Medical Image Analysis 64 (2020) 101724

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection

Murat Seçkin Ayhan[a,*], Laura Kühlewein[a,b], Gulnar Aliyeva[b], Werner Inhoffen[b], Focke Ziemssen[b], Philipp Berens[a,c,d,*]

[a] Institute for Ophthalmic Research, University of Tübingen, Tübingen, Germany
[b] University Eye Clinic, University of Tübingen, Tübingen, Germany
[c] Bernstein Center for Computational Neuroscience, University of Tübingen, Tübingen, Germany
[d] Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany

ARTICLE INFO

Article history:
Received 24 July 2019
Revised 5 May 2020
Accepted 11 May 2020
Available online 18 May 2020

Keywords:
Deep neural networks
Diabetic retinopathy
Uncertainty
Calibration

ABSTRACT

Deep learning-based systems can achieve a diagnostic performance comparable to physicians in a variety of medical use cases including the diagnosis of diabetic retinopathy. To be useful in clinical practice, it is necessary to have well calibrated measures of the uncertainty with which these systems report their decisions. However, deep neural networks (DNNs) are being often overconfident in their predictions, and are not amenable to a straightforward probabilistic treatment. Here, we describe an intuitive framework based on test-time data augmentation for quantifying the diagnostic uncertainty of a state-of-the-art DNN for diagnosing diabetic retinopathy. We show that the derived measure of uncertainty is well-calibrated and that experienced physicians likewise find cases with uncertain diagnosis difficult to evaluate. This paves the way for an integrated treatment of uncertainty in DNN-based diagnostic systems.

1. Introduction

Deep neural networks (DNNs) are emerging as powerful tools for medical image analysis and disease diagnosis (reviewed by Esteva et al., 2019; Topol, 2019). It has been shown that DNNs can detect diabetic retinopathy (DR) from fundus images (Gulshan et al., 2016) or skin cancer from dermoscopic images (Esteva et al., 2017; Haenssle et al., 2018) with high accuracy. These and similar studies (Ardila et al., 2019; De Fauw et al., 2018; Hannun et al., 2019; McKinney et al., 2020) established that DNNs can achieve or even surpass human level performance on challenging diagnostic tasks, raising the hope that deploying DNNs in clinical settings may improve clinical workflows by automating certain tasks. For instance, a CNN-based image mining tool for DR detection has been integrated into a mobile screening system (Quellec et al., 2017) and first DNN-based systems have been approved by the U.S. Food and Drug Administration (FDA) (FDA news). With the help of DNNs, opportunistic screening could be replaced by a widespread systematic screening approach and carried out directly by family doctors and primary care physicians (Abràmoff et al., 2018; Kanagasingam et al., 2018; Verbraak et al., 2019).

Despite these impressive achievements, the probabilistic outputs of DNNs are not reliable estimates of the true probability of their predictions being correct (Guo et al., 2017; Vaicenavicius et al., 2019). Thus, DNNs do not generate well-calibrated, reliable uncertainty estimates regarding their decisions (Gal and Ghahramani, 2016; Guo et al., 2017; Kendall and Gal, 2017; Lakshminarayanan et al., 2017; Malinin and Gales, 2018). In fact, the deeper the network, the less-well is it typically calibrated (Guo et al., 2017). In addition, the interobserver variability in human readings of medical images is high (Elmore et al., 1994; 2015; Krause et al., 2018; Sayres et al., 2019; Sussman et al., 1982), which suggests that traditional screening methods can miss significant numbers of disease cases, especially in large cohorts. Considering that human annotations (*labels*) are also used to supervise the training of DNNs, it is imperative to obtain reliable uncertainty estimates (Bhise et al., 2018) so that clinical professionals can properly judge whether an automated diagnosis can be trusted and integrated into the clinical workflow (Grote and Berens, 2019). For instance, a recent study (Kiani et al., 2020) showed that doctors using DNN-based decision support make more errors for cases in which the DNN is wrong. By additionally reporting the predictive uncertainty of the DNN, the fraction of wrong diagnosis could be reduced in such settings. But this would only work with well-calibrated uncertainty measures,

* Corresponding authors.
E-mail addresses: murat-seckin.ayhan@uni-tuebingen.de (M.S. Ayhan), philipp.berens@uni-tuebingen.de (P. Berens).

https://doi.org/10.1016/j.media.2020.101724
1361-8415/© 2020 Elsevier B.V. All rights reserved.

that is, the uncertainty needs to be higher when more errors are made.

Ideally, one would formulate DNNs in a Bayesian framework to obtain well-calibrated uncertainty estimates. While this is mathematically straightforward (Bishop, 2006; Murphy, 2012; Neal, 2012), such Bayesian DNNs are computationally challenging for many real world problems (Gal and Ghahramani, 2016; Kendall and Gal, 2017; Lakshminarayanan et al., 2017). Recently, tractable solutions to this problem have been proposed. For example, it has been shown that using dropout – originally developed as a regularizer – at test time can provide viable uncertainty estimates (Gal and Ghahramani, 2016), also in a medical setting (Leibig et al., 2017). This method is called Monte Carlo Drop-Out (MCDO). However, state-of-the-art network architectures (He et al., 2016a; 2016b; Howard et al., 2017; Huang et al., 2017; Szegedy et al., 2017; Xie et al., 2017) use Batch Normalization (BN) (Ioffe and Szegedy, 2015; Ioffe, 2017) as an implicit regularizer (Luo et al., 2019; Zhang et al., 2017) instead of explicit dropout layers. BN has also been cast as an approximate Bayesian inference method called Monte Carlo Batch Normalization (MCBN) (Teye et al., 2018) where one uses the stochasticity of minibatch statistics in order to obtain uncertainty estimates. However, the quality of such Bayesian uncertainty estimates strictly depends on the suitability of the prior and the approximation efficacy which is usually tied to computational constraints (Lakshminarayanan et al., 2017). Moreover, MCBN trades the network's maximum classification performance for uncertainty estimates (see in Section 2.2).

