Comparison of Uncertainty Estimation Methods for Diabetic Retinopathy Classification using Deep Learning

Minor Research Project

MSc Medical Imaging

Sara Guillén Fernández-Micheltorena

Examination committee:

Supervisor: dr. Leticia Gallardo Estrella

Senior Deep Learning Engineer, Thirona B.v.

Daily Supervisor: Bálint Hompot

Deep Learning Engineer, Thirona B.v.

Examiner: dr. Alberto de Luca

Assistant professor, UMC Utrecht

Nijmegen, 3rd July 2023

Comparison of Uncertainty Estimation Methods for Diabetic Retinopathy Classification using Deep Learning

Sara Guillén Fernández-Micheltorena
Thirona B.v

***Abstract*—Deep learning models have shown potential in automated diabetic retinopathy classification, but they lack certainty in their predictions, which is crucial in clinical settings. Thus, uncertainty estimation is receiving increased attention in this field. This work compares uncertainty estimation methods for the classification of diabetic retinopathy in a 5-class scheme, including Evidential Deep Learning. To address the absence of ground truth for uncertainty, a novel evaluation framework is proposed. The framework utilizes a threshold-based system that assumes higher uncertainty for images from distributions other than the training distribution. It aims to distinguish between training distribution images and those from other distributions based on uncertainty estimates. Experiments evaluate the performance of the models in scenarios representing aleatoric and epistemic uncertainties. The results reveal the varying behavior of the methods based on the severity of the shift and the type of uncertainty. While ethnicity and disease shifts, as well as low-quality images, pose challenges as models confidently classify them, artificial noisy images and out-of-distribution samples are correctly identified as uncertain. Notably, Evidential Deep Learning demonstrates effective uncertainty modeling even in challenging scenarios. Overall, this work emphasizes the importance of uncertainty estimation for diabetic retinopathy classification, addresses limitations for its clinical applicability, and provides insights for future research in this domain.**

***Index Terms*—uncertainty estimation, diabetic retinopathy, deep learning, bayesian neural networks, evidential deep learning**

1. Introduction

The introduction of artificial intelligence has revolutionized many different fields including image classification [1], medical image segmentation [2], and natural language processing [3]. However, standard methods have shown a tendency towards overconfidence in their predictions [4], making their application in high-risk fields dangerous. This concern is particularly significant in the medical field, where ensuring the reliability of confident model predictions is crucial for automated screening processes. Consequently, there has been a surge in research aimed at exploring uncertainty estimation methods, which enable models to provide not only predictions but also certainty estimates that accurately reflect the confidence in their predictions.

In recent years, there has been increasing interest in the automatic classification of diabetic retinopathy (DR), an eye condition that can lead to vision loss and blindness, using deep neural networks [5]–[8]. More recently, researchers have turned their attention to studying uncertainty estimation techniques [9]–[11]. These studies have primarily centered around Bayesian Neural Networks (BNNs) and have been focused on classifying DR using binary classification schemes: "referable vs non-referable" (RDR) or "healthy vs any DR". However, clinically-oriented approaches have transitioned towards adopting the internationally proposed 5-class DR classification system [12]. The first attempt to estimate uncertainty in a multi-class classification scheme was reported in [11].

The evaluation process plays a critical role in assessing the effectiveness of uncertainty estimation techniques, primarily due to the absence of a definitive ground truth for uncertainty. There exists 3 typical evaluation techniques employed in the study of uncertainty estimation: ranking, calibration, and thresholding. The ranking approach evaluates uncertainty by observing the variation in error when removing predictions with the highest levels of uncertainty [13]. The calibration method assesses the reliability of the uncertainty estimates by examining how well they correspond to ground truth correctness likelihood [14]. On the other hand, the threshold-based method focuses on determining an appropriate threshold for uncertainty, beyond which predictions are referred to clinical specialists (refer to Figure 1). This evaluation method stands out as more clinically reliable in this context. By setting a threshold for uncertainty, the evaluation process ensures that predictions deemed highly uncertain are not solely relied upon and are instead subject to further scrutiny by experts. This approach enhances patient safety by reducing the risk of misdiagnosis or missed diagnoses.

In the aforementioned recent studies on uncertainty estimation for DR, the threshold-based method is commonly used. However, the primary concern with this evaluation process lies in its focus on whether accuracy improves by removing the most uncertain predictions, rather than determining whether the methods effectively model uncertainty. Additionally, these studies often establish the threshold for uncertainty using test set-specific thresholds, rather than employing a previously established threshold that can be generalized to other datasets.

This study aims to address the existing challenges in recent uncertainty estimation studies on DR while proposing a novel evaluation model. The key objectives and contributions of this research are outlined as follows:

- Adoption of a 5-class classification scheme: In this work,

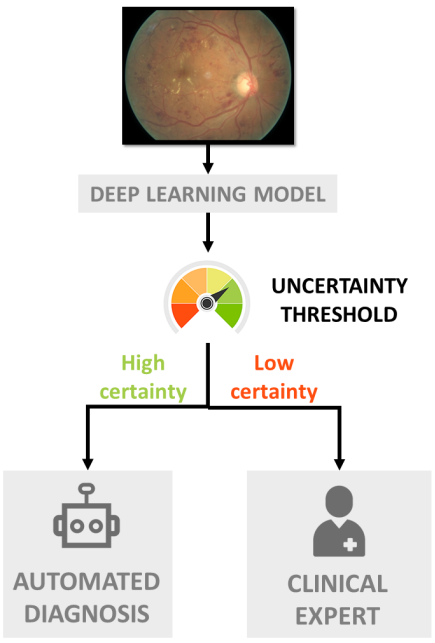


Fig. 1: Overview of the evaluation process simulating real-world clinical setting in which a sample is flagged as uncertain and referred to a clinical expert if the uncertainty value exceeds a predetermined threshold.

we utilize the internationally proposed PIRC scheme for the classification of DR. This approach allows for a more comprehensive and clinically relevant assessment of the disease.

- Comparison of uncertainty estimation methods: We compare various uncertainty estimation techniques, including some that have not been previously applied in the field of DR, such as Evidential Deep Learning. These selected methods effectively model different sources of uncertainty.
- Proposal of a new evaluation framework: We assess the performance of the uncertainty estimation methods by how well they can distinguish normal predictions (representing the training distribution) and corrupted predictions (representing a different distribution), assuming that high uncertainty would correspond to the latter. A threshold-based system is employed where predictions are flagged as uncertain if their uncertainty surpasses a threshold determined using the validation set.
- Generation of evaluation datasets with different uncertainty sources: To address different scenarios, evaluation datasets that exhibit diverse uncertainty types are generated. This ensures a robust evaluation of the uncertainty estimation methods and their applicability in real-world scenarios.

2. Theoretical Background

2.1. DR Classification

DR is a progressive eye disease that occurs as a complication of diabetes. It is characterized by damage to the blood vessels in the retina, leading to vision impairment if left untreated. The severity and progression of DR are typically classified using the PIRC (Proliferative, Ischemic, Retinopathy Classification) scheme, which classifies DR according to the severity into 5 classes (see Fig. 5): no DR (class 0), mild DR (class 1), moderate DR (class 2), severe DR (class 3), and proliferative DR (class 4). This classification system provides a standardized approach to assessing the condition based on specific features observed in fundus images.

Detecting DR poses a significant challenge due to the time, cost, and effort involved in manual diagnosis [15]. However, advancements in machine learning-based medical image analysis have demonstrated promising capabilities in evaluating retinal fundus images [16]. Specifically, the integration of deep learning algorithms has significantly contributed to the early detection of DR [17]. In [18], authors compare and analyze the recent state-of-the-art methods for the detection and classification of DR color fundus images using deep learning techniques, mainly already existing convolutional neural network (CNN) structures such as VGG, ResNet, or AlexNet. They also highlight the importance of using big datasets and the use of data augmentation to reduce overfitting in model training.

2.2. Uncertainty estimation in DR

The concept of uncertainty estimation has gained significant attention in the field of DR research, as it allows for the development of models that provide confidence values alongside their predictions. Recent studies have focused on comparing various methods based on BNNs, such as Mean Field Variational Inference (MFVI), Monte Carlo Dropout (MC Dropout), Radial BNN, Deep Ensembles, and ensembles of these techniques, in order to identify the most effective approach. Furthermore, a recently published study [10] sought to benchmark BNNs and ensembles of BNNs in task-specific situations. They also provided a user-friendly codebase[1] including several implementations.

The findings from these studies highlight the potential of Radial BNN and ensembling BNNS in modeling uncertainty [9]–[11]. These results are reported using a threshold-based evaluation process employed to assess uncertainty estimation in these works. The system acts as a referral process in which the most uncertain predictions are flagged for further examination. The idea is that the removal of the most uncertain samples increases the overall accuracy.

