oct 2020)
16. Romano, Y., Patterson, E., Candès, E.J.: Conformalized quantile regression (2019)
17. Romano, Y., Sesia, M., Candès, E.J.: Classification with valid and adaptive coverage. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H.T. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html
18. Sadinle, M., Lei, J., Wasserman, L.: Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association **114**(525), 223–234 (jun 2018). https://doi.org/10.1080/01621459.2017.1395341, https://doi.org/10.1080.2F01621459.2017.1395341
19. Shafer, G., Vovk, V.: A tutorial on conformal prediction. CoRR **abs/0706.3188** (2007), http://arxiv.org/abs/0706.3188
20. Vovk, V., Gammerman, A., Saunders, C.: Machine-learning applications of algorithmic randomness. In: Sixteenth International Conference on Machine Learning (ICML-1999) (01/01/99). pp. 444–453 (1999), https://eprints.soton.ac.uk/258960/

21. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World (01 2005). https://doi.org/10.1007/b106715
22. Weisschuh, N., Obermaier, C.D., Battke, F., Bernd, A., Kuehlewein, L., Nasser, F., Zobor, D., Zrenner, E., Weber, E., Wissinger, B., Biskup, S., Stingl, K., Kohl, S.: 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**(9), 1514–1527 (Sep 2020)

Compliance with Ethical Standards

The retrospective analysis of retinal images is approved by Moorfields Eye Hospital NHS Foundation Trust and covered by the Institutional Review Board and the UK Health Research Authority (HRA) Research Ethics Committee (REC) reference 22/WA/0049 "Eye2Gene: accelerating the diagnosis of inherited retinal diseases" (Integrated Research Application System (IRAS) project ID: 242050). The conducted research reported in this paper is in accordance with this approved IRB protocol and the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects.

Conflict of Interest Statement

The authors declare no competing interests.

Author Contributions

B.G., B.M., W.W. and N.P. contributed to the conception and design of the work. M.M, S.M.D, S.M, K.F., W.W. and N.P. contributed to the data acquisition and organization. B.G. and B.M. contributed to the technical implementation. M.M., S.A.K, T.A.C.G, M.D.V, S.S and K.B. provided the clinical inputs to the research. B.G. W.W. and B.M. contributed to the evaluation pipeline of this work. W.W. and N.P. provided suggestions on analysis framework. All authors contributed to data processing, drafting and revising of the manuscript.

Funding

N.P, W.W, K.B, S.A.K, S.M.D, M.S, K.F, S.M and M.M are funded by an National Institute for Health Research (NIHR) AI Award (AI_AWARD02488).

Acknowledgments

The views expressed are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

Data Availability

The dataset analysed in the current study are described on the protocol [13] and are not publicly available to protect patient privacy and intellectual property.