Non-Bayesian alternatives can offer simpler yet effective means to quantify the uncertainties of DNNs. For instance, an ensemble of DNNs, each of which is randomly initialized, can sample diverse and accurate predictors from a function space and improve upon the single network performance both in accuracy and uncertainty (Lakshminarayanan et al., 2017; Fort et al., 2019; Ovadia et al., 2019). While the ensemble approach is simple and easy to implement, it is typically costly to work with many networks during both training and inference. Here, we build on recent ideas (Ayhan and Berens, 2018; Wang et al., 2019) using test-time data augmentation (TTAUG) to estimate the diagnostic uncertainty in DNNs in an intuitive and data-driven way. Briefly, the network is presented with a number of subtle variations of the input images and we record the network's responses to such changes in the vicinity of examples in the input spaces. From this distribution over network outputs, we can derive an estimate of predictive uncertainty. Importantly, this technique can be applied to state-of-the-art networks regardless of their design choices or regularization methods. We apply TTAUG to the case of detecting DR from fundus images, a well understood diagnostic task, for which the high performance of DNNs has been demonstrated. We find that TTAUG not only yields well-calibrated uncertainty estimates, but also makes the network more robust on previously unseen data. Furthermore, we validate our uncertainty measures clinically, by showing that disagreement between clinicians is higher for images for which the network reported high uncertainty. This is a crucial prerequisite for making such uncertainty estimates useful in practice.

2. Methods

2.1. Bayesian deep neural networks

In a supervised scenario, a DNN is a sophisticated function that maps inputs to outputs: $y = f_\theta(\mathbf{x})$, where θ represents the network's free parameters. In order to infer an optimal configuration of θ, the network is typically trained on a finite dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$. To obtain a probabilistic view on this issue, DNNs can be viewed from a Bayesian perspective:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \tag{1}$$

where a prior $p(\theta)$ combined with the likelihood of data leads to a posterior distribution. Then, the network can be used for predictions on new examples as follows: $\mathbb{E}[y] = \int f_\theta(\mathbf{x})p(\theta|\mathcal{D})d\theta$. Despite the elegance of the Bayesian framework, it is intractable to directly work with this integral, as the parameter space for DNNs is extremely high-dimensional. Typically, we do not have access to the true posterior over the network parameters: $p(\theta|\mathcal{D})$. Instead, we resort to a single point estimate, which leads us to point predictions. A DNN achieves classification typically via a *softmax* function in its classification layer and estimates the class probabilities of an image $\mathbf{x}$ as follows: $p(y = k|\mathbf{x};\theta) = p_k = e^{f_{\omega_k}(\hat{\mathbf{x}})}/\sum_j e^{f_{\omega_j}(\hat{\mathbf{x}})}$, where k is an index into K classes, $\omega_k \subset \theta$ represents the weights and bias for the kth class in the softmax layer, $\hat{\mathbf{x}}$ is the feature representation by the network's penultimate layer, and outputs are multinomial distributions: $\sum_k p_k = 1$.

2.2. Quantifying uncertainty in DNNs

The predictive uncertainty of DNNs can be decomposed into two parts: *epistemic uncertainty* and *aleatoric uncertainty* (Kendall and Gal, 2017; Malinin and Gales, 2018). Epistemic uncertainty can be formalized by means of a probability distribution over the model parameters and accounts for our ignorance about them. It is also known as model uncertainty and can be explained away given enough data (Kendall and Gal, 2017; Malinin and Gales, 2018). The remaining aleatoric uncertainty is caused by the variation or noise in the observations, corresponding to the model's input-dependent (data) uncertainty. Aleatoric uncertainty is *irreducible*, even with more training data (Kendall and Gal, 2017; Malinin and Gales, 2018).

To capture the model uncertainty (Kendall and Gal, 2017), MCDO (Gal and Ghahramani, 2016) performs approximate Bayesian inference via T forward passes with random selections of active neurons, given a test case. Similarly, MCBN (Teye et al., 2018) exploits the stochasticity of minibatch statistics in order to obtain predictive distributions. T forward passes with random minibatches sampled from training data during inference leads to a predictive distribution for a test case. In this respect, MCBN also captures model uncertainty (Teye et al., 2018). However, MCBN explicitly trades the well-informed moving averages of population statistics with those obtained from individual minibatches for the sake of quantifying the stochasticity in predictions. This leads to a less than optimal discrimination performance of the network (Ioffe and Szegedy, 2015), especially given the whole purpose of *Batch Renormalization (BReN)* (Ioffe, 2017) that makes a deliberate use of the moving averages and improves on BN. Also, MCBN requires the training data to be available for sampling during inference. The procedure further requires that the same minibatch size be specified during training and inference; otherwise, the respective approximate posteriors would be inconsistent (Teye et al., 2018). If a pretrained network is used for inference, it is virtually impossible to figure out the training minibatch size, unless it is documented.

Considering the epistemic and aleatoric uncertainties as properties of the model and the data, respectively, a Bayesian deep learning framework that jointly models both types of uncertainties has been recently proposed (Kendall and Gal, 2017). While simultaneous modeling of both enables the best uncertainty calibration, the overall calibration quality has been found to be dominated by the explanation of aleatoric uncertainty. The contribution from the epistemic component is marginal; hence, in the regime of big data, it is more effective to tackle aleatoric uncertainty (Kendall and Gal, 2017).

Table 1
Data characteristics.

		Kaggle DR			IDRiD
	Partition # of images (%)	Training 35,126 (100)	Validation 10,906 (100)	Test 42,670 (100)	Out of Distribution 516 (100)
Severity	*0-No DR*	25,810 (73.48)	8130 (74.55)	31,403 (73.60)	168 (32.56)
	1-Mild DR	2443 (6.95)	720 (6.60)	3042 (7.13)	25 (4.84)
	2-Moderate DR	5292 (15.07)	1579 (14.48)	6282 (14.72)	168 (32.56)
	3-Severe DR	873 (2.49)	237 (2.17)	977 (2.29)	93 (18.02)
	4-Proliferative DR	708 (2.02)	240 (2.20)	966 (2.26)	62 (12.02)

Recently, a non-Bayesian *ensemble* approach has been combined with *adversarial* training (Lakshminarayanan et al., 2017). The ensemble consists of multiple neural networks and it is diversified by their random initialization and random shuffling of training examples. The ensemble effectively samples diverse and accurate predictive functions from the training trajectories of its members (Fort et al., 2019) and readily provides predictive distributions and uncertainty estimates during inference (Lakshminarayanan et al., 2017). Adversarial examples, which are essentially augmented training examples (Goodfellow et al., 2014; Lakshminarayanan et al., 2017), explore the local neighborhood of the original training examples to test the robustness of neural networks. As a result, the likelihood of the data smooths out in the ε-neighborhood of examples and the ensemble generates well-calibrated outputs (Lakshminarayanan et al., 2017), thanks to the capturing of the both uncertainty types together. Ensembles typically work better than MCDO (Ovadia et al., 2019). Nevertheless, the input-space exploration is guided by gradients and the adversarial examples are attracted towards the objective function (Ayhan and Berens, 2018). While the procedure avoids the computational burden of the ideal exploration in all directions (Lakshminarayanan et al., 2017), it leaves some areas of the input space underexplored (Ayhan and Berens, 2018).