2.3. Types of uncertainty

There are two types of uncertainty described in the literature: epistemic and aleatoric [19]. Epistemic or model uncertainty is caused by the ignorance of the model on certain

[1] https://github.com/google/uncertainty-baselines

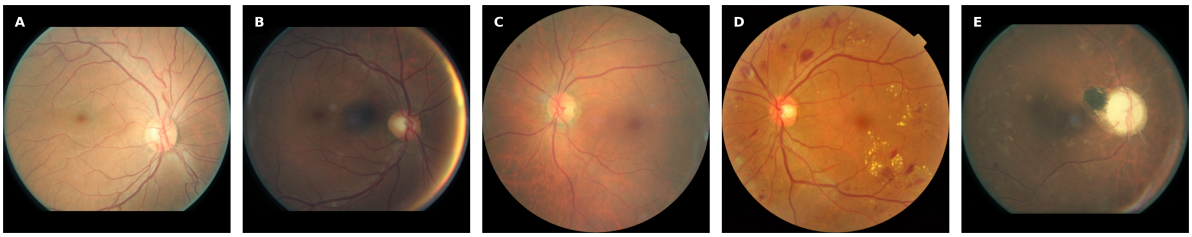


Fig. 2: Standard photographs from the patients with diabetic (A) No DR, (B) mild NPDR, (C) moderate NPDR, (D) severe NPDR, (E) PDR.

data regimes. Therefore, it can be explained away by adding more data covering unseen regimes to its training. For instance, in the context of DR, different ethnicity, or images containing another disease. On the other hand, aleatoric uncertainty is caused by variability and randomness in the data generation. This uncertainty is inherent to the data itself and thus, cannot be reduced. An example of aleatoric uncertainty can be a noisy or low-quality retina image. A representation of both types can be seen in Fig. 3.

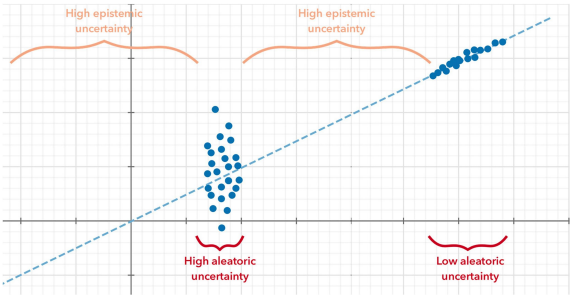


Fig. 3: The graph depicts the relationship between epistemic and aleatoric uncertainty. The dashed line represents an estimated function by a model, and the blue dots, the training data samples. Source: [20]

Aleatoric uncertainty can be further divided into two types: homoscedastic and heteroscedastic uncertainty [21]. Homoscedastic uncertainty is characterized by a constant level of variability in the errors of a model across the entire range of input data, whereas heteroscedastic uncertainty occurs when levels of variability in the errors differ across the input data range.

2.4. Uncertainty estimation

2.4.1. Bayesian Neural Network (BNN)

The objective of conventional neural networks is to classify a given data point into its corresponding class. This is achieved through a process of weight adjustment, where the network adapts its weights w to maximize the likelihood $p(D|w)$ of observing the data D. This optimization approach is known as Maximum Likelihood Estimation (MLE), as it aims to maximize the probability of the observed data D given the weights w. However, MLE does not provide a direct means to calculate the probability of the prediction being correct. To address this, it becomes necessary to determine the posterior distribution of the weights $p(w|D)$. This is where BNNs prove advantageous, as they enable the approximation of this distribution.

BNNs represent the weights and biases of neural networks as probability distributions [22]. The posterior probability of the model parameters given the observed data can be computed using Bayes' theorem, which is expressed as:

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)} = \frac{P(D|w)P(w)}{\int_w P(D|w)dw} \tag{1}$$

In practice, calculating the posterior distribution of the model parameters is intractable. Therefore, approximation techniques are used to estimate the posterior probability such as Markov chain Monte Carlo (MCMC) sampling or Varinal Inference (VI). As MCMC is computationally expensive, VI techniques are often employed which include Gaussian, or Radial-Gaussian. The goal is to make the approximate posterior distribution $q(w|\theta)$ as close as possible to the true Bayesian posterior $p(w|D)$. The parameters θ are optimized to minimize the Kullback-Leibler (KL) divergence [23] between the two distributions. In Gaussian VI or MFVI, the prior and posterior distributions are assumed to be multivariate standard normal distributions.

These methods can inexpensively estimate the posterior predictive distribution using Monte Carlo sampling and the uncertainty associated with the predictions can be quantified by computing its variance.

The training of BNNs is computationally more expensive as the models occupy more space than regular models.

2.4.2. Radial BNN

$$w_k = \mu_k + \sigma_k \times \epsilon \tag{2}$$

The multivariate normal distribution exhibits a phenomenon known as 'soap bubble' pathology in high-dimensional spaces, meaning that most of the probability mass is concentrated on a thin shell far from the mean. Consequently, samples drawn from this distribution show high norms. In [24], it is suggested that these high norms pose a challenge when training deep neural networks using the MFVI approach with Gaussian posteriors. To tackle this issue, a novel posterior distribution called the Radial BNN is proposed. This distribution is designed to ensure that the expected norm of the samples matches that of a univariate standard normal distribution, irrespective of the dimensionality. The sampling process for a single weight vector is defined as follows:

$$w_k = \mu_k + \sigma_k \times \frac{\epsilon}{||\epsilon||} \times r \tag{3}$$

The resulting random variable w avoids the soap bubble pathology. However, it has no closed-form probability density function or KL-divergence. The authors noted that an estimation of the KL-divergence can be computed stochastically, as shown in [25]. This enables the optimization of the ELBO up to a constant. The prior chosen for this purpose is a multivariate standard normal distribution as for MFVI.

2.4.3. MC Dropout

MC Dropout, introduced in [26], is a technique utilized to estimate uncertainty in deep learning models that have been trained using dropout regularization. The method consists of activating dropout during inference mode. During dropout, a binary mask is created from a Bernoulli distribution with probability p which is then applied to the activations. Activating dropout during test time allows the model to function as a simplified form of BNN where the dropout distribution is the approximate posterior distribution. The posterior predictive distribution can be then computed using MC sampling. MC Dropout presents a simple and computationally efficient approach for uncertainty estimation, which can be easily integrated into existing deep learning pipelines.

2.4.4. Deep Ensemble

Deep Ensembles consist of collections of multiple neural networks [27]. These models are individually trained using data shuffling and random initialization techniques to ensure variability among the models. This way the models converge toward a distinct local optimum. This is due to the highly non-convex nature of the neural network loss surface. After training, the uncertainty can be calculated by combining the predictions of these models for a given input data point. As suggested in several works [28], using an ensemble of N = 5 is sufficiently large to achieve favorable results.

This technique is simple to implement, but it requires parallel training of the models which is computationally and timely expensive.

2.4.5. Evidential Deep Learning (EDL)

Evidential deep learning (EDL) [29] seeks to model uncertainty using the framework of the Theory of Evidence. In this framework, uncertainty is expressed as the degree of belief or evidence associated with different outcomes. Subjective Logic (SL) provides a formalization of the belief assignments by representing them as a Dirichlet Distribution. This utilization of the Dirichlet Distribution enables the application of evidential theory principles to accurately quantify belief masses and uncertainty within a clearly defined theoretical framework. Unlike point estimates provided by traditional deep learning models, these probability distributions capture the range of possible outcomes and their associated likelihoods.

SL considers a frame of K class labels by assigning a belief mass b_k to each class and an uncertainty mass of u. This extra mass value is the belief that the truth can be any class label, the "I do not know class". These $K + 1$ belief masses are non-negative and sum to 1: $u + \sum_k^K b_k = 1$. Evidence (e_k) is the amount of support collected from data for a sample to be classified into a certain class and i is related to the Dirichlet parameter α_k by $\alpha_k = e_k + 1$. The belief masses can then be derived from these parameters using $b_k = (\alpha_k - e_k)/S$ where $S = \sum_k^K \alpha_k$ is the Dirichlet strength. Therefore, we can calculate the corresponding uncertainty estimate of a sample.

The method described in this work has not been considered in previous studies on uncertainty estimation in DR classification and will be incorporated into our research. Implementation of this method is straightforward, involving a simple modification of the loss function.

2.4.6. Auxiliary Output

Aleatoric uncertainty can be predicted through a method described in [21]. This method consists of adjusting an observation noise parameter, which can be perceived as corrupting the model with a Gaussian distribution. In the case of homoscedastic uncertainty, this noise parameter is a scalar number. On the contrary, for modeling heteroscedastic uncertainty, the noise parameter is tuned, as it depends on the data.

The network is trained to output the parameters of the Gaussian distribution: the mean μ and standard deviation σ. Subsequently, the logits are approximated by sampling from this distribution. The aleatoric uncertainty can be attributed to the noise parameter (σ) associated with the input sample.

The implementation of this method is straightforward and does not require any additional time, as the sampling process is performed after a single pass for obtaining the distribution parameters.

3. Methods

3.1. Baseline

The base model architecture used was the EfficientNet [1]. It consists of a scaled-up neural network, where the scale is implemented in all dimensions (width, depth, and resolution) with a fixed ratio, a method known as compound scaling [30]. The EfficientNet architecture includes 7 MBConv or inverted residual blocks before outputting the feature map, see Fig. 4. The 8 models of EfficientNet (B0 - B7) share common blocks with subtle complexities in their architectures.

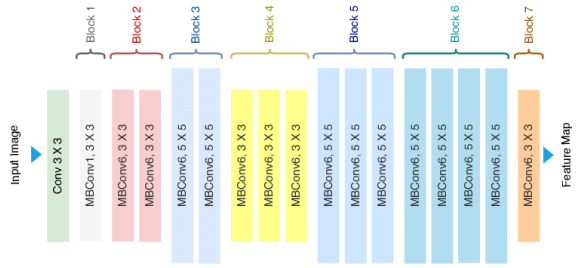


Fig. 4: Architecture of the EfficientNet with MBConv as the basic building block. Source: [31]

The chosen model was the EfficientNetB4 and it was trained using the categorical cross-entropy loss for the multi-label classification task addressed in this work.

The baseline method does not include an uncertainty estimation method, but the uncertainty (u) is calculated using the Softmax output by obtaining the missing probability of the predicted class:

$$u_i = 1 - argmax(f(x_i)) \tag{4}$$

3.2. Deep Ensemble

The Deep Ensemble method is applied by training N separate baseline models with the same architecture described in the previous section. The prediction ($\hat{y}$) is the mean of the outputs of the N models and the uncertainty (u) is calculated as the variance of these outputs.

$$\hat{y}_i = \frac{1}{N}\sum_{n}^{N} f_n(x_i) \tag{5}$$

$$u_i = \frac{1}{N}\sum_{n}^{N} (f_n(x_i) - \hat{y}_i)^2 \tag{6}$$

where each f_n corresponds to a different model. In this work, $N = 5$ models are used.

3.3. MC Dropout

MC Dropout is implemented in an already trained model by activating the dropout during inference that can be easily manipulated by changing the *training* argument in the dropout layers. The prediction is obtained by sampling through the posterior distribution. In practice, this is done by averaging the outputs of T passes through the model, similar to the Deep Ensemble method. However, in this case, we use one single model as each pass will result in a different output. In this work, $T = 25$ passes were used.

$$\hat{y}_i = \frac{1}{T}\sum_{t}^{T} f(x_i) \tag{7}$$

$$u_i = \frac{1}{T}\sum_{t}^{T} (f(x_i) - \hat{y}_i)^2 \tag{8}$$

3.4. Auxiliary Output

In order to apply the Auxiliary Output method, the single output layer of the neural network is changed into two output layers, each corresponding to the parameters μ and σ of the distribution. Then, S logit samples are obtained to generate a final output which is then passed through the Softmax activation function.

$$x_{i,s} = \mu + eps_s * \sigma \tag{9}$$
$$eps \sim N(0, I)$$

$$\hat{y}_i = \frac{1}{S}\sum_{s}^{S} f(x_{i,s}) \tag{10}$$

In this case, the uncertainty calculation is straightforward as σ corresponds to the aleatoric uncertainty of the input sample.

3.5. Evidential Deep Learning (EDL)

The implementation of the EDL method requires a small change in the model. In this method, the model outputs a distribution of categorical probabilities (Dirichlet distribution). In practice, the Softmax layer is replaced by an exponential function which ensures that the class parameters are non-negative. The expected class probabilities ($\hat{p}_k$) are obtained using the calculated Dirichlet strength (S) $\hat{p}_k = \frac{\alpha_k}{S}$. The prediction is calculated by obtaining the highest expected probability:

$$\hat{y} = argmax(\hat{p}_k) \tag{11}$$

The loss function is chosen to be the modified mean squared error that the authors determined was best suitable (see Eq. 5 in [29]):

$$Loss_{EDL} = \sum_{k}^{K}(\hat{y} - \frac{\alpha}{S})^2 + \frac{\alpha(S-\alpha)}{S^2(S+1)} \tag{12}$$

The Dirichlet parameters can be seen as the beliefs of each class and are calculated using S by $b_k = \frac{\alpha_k - 1}{S}$. The belief not assigned to any of the K classes can be interpreted as the epistemic uncertainty:

$$u = 1 - \sum_{k}^{K} b_k = \frac{K}{S} \tag{13}$$

3.6. Radial BNN

In radial BNN, the parameters are no longer point estimates but distributions having parameters μ and σ. In practice, the main modification needed for the model is to change the dense and convolutional layers to accept weight distributions, specifically, the implemented radial distribution as posterior and multivariate normal as prior.

During inference, the prediction is obtained by sampling from the predictive distribution using T forward passes of the input sample, like the MC Dropout method (Eq. 7 and 8).

Unfortunately, due to limited computational resources, this implementation was not utilized in the final analysis of this paper.

4. Experiments

4.1. Datasets and processing

In this section, a summary of each of the datasets employed in the training and evaluation of the models is described. More details regarding the class distribution, total images, and uncertainty type related to each dataset are depicted in Table I in the Appendix.

EyePACS. The EyePACS dataset, previously utilized for the Kaggle Diabetic Retinopathy Detection Challenge [32], was selected as the training dataset for this study. This dataset comprises 88,702 high-resolution RGB retina images, which exhibit varying degrees of DR according to the PIRC scale. Each patient is represented by two images (right and left eyes). The images were then divided into three subsets: 66,562 images for training, 8,285 images for validation, and 8,285 images for testing. Particular attention was given to ensuring that both images from the same patient were placed in the same subset.

Low-quality EyePACS. A set of 5,400 low-quality images, following the annotations presented in [33], was excluded from the total training and validation sets, forming the Low-quality EyePACS dataset.

Noisy EyePACS. To test the model against *noisy images*, another was created by manipulating the 8,285 images from the EyePACS test set, adding Gaussian noise.

Blurry EyePACS. Similarly, this dataset was formed by manipulating the EyePACS tes set, adding Gaussian blur, to test the model against *blurred images*.

One channel EyePACS. This other dataset was created by setting the green and blue channels to 0 of the EyePACS testset images, to test the model against *missing information*.

IDRID. In order to create a task that effectively measures model performance in the presence of *ethnicity shift*, the Indian Diabetic Retinopathy Image Dataset (IDRID) dataset was used. The IDRID dataset [34] contains 516 fundus images with information regarding the disease severity level of DR graded by medical experts according to the PIRC convention.

Glaucoma (ODIR). The Ocular Disease Intelligent Recognition (ODIR) dataset was used to evaluate the model in the presence of *another eye disease*. The ODIR dataset [35] contains color fundus photographs from 5,000 patients from different hospitals in China. Each image includes diagnostic keywords from doctors which were labeled by trained human readers into eight labels: Normal (N), Diabetes (D), Glaucoma (G), Cataract (C), Age-related Macular Degeneration (AMD) (A), Hypertension (H), Pathological Myopia (M), Other diseases/abnormalities (O). A dataset containing images of glaucoma disease was derived from the ODIR dataset. The classification of images within this dataset was based on the PIRC scheme. Specifically, images that exhibited glaucoma but lacked DR were assigned to class 0 while images depicting both glaucoma and varying degrees of DR severity were categorized into classes 1 to 4, in accordance with the severity level described in the provided diagnostic keywords.

AMD (ODIR). Another dataset containing images of AMD was derived from the ODIR dataset to check the model against a *different eye disease*. The creation of this dataset followed the same procedure as the glaucoma dataset.

Tsukazaki. The Tsukazaki Optos Public (TOP) dataset [36] was another dataset employed to study the effect of images captured with a *different camera*. This dataset encompasses 13,047 ultra-widefield (UWF) retina images obtained from 8,588 eyes of 5,389 patients in Japan. Each image has binary labels for eight retinal diseases. However, only the DR binary labels were used. To ensure consistency in image size, the UWF retina images from the TOP dataset were subjected to centered cropping, resulting in a standardized resolution of 512x512 pixels.