Selective prediction has been recently discussed in the context of DNNs (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017; 2019). The proposed framework constructs a well-calibrated classifier, given an uncertainty estimation function for the classifier. While the initial selective classifiers adopted pretrained networks and learned selection functions separately (Geifman and El-Yaniv, 2017), *SelectiveNet* trains a classifier and its selection function simultaneously (Geifman and El-Yaniv, 2019). Thus, the classifier focuses on the most relevant examples and offers risk guarantees for its predictions at a given threshold for its coverage of data (Geifman and El-Yaniv, 2017; 2019). So far, this framework has been shown to work with uncertainty estimates from softmax outputs and MCDO (Geifman and El-Yaniv, 2017; 2019).

2.3. Datasets and disease detection tasks

We evaluate our method using two collections of fundus images: (i) a data set from a Kaggle competition (Kaggle.com, 2015) and (ii) the Indian Diabetic Retinopathy Image Dataset (IDRiD) (Porwal et al., 2018) (see Table 1). All images in both datasets are graded according to the International Clinical Diabetic Retinopathy Severity Scale (Int. Council of Ophth; Wu et al., 2013). The Kaggle DR dataset is severely imbalanced and the examples of No DR dominate all partitions for training, validation and test. In our study, we adhere to the same partitions of data as during the Kaggle competition. The IDRiD images are acquired via a Kowa VX-10α digital fundus camera with 50° field of view (Porwal et al., 2018). In general, IDRiD images are of better quality and contain much less camera artifacts than Kaggle DR images. We use the entire IDRiD data to test the generalization performance of our network and uncertainty measures on *out of distribution (OoD)* examples.

We trained a network to discriminate all five levels of the DR scale in Table 1. We evaluate it for two binary disease detection tasks, considering either Mild DR or Moderate DR as the disease onset. To this end, we dichotomized the network outputs by summing up the softmax values accordingly, resulting in the following groupings: {0} vs. {1,2,3,4} and {0,1} vs. {2,3,4}, respectively. In these discrimination tasks, $p(y = 1|\mathbf{x};\theta)$, where 1 marks the presence of disease, is sufficient to indicate the most likely decision.

2.4. Deep neural network architecture and training

We implemented a CNN based on the residual learning with bottleneck design (He et al., 2016a; 2016b) (Fig. 1a, Table 1 in supplements). We modified the original architecture (He et al., 2016a) using an additional fully connected layer before softmax. Also, our network uses parametric ReLUs (PReLUs) (He et al., 2015) in the first convolutional stack and fully connected layer. Since these parts of the network have no residual connections, PReLUs promote gradient propagation through those layers. The intermediate convolutional stacks adopt the residual architecture with ReLUs. Head nodes use max pooling, if necessary, to downsample their features maps as well as 1 × 1 convolution in order to match the number of channels to that of the corresponding residual parts. Our residual blocks also adopt EraseReLU (Dong et al., 2017), which eliminates the last ReLU in each residual block and reduces nonlinearity in the hopes of easing optimization and improving the generalization performance. All weight layers use *Batch Renormalization (BReN)* (Ioffe, 2017). We concatenate the max and average pooled features from the convolutional stack before the fully connected layer. Finally, 512 features from the penultimate layer are fed into a 5-way softmax for classification.

We construct a 15-layer network and train it with softmax cross-entropy loss in an *end-to-end* fashion for 500,000 iterations with SGDR (Loshchilov and Hutter, 2016) with cosine decay. For the first 50,000 iterations, we use balanced minibatches of 20 images account for severe class imbalance. Then, we increase the minibatch size to 23 and sample stratified minibatches until the end of training, which lets the network adjust itself to the true distribution of classes with an incentive of oversampling[1] Also note that we apply data augmentation with the probability of 0.9 at each iteration. As a result, the network is exposed to 11,350,000 images, about 10,215,000 of which are randomly generated on the fly.

Weights are initialized via *He's initialization* (He et al., 2015) and we use weight decay ($L2$-reg.) with $\lambda = 1e-5$ for all weight layers except the fully connected layer that adopts $L1$-reg. with the same λ to promote the sparsity of features in the penultimate layer. Initial learning rate is 0.003 and it decays with a rate of 0.8 according to the cosine decay schedule, which starts with a period of 10,000 iterations and doubles it after each cy-

[1] Minibatches of 23 consists of 15,2,4,1 and 1 examples from classes in respective order, which still induces an oversampling of the minority classes; however, it preserves the majority of the No DR examples. Also note that 23 is the maximum number of examples that fit into the GPU memory per iteration.