ImageNet. The final evaluation dataset utilized in this study was a subset of the ImageNet dataset to test the model against *out-of-distribution (OOD) samples*. The complete ImageNet dataset comprises more than 1 million images covering a wide range of 1,000 object classes [37]. However, only a smaller subset of images was selected where all images were labeled as class 0. To ensure uniformity in image resolution, the images were rescaled to a standardized size of 512x512 pixels.

4.2. Evaluation

4.2.1. Model performance

The performance of the DR classification models is assessed using the accuracy metric which measures the fraction of correctly classified samples.

4.2.2. Proposed threshold-base evaluation system

The proposed evaluation system is based on a referral process (see Fig. 1) that emulates a real clinical scenario, similar to the evaluation techniques implemented in other uncertainty estimation studies. In this work, a new approach is utilized in order to address the lack of ground truth for uncertainty. The methods are evaluated on mixed datasets containing normal samples (representing the training distribution) and corrupted samples (representing other distributions). The objective is to distinguish between normal and corrupted samples based on their uncertainty, assuming that corrupted samples present high uncertainty in their predictions, using a predetermined uncertainty threshold where certain predictions are retained and uncertain predictions are flagged. In other words, it is a binary classifier where true positives (TP) refer to certain normal predictions, while true negatives (TN) correspond to correctly flagged corrupted predictions as uncertain.

The mixed datasets are generated as follows: 50% of the dataset comprises normal samples from the training distribution (EyePACS test set) and the remaining 50% corresponds to corrupted samples from other distributions, corresponding to the testing datasets described in the previous section.

When determining the uncertainty threshold, it is important to note that the uncertainty threshold values are derived from the validation dataset. This approach differs from other studies where the thresholds were implemented directly in the

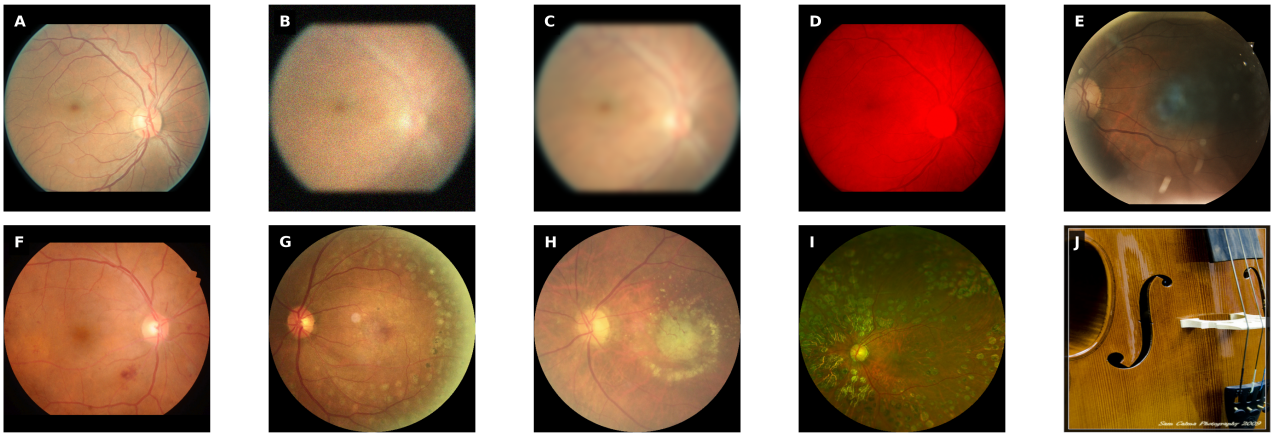


Fig. 5: Examples of images from each dataset. (A) EyePACS, (B) noisy EyePACS, (C) blurred EyePACS, (D) one-channel EyePACS, (E) low-quality EyePACS, (F) IDRID, (G) Glaucoma (ODIR), (H) AMD (ODIR), (I) Tsukazaki, (J) ImageNet.

evaluation dataset. This allows for greater flexibility and generalizability in assessing uncertainty across different scenarios and datasets. The process unfolds in the following manner: uncertainty estimates of the validation set are sorted, and thresholds are established for varying percentages of certain and uncertain samples. These percentages range from 5% to 95% at 5% intervals. For instance, by considering 5% of the uncertainty estimates from the validation predictions as certain, we can identify the specific uncertainty value that distinguishes them from the uncertain samples. These thresholds are saved and applied to every evaluation dataset.

In real-life situations, the main objective is to prioritize the identification of normal data predictions as trustworthy outcomes, while ideally flagging all corrupted data points. Hence, it is crucial to determine the number of normal samples correctly identified as certain, i.e., the precision of the system:

$$precision = \frac{TP}{TP + FP} \tag{14}$$

5. Results

In this section, the results of the proposed evaluation in each dataset are shown. The results are summarized into 2 plots:

- Plot 1: the percentage of certain predictions of the combined dataset for each of the thresholds obtained using the validation set and their corresponding accuracy. This plot represents the performance of the DR classification in the retained predictions.
- Plot 2: the number of certain predictions and the percentage of these predictions being non-corrupted. This plot represents the precision of the evaluation system with respect to the ideal scenario which is denoted with a dashed line. In the ideal scenario, only normal samples are retained as certain predictions, i.e. the precision is 1 for all thresholds, and the limit in the number of certain predictions is the number of normal samples in the testset. It should be noted that the ideal scenario assumes that corrupted predictions present high uncertainty and thus, would ideally be flagged.

EyePACS testset. The testset comprising data from the training distribution reveals a correspondence between the percentage of specific predictions and the uncertainty estimate threshold obtained from the validation set (see Fig. 6. The accuracy of the models for the predictions flagged as certain at a threshold of 95% show the overall performance of the models. The Baseline (0.883), Deep Ensemble (0.879), Auxiliary Output (0.867) and EDL (0.874) models achieve similar classification performance while the MC dropout predictive performance is lower (0.746).

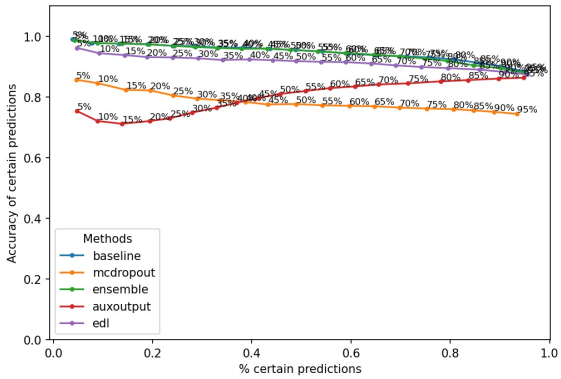
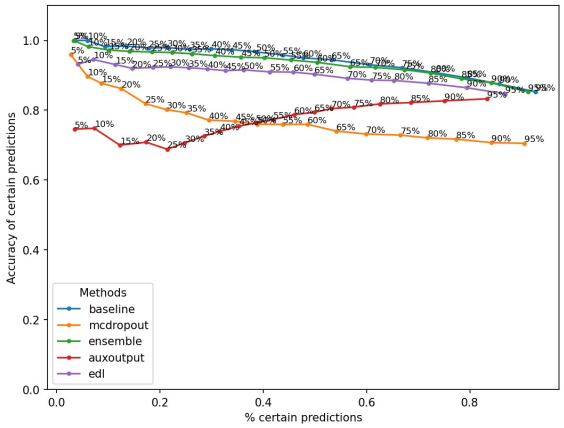


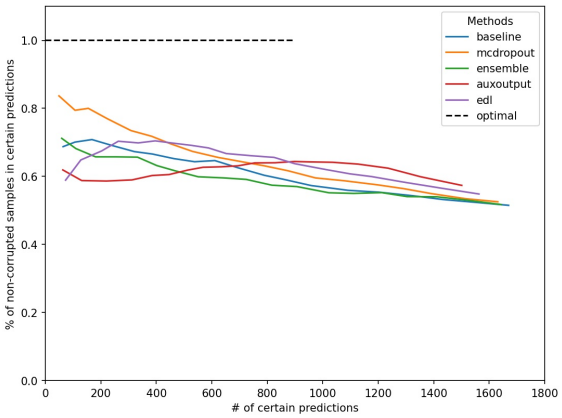
Fig. 6: Accuracy of certain predictions for each threshold of the EyePACS test dataset.

Low-quality EyePACS.The results for the dataset containing low-quality images from the EyePACS dataset are shown in Fig 7. Plot 1 shows a similar behavior as for the original testset (see Fig. 6). Plot 2 shows that the precision of the methods ranges from 0.6 to 0.7 in all cases, being EDL, MC Dropout, and Auxiliary Output the ones with better results.

Noisy EyePACS. In Fig. 8, the results obtained for the dataset containing noisy images are shown. An almost-perfect performance can be observed in Plot 2 for EDL and Auxiliary

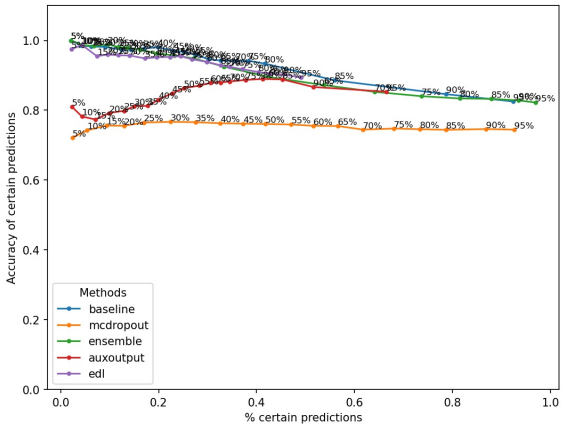


(a) Plot 1: accuracy of retained samples vs percentage of certain predictions for each threshold.

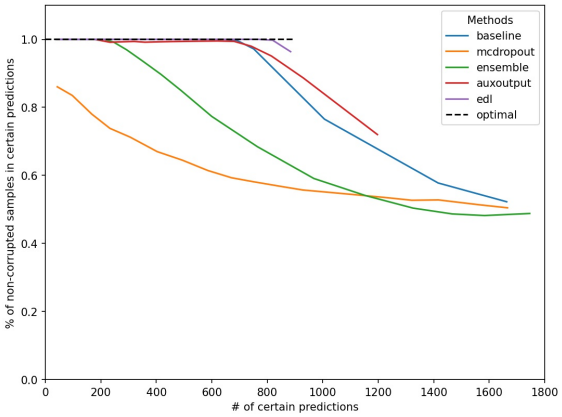


(b) Plot 2: number of certain predictions and the percentage of non-corrupted samples in these predictions.

Fig. 7: Results for the low-quality EyePACS dataset.



(a) Plot 1: accuracy of retained samples vs percentage of certain predictions for each threshold.



(b) Plot 2: number of certain predictions and the percentage of non-corrupted samples in these predictions.

Fig. 8: Results for the noisy EyePACS dataset.

Output methods, which are able to flag most corrupted predictions as uncertain, achieving high accuracy in the retained ones (see Plot 1).

Blurred EyePACS. Fig. 9 depicts the results obtained for the dataset containing blurred images. The first plot shows a similar behavior as the original test set. In terms of distinguishing the normal from the corrupted samples, EDL and Auxiliary Output achieve the best results reaching a precision of 0.9.

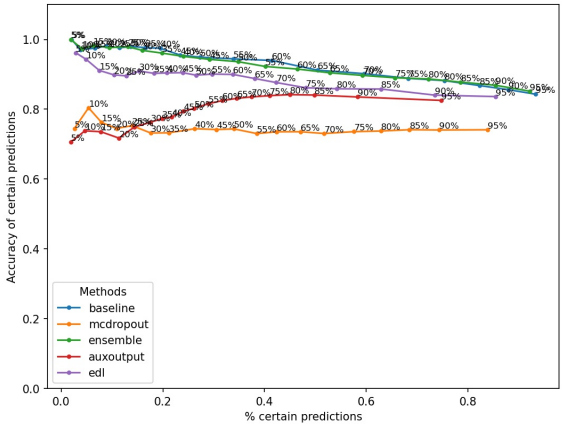
One-channel EyePACS. The results for the dataset with images containing only red channel information show that for MC Dropout, EDL and Auxiliary Output, the number of predictions flagged as uncertain is above 40% (see Plot 1 in Fig. 10). The precision of the system for these 3 methods is almost perfect (see Plot 2 in Fig. 10).

IDRID. The IDRID combined dataset shows a similar behavior as the EyePACS testset (left plot in Fig. 11). The right plot represents the precision of the system where the EDL method obtains better results.

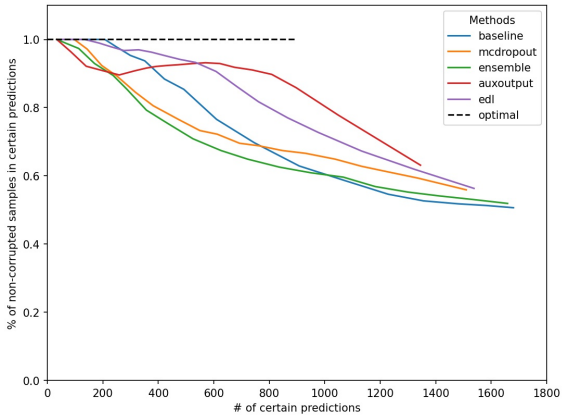
Glaucoma and AMD (ODIR). The summary plots for the results of the dataset containing glaucoma and AMD images can be observed in Figs. 12 and 13, respectively. In both cases, the performance of the models is similar to the EyePACS testset except for a visible drop in the accuracy of Auxiliary Output for the first thresholds. Regarding the performance of the evaluation system, EDL shows a higher precision in both scenarios.

Tsukazaki. The Tsukazaki-containing dataset has an overall lower classification accuracy (see Plot 1 in Fig. 14). Regarding the precision of the system, despite MC Dropout showing higher precision, EDL is the method that flags most of the cases as uncertain for smaller thresholds (see Plots 1 and 2 in Fig. 14)

ImageNet. The results for the dataset with ImageNet images are summarized in 15. The plots show an overall good accuracy when classifying images with all methods. According to

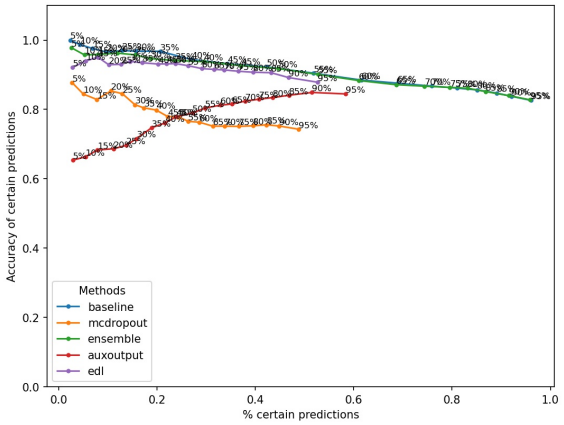


(a) Plot 1: accuracy of retained samples vs percentage of certain predictions for each threshold.

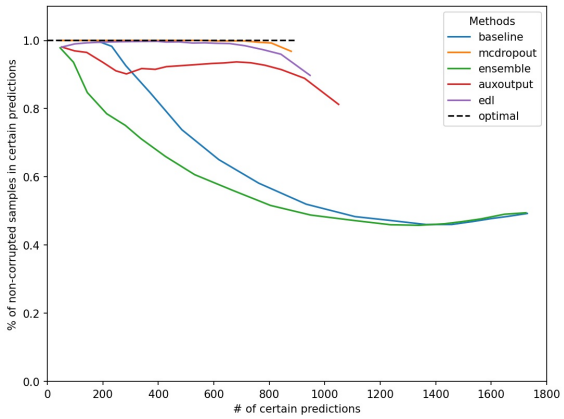


(b) Plot 2: number of certain predictions and the percentage of non-corrupted samples in these predictions.

Fig. 9: Results for the blurred EyePACS dataset.

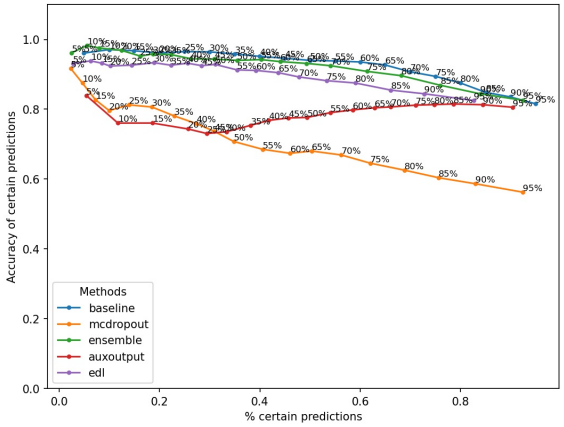


(a) Plot 1: accuracy of retained samples vs percentage of certain predictions for each threshold.