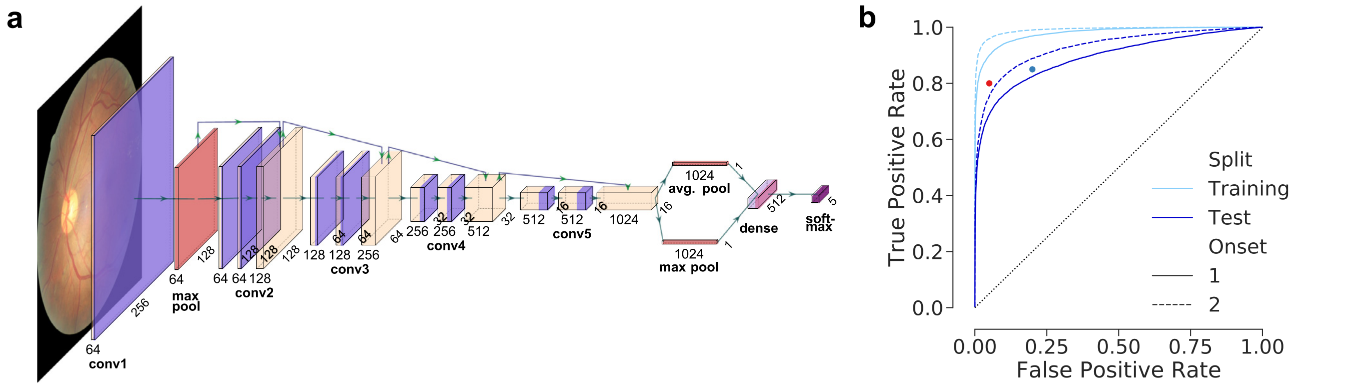

Fig. 1. Network architecture and single prediction performance on disease detection tasks. **(a)**: A CNN model with residual connections and 5-way *softmax* function, plotted with PlotNeuralNet (Iqbal, 2018). **(b)**: Receiver Operating Characteristic (ROC) curves w.r.t. onset levels 1 and 2. Validation results are excluded for clarity. Thresholds (Younis et al., 2003; Leibig et al., 2017) suggested for the detection of sight-threatening (moderate) DR by the British Diabetic Association (BDA) (80%/95% sensitivity/specificity) and the NHS Diabetes Eye Screening programme (85%/80% sensitivity/specificity) are also given as *red* and *blue* dots, respectively. Sensitivity=TPR, specificity=1-FPR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cle. The minimum learning rate is set to 1% of the initial rate. The momentum coefficient μ is initialized to 0.5 and updated w.r.t. $\min(1-2^{-1-\log_2(\lfloor t/250\rfloor+1)}, \mu_{\max})$ (Sutskever et al., 2013), where t is the iteration number and $\mu_{\max}=0.9$, until the 95% of training is completed. Then, we set $\mu=0.5$ once again to allow for finer convergence towards the end of training. The learnable parameters of the BReN layers are initialized as $\gamma=1, \beta=0$ with the exception that the last BReN layers in residual blocks are initialized with $\gamma=0, \beta=0$, which emphasizes the information propagation via the identity connections and eases the optimization, especially in the early stages of training (Goyal et al., 2017). After all, the performance of the network is measured on the validation set once in 5000 iterations. The best performing configuration is saved and used for inference. Training curve and model selection based on validation performance are shown in Fig. 1 in supplements.

We used Tensorflow 1.9-1.13 (Abadi et al., 2015) for the development, training and validation of our networks, all of which were performed on an NVIDIA Titan Xp GPU with 12GB memory and CUDA versions 8/9/9.2 and cuDNN 7. Code and models will be available at https://github.com/berenslab/ttaug-DR-uncertainty upon publication.

2.5. Image preprocessing and data augmentation

All images were cropped to a squared center region, resized to 512 × 512 pixels and locally color-normalized for contrast enhancement (Leibig et al., 2017). For the sake of floating point representation of images, pixel values were mapped from [0, 255] into [0, 1]. Also, all images that are fed to the network were standardized w.r.t. the global image statistics.

In addition to image processing, we perform data augmentation before providing images to the network during both training and testing. To this end, we first adopt a recently proposed image acquisition model (Wang et al., 2019) and its transformation function $\mathcal{T}$, which is parameterized by $\Phi=[\phi_1,\ldots,\phi_n]$ for n different transformations: $\mathbf{x}=\mathcal{T}_\Phi(\mathbf{x}_0)$. In the end, an augmented image $\mathbf{x}$ is a result of a series of transformations embedded into $\mathcal{T}$ and applied onto $\mathbf{x}_0$. Assuming prior distributions over the parameters of transformations, the distribution of augmented images $p(\mathbf{x})$ covers the distribution of observations $p(\mathbf{x}_0)$ (Wang et al., 2019). However, in a classification setting, where the goal is to learn a mapping from images to class labels, y and y_0 are ideally invariant to $\mathcal{T}$. In other words, one should refrain from excessive transformations that might severely alter the structure in the data.

We use the following operations in our data augmentation pipeline.

1. Crop and resize: A random variable *crop* controls whether an image will be cropped or not: *crop* ~ *Bern*(ϕ_{11}) where $\phi_{11}=0.5$. If yes, the lower left and upper right corners of the box to crop into are independently and randomly sampled from the margins of $\phi_{12}=0.15$ of height or width from each side: $x_1\sim U(0,\phi_{12}), y_1\sim U(0,\phi_{12}), x_2\sim U(1-\phi_{12},1), y_2\sim U(1-\phi_{12},1)$. Then, the image is resized to its previous dimensions. Note that $\phi_1=[\phi_{11},\phi_{12}]$.
2. Color transformations are applied w.r.t. factors sampled as follows:
   - Brightness: b ~ $U(\phi_{21}, \phi_{22})$ where $\phi_{21}=-0.15$ and $\phi_{22}=0.15$
   - Hue: h ~ $U(\phi_{31}, \phi_{32})$ where $\phi_{31}=-0.15$ and $\phi_{32}=0.15$
   - Saturation: s ~ $U(\phi_{41}, \phi_{42})$ where $\phi_{41}=0.5$ and $\phi_{42}=2.5$
   - Contrast: c ~ $U(\phi_{51}, \phi_{52})$ where $\phi_{51}=0.5$ and $\phi_{52}=1.5$
3. Geometric transformations include independent horizontal and vertical flips, translation and rotation.
   - Horizontal flip is controlled by f_h ~ *Bern*(ϕ_6) where $\phi_6=0.5$
   - Vertical flip is controlled by f_v ~ *Bern*(ϕ_7) where $\phi_7=0.5$
   - Translation offsets by pixels in 2D: t_x ~ $U(\phi_{81}, \phi_{82})$ and t_y ~ $U(\phi_{81}, \phi_{82})$ where $\phi_{81}=-25$ and $\phi_{82}=25$
   - Rotation angle: r ~ $U(\phi_{91}, \phi_{92})$ where $\phi_{91}=-15$ and $\phi_{92}=15$

2.6. Assessing calibration and alleviating it

Ideally, the probabilistic outputs of a classifier reflects the true probability of its predictions being correct, its accuracy. Such a classifier is said to be well-calibrated and its outputs can be readily interpreted as confidence scores. However, DNNs are notoriously miscalibrated and overconfident about their predictions (Ding et al., 2019; Guo et al., 2017; Vaicenavicius et al., 2019) (Fig. 2b, bottom row). The concordance between confidence and accuracy can be visualized via reliability diagrams (Ding et al., 2019; Guo et al., 2017; Niculescu-Mizil and Caruana, 2005; Vaicenavicius et al., 2019), where a classifier is better calibrated as it gets closer to the diagonal.