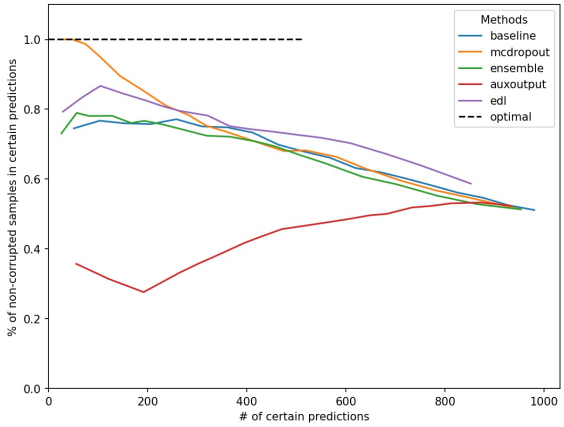


(b) Plot 2: number of certain predictions and the percentage of non-corrupted samples in these predictions.

Fig. 10: Results for EyePACS containing only red channel information.

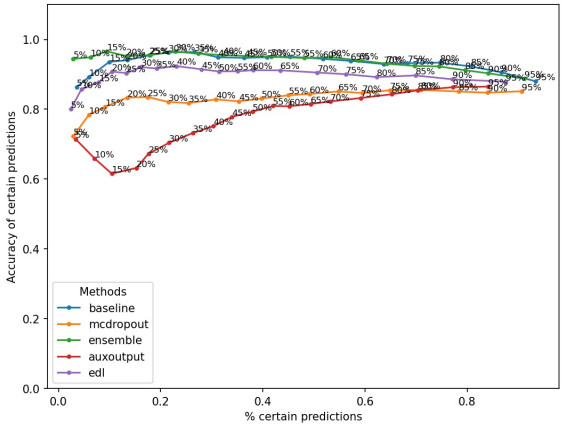


(a) Plot 1: accuracy of retained samples vs percentage of certain predictions for each threshold.

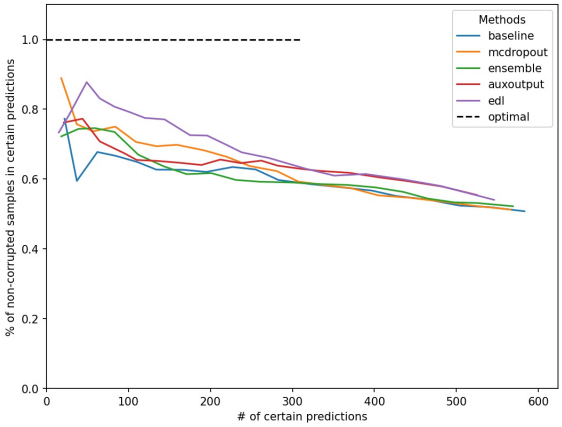


(b) Plot 2: number of certain predictions and the percentage of non-corrupted samples in these predictions.

Fig. 11: Results for IDRID dataset.

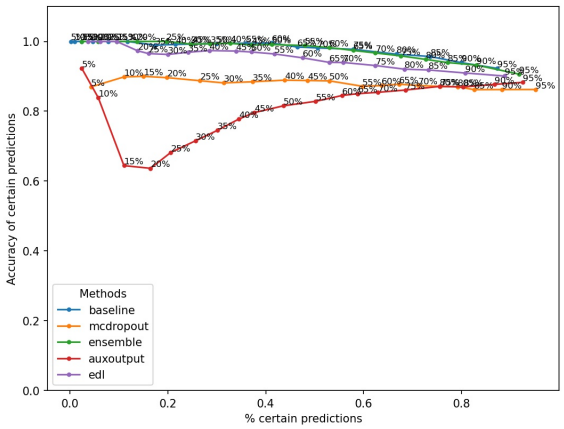


(a) Plot 1: accuracy of retained samples vs percentage of certain predictions for each threshold.

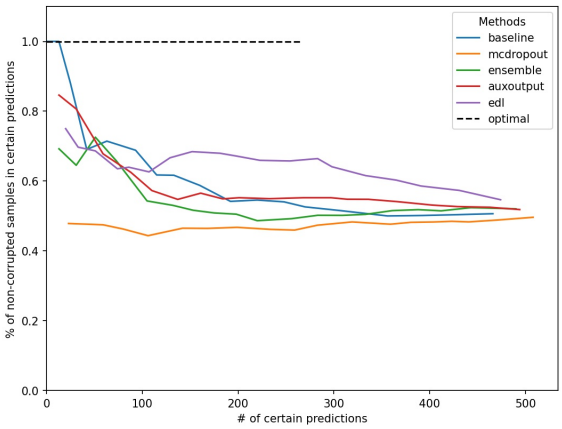


(b) Plot 2: number of certain predictions and the percentage of non-corrupted samples in these predictions.

Fig. 12: Results for Glaucoma dataset.

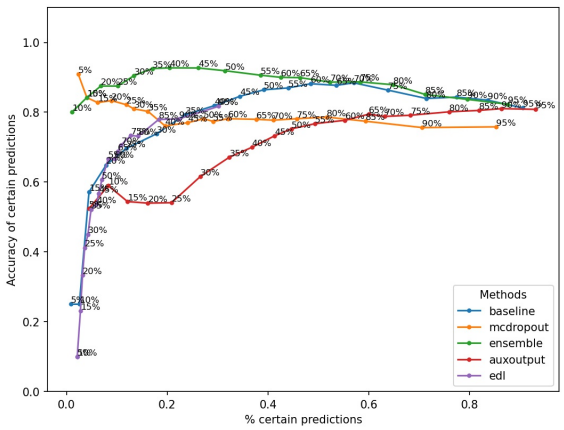


(a) Plot 1: accuracy of retained samples vs percentage of certain predictions for each threshold.

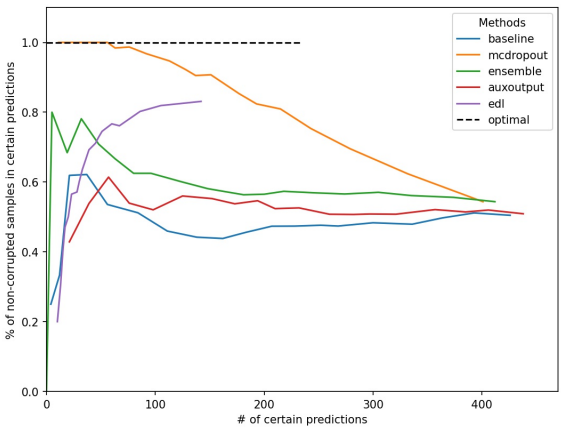


(b) Plot 2: number of certain predictions and the percentage of non-corrupted samples in these predictions.

Fig. 13: Results for AMD dataset.

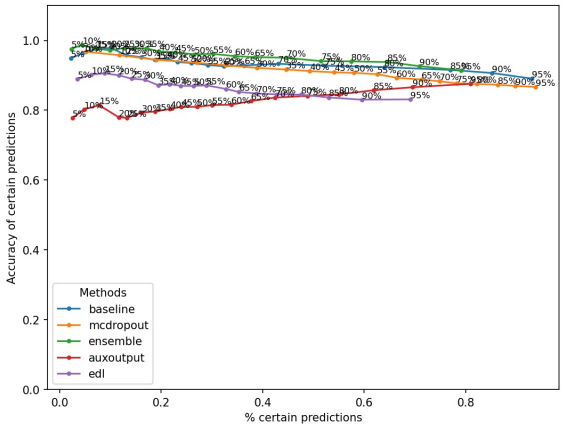


(a) Plot 1: accuracy of retained samples vs percentage of certain predictions for each threshold.

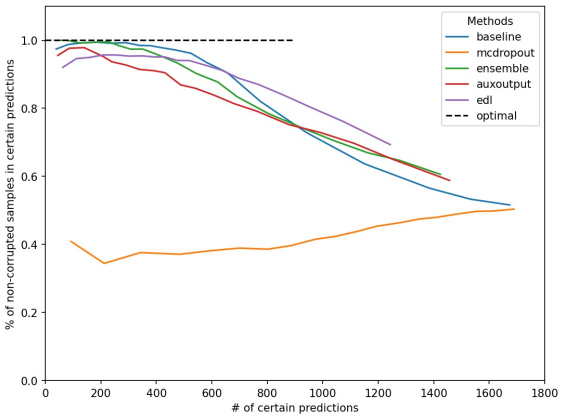


(b) Plot 2: number of certain predictions and the percentage of non-corrupted samples in these predictions.

Fig. 14: Results for Tsukazaki dataset.



(a) Plot 1: accuracy of retained samples vs percentage of certain predictions for each threshold.



(b) Plot 2: number of certain predictions and the percentage of non-corrupted samples in these predictions.

Fig. 15: Results for ImageNet dataset.

Plot 2, EDL is the method that flags more images as uncertain with higher precision.

6. Discussion

6.1. Model performance

In this work, five uncertainty estimation methods are implemented for comparison: Baseline, Deep Ensemble, MC Dropout, EDL, and Auxiliary Output. These models achieve similar accuracies on the EyePACS testset except for the MC Dropout model, which presents a noteworthy lower accuracy. This is an important finding and suggests that enabling dropout during inference has a negative impact on the model's predictive performance. The reason behind this effect lies in the disruption of batch normalization behavior when dropout is activated during inference [38]. Activating dropout during inference leads to the deactivation of certain units, resulting in a different distribution of activations. Consequently, batch normalization, which assumes all units to be active, normalizes the activations differently. The inconsistency between the statistics calculated during training and the activations during inference contributes to the decrease in accuracy. Using batch normalization in inference mode increased the prediction performance (refer to Fig. 16 in the Appendix for comparison) as the inputs are normalized in the current batch with dropped activations. However, the uncertainty estimation performance decreased when using batch normalization. Given our focus on uncertainty modeling and the real-world scenario of single input inference, batch normalization was set as the default approach.

Additionally, it is worth noting the peculiar behavior of the Auxiliary Output model in terms of predictive performance. Unlike the other models, the accuracy of the Auxiliary Output increases when increasing the uncertainty threshold.

Another observation to emphasize is that the Deep Ensemble method does not serve as an effective uncertainty estimation approach in this context. One potential explanation could be a low diversity among the models employed in the Deep Ensemble. However, further investigation is warranted to ascertain the underlying issue.

6.2. Aleatoric vs epistemic uncertainty

The evaluated uncertainty estimation methods were assessed on datasets containing images reflecting uncertainties from different sources. The datasets containing images exhibiting aleatoric uncertainty are the following: low-quality, noisy, blurred, and one-channel datasets. In these scenarios, the EDL and Auxiliary Output methods outperformed the other techniques when modeling uncertainty. The remaining datasets represent epistemic uncertainty: IDRiD (ethnicity shift), glaucoma and AMD (another disease), Tsukazaki (another camera), and ImageNet (random images). EDL tends to have better precision when retaining the normal predictions as certain.

The strong performance of the Auxiliary Output method is in line with its intended purpose of modeling aleatoric uncertainty. On the other hand, EDL has demonstrated efficiency in modeling both aleatoric and epistemic uncertainties.

6.3. Realistic vs fake scenarios

An interesting finding was the different behavior of the models when tested against realistic or fake scenarios. In the aleatoric uncertainty datasets, distinguishing the corrupted samples in manually manipulated datasets (noisy, blurry, and one-channel datasets) tended to be relatively easier compared to identifying specifically determined low-quality images, especially for noisy and one-channel images, as they exhibit more noticeable differences. It is worth noting that the classification performance in low-quality images was similar to that achieved by the original EyePACS testset, indicating that the models can confidently classify DR in this scenario. Consequently, distinguishing the corrupted samples from the normal ones becomes challenging.

6.4. Dataset shift vs OOD

IDRID, glaucoma, and AMD show dataset shifts while Tsukazaki and ImageNet contain OOD samples. In the case of dataset shifts, the models achieved great predictive performance, similar to the situation observed in the low-quality dataset. This means that classifying these images correctly with high confidence is relatively easy. Therefore, it is challenging to use uncertainty estimates to distinguish between normal and corrupted samples. This discovery challenges the prevailing belief that classifying DR in samples from a different ethnic distribution, such as IDRID, is a difficult task [39]. It is worth noting that EDL was efficient in modeling uncertainty even in these challenging scenarios.

On the contrary, when dealing with OOD images, the uncertainty estimates seemed useful in flagging the corrupted predictions, especially for the Tsukazaki dataset, where the EDL method considers more than half of the predictions uncertain.

6.5. Proposed evaluation system

The evaluation of uncertainty estimation methods plays a critical role in this study due to the absence of ground truth for uncertainty. The evaluation proposed in this study adopts a referral process similar to previous works but with a unique approach. We use combined datasets that encompass both normal data (representing the training distribution) and corrupted data (representing other distributions). The underlying assumption is that normal samples should exhibit high certainty, while corrupted samples should exhibit higher levels of uncertainty. However, a corrupted sample can still be correctly classified with high certainty. This has been observed in the low-quality, IDRID, glaucoma, and AMD cases, where using uncertainty estimates to flag uncertain predictions fails as the models are certain about their predictions.

An advantageous aspect of this effect is the ability to discern the types of images or distribution shifts that carry less significance in the predictive performance, as the models can accurately classify them. Therefore, we can understand how the models handle different distribution shifts.

Lastly, another important limitation is that using a distinct evaluation process prevents direct comparisons with other studies.

6.6. Clinical implications and future work

The next steps in the uncertainty estimation research applied to DR would involve including more uncertainty estimation methods in the comparison such as radial BNNs, which could not be implemented in this study. Furthermore, combinations of the existing methods could be explored to model both aleatoric and epistemic uncertainty, such as combining MC Dropout and Auxiliary Output.

One important clinical implication is the correct identification of OOD samples or fundus images with missing information, as uncertain. This indicates that the models perform well in extreme scenarios, allowing for the effective removal of these images.

The utilization of the PIRC classification scheme is clinically relevant as it offers the ability to analyze the uncertainty based on the predicted class. This option could be further leveraged by analyzing the rest of the classes which would give an additional insight regarding the input image and its class prediction. One potential direction to explore would be to use the belief theory employed in EDL to identify the amount of belief assigned to the other classes representing different degrees of DR.

For the system to be clinically applicable, predictive accuracy must be considered. While this study primarily emphasizes the performance of uncertainty estimation rather than achieving the highest classification accuracy, it is essential to ensure that the predictive accuracy of the uncertainty estimation model closely aligns with the base model that is already being used.

Furthermore, the suggested evaluation framework allows users to meticulously select an optimal threshold aligned with their desired model performance. This decision also influences the proportion of flagged predictions, a crucial factor to consider given the availability of medical experts responsible for assessing these uncertain predictions.

7. Conclusion

This study addressed the challenges in uncertainty estimation works on DR by adopting a 5-class classification scheme, comparing various uncertainty estimation methods, and proposing a threshold-based evaluation model using combined datasets with different uncertainties.

The adoption of the 5-class PIRC scheme allowed for a comprehensive assessment of DR, enhancing the clinical relevance of the classification process. The comparison of uncertainty estimation methods in the varied test datasets provided insights into their performance in modeling different scenarios. The proposed evaluation process has proven to be a more clinically relevant approach to assessing the effectiveness of uncertainty estimation methods in accurately capturing uncertainty. The study has shown that particular consideration should be given to situations where images from distribution shifts are classified correctly with a high level of certainty, such as the case of IDRID or low-quality datasets, invalidating the assumption that these images should have uncertain predictions.

Important findings include the outstanding performance of auxiliary output when modeling aleatoric uncertainty, and EDL in epistemic uncertainty scenarios, especially in challenging ones. Additionally, predictions for OOD samples and fake images are correctly flagged as uncertain which has important clinical implications.

In conclusion, this study contributes to the advancement of uncertainty estimation studies in DR by addressing the existing challenges and providing a new enhanced evaluation framework. Future research directions include exploring additional uncertainty estimation methods, combining existing techniques, and further analysis for its clinical applicability.

References

[1] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020.

[2] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241, (available on arXiv:1505.04597 [cs.CV]). [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[4] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1321–1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[5] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0161642016317742

[6] Z. Gao, J. Li, J. Guo, Y. Chen, Z. Yi, and J. Zhong, "Diagnosis of diabetic retinopathy using deep neural networks," *IEEE Access*, vol. 7, pp. 3360–3370, 2019.