The Expected Calibration Error (ECE) (Eq. 2) measures the degree of miscalibration, given M predictions grouped into G bins.

$$\mathrm{ECE} \;= \sum_{g=1}^{G} \frac{|B_g|}{M} \left| acc(B_g) - conf(B_g) \right|$$

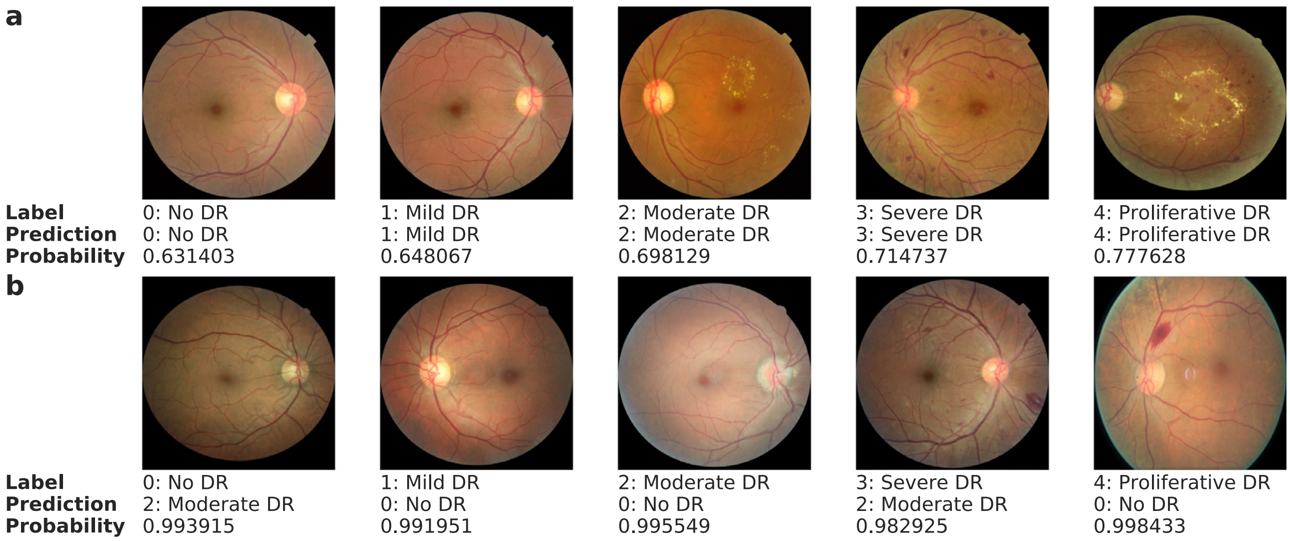

Fig. 2. Examples of test images with the reference labels from the dataset, network's predictions and predictive probabilities from the softmax layer. Top row shows correct predictions, whereas overconfident but wrong predictions are at the bottom.

$$
\begin{aligned}
&= \sum_{g=1}^{G} \frac{|B_g|}{M} \left| \frac{1}{|B_g|} \sum_{\mathbf{x}_i \in B_g} \mathbf{1}\{\hat{y}_i, y_i\} - \frac{1}{|B_g|} \sum_{\mathbf{x}_i \in B_g} p(y_i = \hat{y}_i | \mathbf{x}_i) \right| \\
&= \sum_{g=1}^{G} \frac{|B_g|}{M} \left| \frac{1}{|B_g|} \left(\sum_{\mathbf{x}_i \in B_g} \mathbf{1}\{\hat{y}_i, y_i\} - \sum_{\mathbf{x}_i \in B_g} p(y_i = \hat{y}_i | \mathbf{x}_i) \right) \right|, \qquad (2)
\end{aligned}
$$

where $\hat{y}_i$ and y_i are the predicted and actual class labels, respectively. Also $|B_g|$ is the gth bin size, whereas large bars, e.g., $\left|\ldots\right|$, indicate absolute values.

Due to overconfidence, the distribution of predictive probabilities can be highly *non-uniform*. Thus, reliability diagrams and ECE are strongly tied to the binning decisions (Ding et al., 2019). While large bins may cause the bin errors to be small, a high-resolution histogram with many bins leads to inaccurate estimation of ECE (Ding et al., 2019). Moreover, the gap between the accuracy and confidence of individual predictions may have different signs, which may cause internal compensation inside bins even with uniform distributions (Ding et al., 2019). To avoid such complications, we adopt the Adaptive ECE method (Ding et al., 2019) that relies on adaptive histograms and computes *positive* and *negative* gaps explicitly.

2.6.1. Temperature scaling

Temperature scaling is a parametric approach to calibrating the probabilistic outputs. The basic idea is to rescale the functional values before applying the softmax function so that the probability scores are *softened* and hence better calibrated (Guo et al., 2017; Liang et al., 2018). Only one parameter, τ, needs to be chosen:

$$
p(y = k | \mathbf{x}; \theta, \tau) = p_k = \frac{e^{f_{\omega_k}(\hat{\mathbf{x}})/\tau}}{\sum_j e^{f_{\omega_j}(\hat{\mathbf{x}})/\tau}} \qquad (3)
$$

We fit τ to the validation data minimizing the negative log-likelihood (Guo et al., 2017). Note that τ is not an essential part of our network. We use it only to adjust the functional values of the network in a post-processing fashion, which does not change the final decisions. Thus, the accuracy remains untouched, while the probabilities are calibrated.

2.6.2. Ensemble of networks

The ensemble approach aims at improving the accuracy and uncertainty quantification of DNNs by using multiple networks. In this setting, each network is randomly initialized, follows a different optimization trajectory and explores a different mode in function space; the ensemble decisions are based on multiple predictive functions sampled from these modes (Lakshminarayanan et al., 2017; Fort et al., 2019). Even small ensembles have performed well on standard datasets (Lakshminarayanan et al., 2017; Fort et al., 2019). Using the same network architecture (Section 2.4), hyperparameters and training procedures, we also construct an ensemble of 3 networks. These are diversified by the randomness in not only initialization but also the shuffling of training examples and data augmentation.

2.7. Test-time data augmentation based uncertainty measures

Although data augmentation for training is well-established and has been used to improve the discriminative performance of models during inference, its use in predictive uncertainty estimation is underexplored (but see ref (Ayhan and Berens, 2018; Wang et al., 2019)). In a probabilistic classification setting, where the classifier outputs are estimated class membership values and only suggest likely class assignments, rather than making deterministic predictions, y becomes a random response to examples from $p(\mathbf{x})$. However, due to a mixture of various types of transformations, $p(\Phi)$ can be a sophisticated joint distribution. Therefore, it is difficult to exactly compute the following quantity to arrive at an aggregate decision with regards to the variation in inputs:

$$
\begin{aligned}
\mathbb{E}[p(y = 1 | \mathbf{x}; \theta)] &= \int p(y = 1 | \mathbf{x}; \theta) p(\mathbf{x}) d\mathbf{x} \\
&= \int p(y = 1 | \mathcal{T}_\Phi(\mathbf{x}_0); \theta) p(\Phi) d\Phi \qquad (4)
\end{aligned}
$$

Therefore, we resort to sampling and propose to use TTAUG as an approximate method to evaluate the predictive uncertainty associated with observations. Simply, we sample Φ_t, where $t \in \{1, \ldots, T\}$, and generate T variations ($\mathbf{x}_t$) of a given observation $\mathbf{x}_0$ (Fig. 3a). As a result, we obtain a distribution of predictive probabilities as a proxy for the true but sophisticated one (Fig. 3b). Then, we can examine how much the network output varies in the vicinity of examples in all directions in the input spaces (see Section 2.8 and Fig. 10 in supplements for illustration and discussion).

The measures of spread, e.g., standard deviation (STD), interquartile range (IQR), are sufficient under binary classification

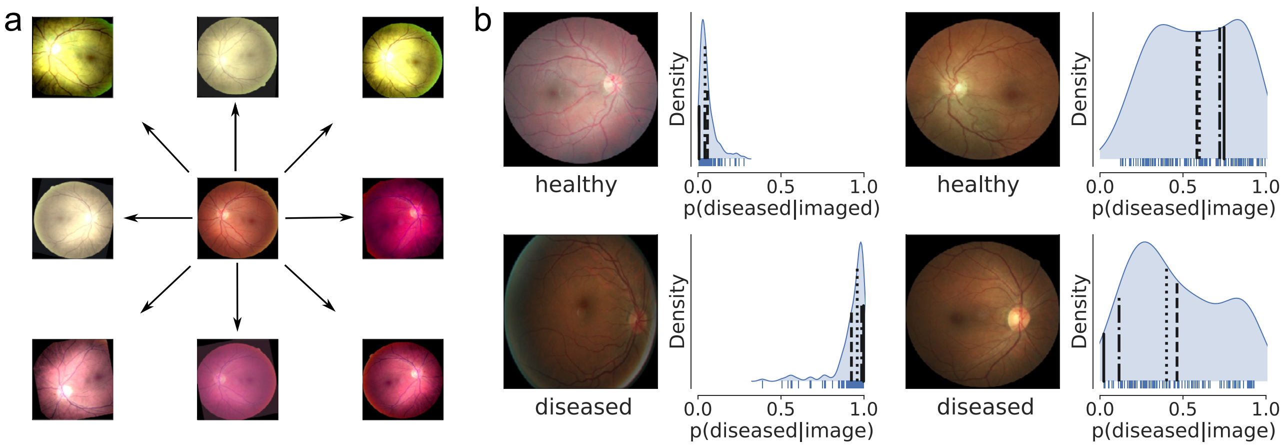

Fig. 3. Test-time data augmentation and uncertainty. **(a)**: Given an image in the center, T variations are randomly generated via a series of geometric and color transformations. This can be understood in analogy to doctors looking at images from different angles and/or under various illumination so that an underlying disease pattern can be examined from multiple views. (**b**): Exemplary fundus images (top: healthy, bottom: diseased) and corresponding distributions of predictive probabilities. Given $T = 128$ predictions under onset 1 scenario, mean and median predictions are shown as dashed and dotted lines, respectively. Solid lines indicate the single point predictions, whereas the dashdot style indicates the ensemble predictions. Label assignment is achieved by thresholding the predictions at 0.5. Above the threshold, the label is `diseased`, otherwise `healthy`. Note that the predictive distribution entropy (or spread) is higher in the cases of wrong predictions.

scenarios. However, these are *undefined* for single point predictions, by which we establish our baseline (Hendrycks and Gimpel, 2017) performances. Therefore, we use *entropy* (Eq. 5) which is universally applicable to probability distributions.

$$H[p(y|\mathbf{x};\theta)]] = -\sum_k p(y=k|\mathbf{x};\theta)\log p(y=k|\mathbf{x};\theta)$$
$$= -\sum_k p_k \log p_k \tag{5}$$

For a binary task, it reduces to $H[p(y|\mathbf{x};\theta)] = -p\log p - (1-p)\log(1-p)$, where $p = p(y=1|\mathbf{x};\theta)$. Given an *ensemble* of T predictions $\{p(y_t = 1|\mathbf{x}_t;\theta)\}_{t=1}^{T}$, the entropy of the expected prediction $H[\mathbb{E}[p(y|\mathbf{x};\theta)]]$ indicates the predictive uncertainty associated with the observation $\mathbf{x}_0$ (Malinin and Gales, 2018; Smith and Gal, 2018). Typically, $\mathbb{E}[p(y|\mathbf{x};\theta)]$ is estimated by a simple mean:

$$\frac{1}{T}\sum_{t=1}^{T} p(y_t = 1|\mathbf{x}_t;\theta) = \frac{1}{T}\sum_{t=1}^{T} p(y_t = 1|\mathcal{T}_{\Phi_t}(\mathbf{x}_0);\theta). \tag{6}$$

However, $p(y=1|\mathbf{x};\theta)$ does not necessarily follow a Gaussian distribution (Fig. 3b). Therefore, we used the median prediction for $\mathbf{x}_0$.

2.8. t-distributed stochastic neighbor embedding (tSNE)

We used tSNE (Maaten and Hinton, 2008), a non-linear dimensionality reduction method that enables the interpretation of high-dimensional data in low-dimensional settings, to visualize how uncertainty changes across the space of fundus images. The goal of tSNE is to map high-dimensional data to a low-dimensional space while preserving the structure in data. The most important tSNE parameter is *perplexity*, which effectively determines the size of neighborhood where examples interact with each other. We follow the approach .of Kobak and Berens (2019) and adopt large perplexity values, e.g., $\sim N/50$, where N is the number of examples. Since this quickly becomes computationally infeasible for large N, we use Fast Fourier Transform-accelerated Interpolation-based tSNE (FI-tSNE) (Linderman et al., 2019). Also, we adopt the PCA initialization (Kobak and Berens, 2019), not only because it promotes the preservation of global structure but also for the reproducibility of tSNE outputs.