[7] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic retinopathy fundus image classification and lesions localization system using deep learning," *Sensors*, vol. 21, no. 11, p. 3704, May 2021. [Online]. Available: https://doi.org/10.3390/s21113704

[8] S. Gayathri, V. P. Gopi, and P. Palanisamy, "Diabetic retinopathy classification based on multipath CNN and machine learning classifiers," *Physical and Engineering Sciences in Medicine*, vol. 44, no. 3, pp. 639–653, May 2021. [Online]. Available: https://doi.org/10.1007/s13246-021-01012-3

[9] A. Filos, S. Farquhar, A. N. Gomez, T. G. J. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. de Kroon, and Y. Gal, "A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks," 2019. [Online]. Available: https://arxiv.org/abs/1912.10481

[10] N. Band, T. G. J. Rudner, Q. Feng, A. Filos, Z. Nado, M. W. Dusenberry, G. Jerfel, D. Tran, and Y. Gal, "Benchmarking bayesian deep learning on diabetic retinopathy detection tasks," 2022. [Online]. Available: https://arxiv.org/abs/2211.12717

[11] J. Jaskari, J. Sahlsten, T. Damoulas, J. Knoblauch, S. Särkkä, L. Kärkkäinen, K. Hietala, and K. Kaski, "Uncertainty-aware deep learning methods for robust diabetic retinopathy classification," 2022. [Online]. Available: https://arxiv.org/abs/2201.09042

[12] J. Krause, V. Gulshan, E. Rahimy, P. Karth, K. Widner, G. S. Corrado, L. Peng, and D. R. Webster, "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, Aug. 2018. [Online]. Available: https://doi.org/10.1016/j.ophtha.2018.01.034

[13] G. Scalia, C. A. Grambow, B. Pernici, Y. Li, and W. H. G. Jr., "Evaluating scalable uncertainty estimation methods for dnn-based molecular property prediction," *CoRR*, vol. abs/1910.03127, 2019. [Online]. Available: http://arxiv.org/abs/1910.03127

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *CoRR*, vol. abs/1706.04599, 2017. [Online]. Available: http://arxiv.org/abs/1706.04599

[15] and and and, *Improving Diagnosis in Health Care*, E. P. Balogh, B. T. Miller, and J. R. Ball, Eds. National Academies Press, Dec. 2015. [Online]. Available: https://doi.org/10.17226/21794

[16] S. Muchuchuti and S. Viriri, "Retinal disease detection using deep learning techniques: A comprehensive review," *Journal of Imaging*, vol. 9, no. 4, p. 84, Apr. 2023. [Online]. Available: https://doi.org/10.3390/jimaging9040084

[17] M. Z. Atwany, A. H. Sahyoun, and M. Yaqub, "Deep learning techniques for diabetic retinopathy classification: A survey," *IEEE Access*, vol. 10, pp. 28 642–28 655, 2022.

[18] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, "Diabetic retinopathy detection through deep learning techniques: A review," *Informatics in Medicine Unlocked*, vol. 20, p. 100377, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914820302069

[19] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. X. Zhu, "A survey of uncertainty in deep neural networks," 2021. [Online]. Available: https://arxiv.org/abs/2107.03342

[20] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa, "Explainable AI in medical imaging: An overview for clinical practitioners – beyond saliency-based XAI approaches," *European Journal of Radiology*, vol. 162, p. 110786, May 2023. [Online]. Available: https://doi.org/10.1016/j.ejrad.2023.110786

[21] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" 2017. [Online]. Available: https://arxiv.org/abs/1703.04977

[22] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," 2015.

[23] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951. [Online]. Available: https://doi.org/10.1214/aoms/1177729694

[24] S. Farquhar, M. A. Osborne, and Y. Gal, "Radial bayesian neural networks: Beyond discrete support in large-scale bayesian deep learning," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. PMLR, 26–28 Aug 2020, pp. 1352–1362. [Online]. Available: https://proceedings.mlr.press/v108/farquhar20a.html

[25] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1613–1622. [Online]. Available: https://proceedings.mlr.press/v37/blundell15.html

[26] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1050–1059. [Online]. Available: https://proceedings.mlr.press/v48/gal16.html

[27] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," 2017.

[28] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf

[29] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," 2018. [Online]. Available: https://arxiv.org/abs/1806.01768

[30] J. Lee, T. Won, T. K. Lee, H. Lee, G. Gu, and K. Hong, "Compounding the performance improvements of assembled techniques in a convolutional neural network," 2020.

[31] T. Ahmed and N. H. N. Sabab, "Classification and understanding of cloud structures via satellite images with EfficientUNet," Jul. 2021. [Online]. Available: https://doi.org/10.1002/essoar.10507423.1

[32] J. W. C. Emma Dugas, Jared, "Diabetic retinopathy detection," 2015. [Online]. Available: https://kaggle.com/competitions/diabetic-retinopathy-detection

[33] H. Fu, B. Wang, J. Shen, S. Cui, Y. Xu, J. Liu, and L. Shao. Springer International Publishing, 2019, pp. 48–56.

[34] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (idrid)," 2018. [Online]. Available: https://dx.doi.org/10.21227/H25W98

[35] "Ocular disease intelligent recognition odir-5k." [Online]. Available: https://odir2019.grand-challenge.org/

[36] J. Engelmann, A. D. McTrusty, I. J. C. MacCormick, E. Pead, A. Storkey, and M. O. Bernabeu, "Detecting multiple retinal diseases in ultra-widefield fundus imaging and data-driven identification of informative regions with deep learning," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1143–1154, Dec. 2022. [Online]. Available: https://doi.org/10.1038/s42256-022-00566-5

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[38] X. Li, S. Chen, X. Hu, and J. Yang, "Understanding the disharmony between dropout and batch normalization by variance shift," 2018.

[39] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. S. Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N. C. Tan, E. A. Finkelstein, E. L. Lamoureux, I. Y. Wong, N. M. Bressler, S. Sivaprasad, R. Varma, J. B. Jonas, M. G. He, C.-Y. Cheng, G. C. M. Cheung, T. Aung, W. Hsu, M. L. Lee, and T. Y. Wong, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, p. 2211, Dec. 2017. [Online]. Available: https://doi.org/10.1001/jama.2017.18152

APPENDIX

A. Additional information on the datasets

TABLE I: Summary of the datasets including the type of uncertainty present and the class distribution.

Mode	Dataset name	Uncertainty type	Class distribution					Total images
			0	1	2	3	4	
Training	EyePACS	-	73.9%	7.2%	14.7%	2.3%	1.9%	66,562
Validation	EyePACS	-	75.1%	6.6%	13.7%	2.5%	2.1%	8,285
Testing	EyePACS test	-	73.7%	6.9%	16%	1.8%	1.6%	8,285
	EyePACS test (noisy)	Aleatoric	73.7%	6.9%	16%	1.8%	1.6%	8,285
	EyePACS test (blurred)	Aleatoric	73.7%	6.9%	16%	1.8%	1.6%	8,285
	EyePACS test (one channel)	Aleatoric	73.7%	6.9%	16%	1.8%	1.6%	8,285
	EyePACS low quality	Aleatoric	66.6%	5.3%	18.6%	3.4%	6.1%	5,400
	IDRID	Epistemic	32.6%	4.8%	32.6%	18%	12%	516
	AMD (ODIR)	Epistemic	98.5%	1.1%	0.4%	0%	0%	267
	Glaucoma (ODIR)	Epistemic	93.6%	1%	5.4%	0%	0%	312
	ImageNet	Epistemic	100%	0%	0%	0%	0%	900
	Tsukazaki	Epistemic	74.9%	25.1%				235

B. Comparison of MC Dropout accuracy with and without batch normalization in inference mode

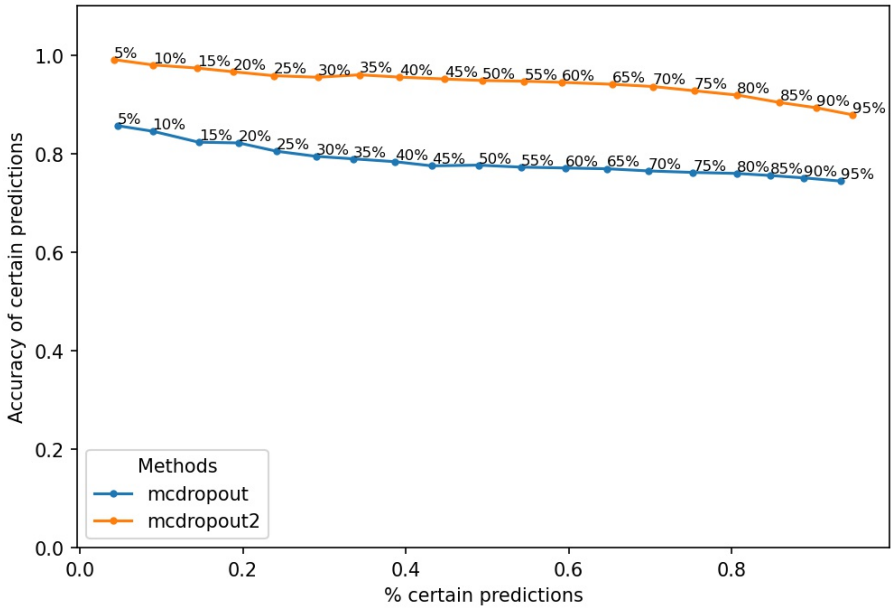


Fig. 16: Plot 1 comparing the accuracy results of the original MC Dropout model (mcdropout) and including batch normalization in inference mode (mcdropout2).