2.9. Expert validation

In order to validate our uncertainty estimates, we selected a subset of test images and had them re-evaluated by four ophthalmologists according to the same five DR grades provided in the Kaggle data set. One ophthalmologist (GA) was in training with 3 years of clinical experience. The other three are fully trained ophthalmologists or retina specialists with 6–10 years of clinical experience in medical retina. The subset consisted of 65 images selected based on the predictive uncertainty under the onset 1 scenario as well as heuristics for colorfulness (Hasler and Suesstrunk, 2003) and brightness for the sake of quality in visual inspection. Given the distribution of uncertainty among `correct` and `wrong` predictions (see Fig. 6 in supplements), 5 images were sampled from the following categories: i) wrong and uncertain: 0-No DR, 1-Mild DR, 2-Moderate DR, ii) wrong and confident: 2-Moderate DR, 3-Severe DR and 4-Proliferative DR, iii) correct and uncertain: 0-No DR, 1-Mild DR and 2-Moderate DR, iv) correct and confident: 0-No DR, 1-Mild DR, 2-Moderate DR and 3-Severe DR. We divided the images into two groups based on their predictive uncertainty, which resulted in low and high uncertainty examples. Then, we compared the DNN decisions with experts' as well as expert decisions with each other for their coherence. To this end, we used Cohen's kappa score (Cohen, 1960) that measures agreement between two annotators. Its range is $[-1, 1]$, where 1 indicates a full agreement, whereas lower scores mean less agreement. A negative score can also be interpreted as the degree of disagreement.

3. Results

We developed a DNN based on the ResNet architecture to detect diabetic retinopathy from fundus images (Fig. 1a, see Methods). The network was trained with five different categories (healthy and four disease stages), but evaluated only in a binary task, where we considered either mild (level 1) or moderate (level 2) DR as disease onset. The prediction performance of our network was competitive with the recommendations of the British Diabetic Association (BDA) and the NHS Diabetes Eye Screening programme (Fig. 1b), especially given that the recommendations are for the sight-threatening (moderate) DR (Leibig et al., 2017; Younis et al., 2003). However, like many state-of-the-art networks (Guo et al., 2017), it provided miscalibrated reports of its uncertainty based on the softmax output (see Methods): While it can correctly classify

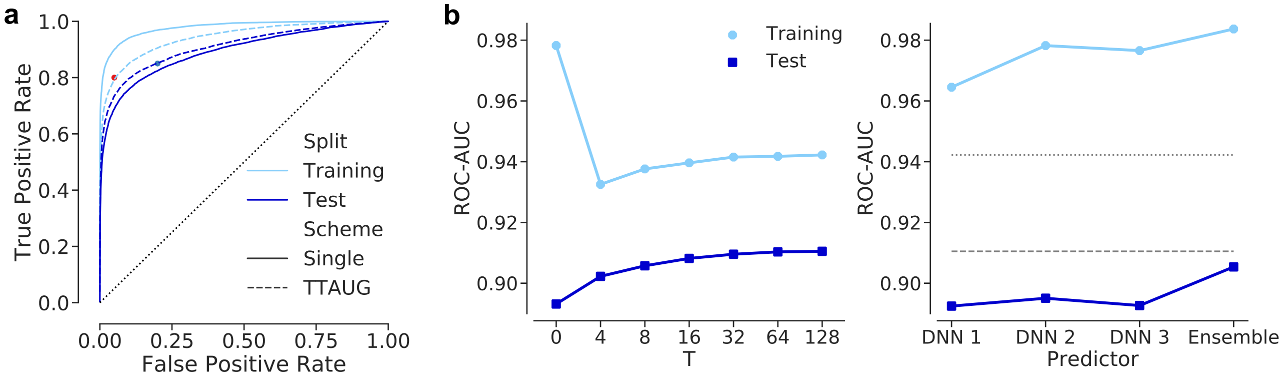

Fig. 4. The impact of TTAUG on disease detection performance under onset 1 scenario and comparison with the ensemble performance. Validation results are excluded for clarity. Onset 2 results are given in Fig. 2 in supplements. **(a)**: Receiver Operating Characteristic (ROC) curves for the binary classification via the single and *median* of $T = 128$ predictions. **(b)**: ROC-AUC scores of the single ($T = 0$) and median predictions w.r.t. various T values (*left*) as well as the ensemble and its members (*right*).

many fundus images with reasonably high confidence (Fig. 2, top row), it also creates wrong predictions with very high confidence (Fig. 2, bottom row).

To obtain well calibrated predictive probabilities, we generated T variations of a given example via subtle modifications (Fig. 3a) and used our network to make a prediction for each augmented example, a procedure we call test-time data augmentation (TTAUG) (Ayhan and Berens, 2018). This resulted in a distribution of predictive probabilities (Fig. 3b). For two examples of correct predictions, this distribution was narrow, indicating that all augmented examples were classified similarly (Fig. 3b, left). For two examples of incorrect predictions, it was broad and even multimodal, suggesting that the augmentation operations lead to widely varying outputs (Fig. 3b, right). We computed the mean or the median of this distribution as a smoothed estimate of the predictive probability.

Interestingly, these smoothed predictive probabilities obtained through TTAUG reduced the generalization gap between training and test accuracy and slightly improved the discrimination performance of the network on the unseen test data compared to the single predictions (Fig. 4a), even with small values of T (Fig. 4b, left). We found that increasing T up to 128 continued to increase performance (also see Fig. 4 in supplements for calibration and uncertainty). Of course, there is a trade-off between the efficiency of creating predictions with high T and the slight increases in prediction performance. Nevertheless, the computational cost induced by large T is redeemed in the face of the ensemble's large generalization gap and lower discrimination performance on the test data (Fig. 4b, right).

Apart from the by-product of improved generalization, we mainly wondered whether the TTAUG-smoothed estimates of the predictive probability were actually better calibrated (Fig. 5). Even though our network consisted of only 15 layers, using the single predictive probability led to overconfident predictions with the largest calibration error (Fig. 5a), as suggested by the examples shown previously (Fig. 2b). In the literature, temperature scaling has been suggested to correct for this overconfidence (Guo et al., 2017; Liang et al., 2018) (see Methods). Also in our case, applying temperature scaling led to better calibrated predictive probabilities (Fig. 5b). Similarly, ensemble predictions yielded also uncertainty estimates comparable to temperature-scaled outputs in terms of calibration error (Fig. 5c). While the mean of the TTAUG distribution of predictive probabilities induced larger calibration error than both the temperature-scaled single predictions and ensemble predictions, with highly underconfident predictions at higher accuracies (Fig. 5d), the median of the TTAUG distribution of predictive probabilities led to the best overall calibration, both visually and quantitatively (Fig. 5e). Even with small values of T, the median prediction maintained its calibration quality and achieved smaller errors than alternatives (see Fig. 4a in supplements).

As the median of the TTAUG-smoothed predictive probabilities was calibrated the best, we computed its entropy to quantify the uncertainty of the diagnostic decision of the network (see Methods). To study the properties of this measure more intuitively, we embedded all test examples into one dimension using t-distributed stochastic nearest neighbour embedding (tSNE) based on the network activations in the penultimate layer (Fig. 6a, see Methods). Although the network was presented only categorical labels and has no explicit knowledge of the underlying disease pathology or progression, the map indicates that the network successfully reconstructed the disease continuum and ordered the classes accordingly (compare to a similar 3D tSNE map by Eulenberg et al. (2017)) with healthy cases being placed to the left and diseased cases to the right. The cluster of 2064 examples mapped to the leftmost (tSNE < -15) consists essentially of examples with poor image quality, noise or missing content despite the reference labels (see Fig. 7 in supplements for example images from this cluster).

We found that the average uncertainty was largest at the boundary between disease stages (Fig. 6b). For example, onset 1 uncertainty was highest around t-SNE coordinate 4, where the transition between no DR/mild and moderate DR and the remaining classes could be found. In addition, samples with higher uncertainty were more likely to be erroneously classified (Fig. 6c). This suggests to use the uncertainty measure for selective prediction (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017; 2019), referring predictions on difficult cases for further evaluation by an expert (Leibig et al., 2017). To compare different uncertainty measures with respect to their ability to refer erroneous predictions efficiently, we ranked the predictions by their uncertainty and measured the network's performance on the remaining cases, given a certain referral rate. This procedure yielded monotonic improvements in prediction performance, with the median of the TTAUG distribution of predictive probabilities resulting in the most efficient referrals (Fig. 6d). With this, our network achieved a ROC-AUC score of 0.959 among the remaining data, when 50% of the cases were referred, compared to 0.940 at the same referral rate via the VGG-like Bayesian CNN in ref. Leibig et al. (2017) which used MCDO-based uncertainty estimates for the same disease detection and referral tasks on the same data. Our network can match their performance at the referral rate of only 32%.

We next evaluated the generalization performance of our network and uncertainty measure on the independent IDRiD data set (see Methods). The network performed better on the previously unseen IDRiD images than on the Kaggle data, (Fig. 7), where it

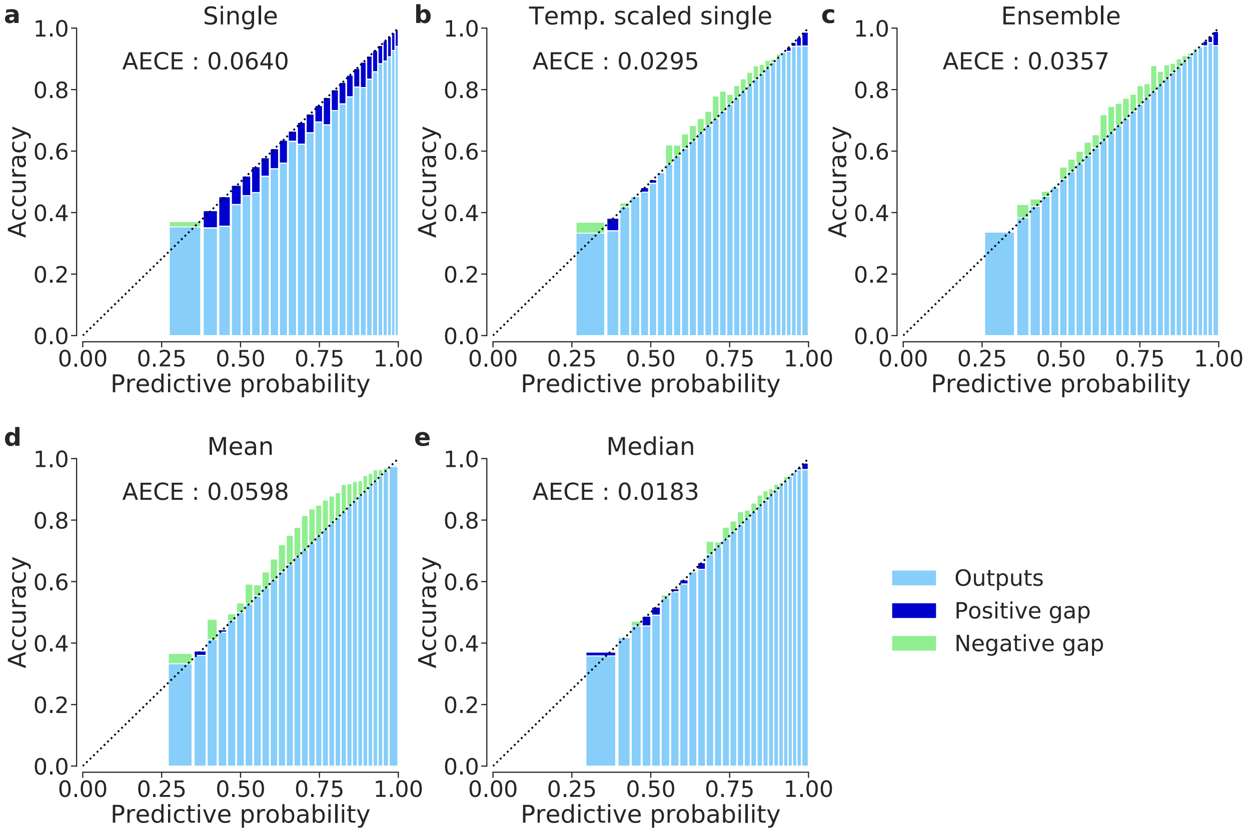

Fig. 5. Reliability diagrams and calibration quality via adaptive ECE (AECE) on test data. Positive gap (blue) indicates overconfidence, whereas negative gap (green) implies the lack of confidence. The mean or median of $T = 128$ softmax outputs are first computed class channel-wise and then renormalized so that the outputs constitute valid probability distributions. **(a)**: Single prediction. **(b)**: Temperature-scaled single prediction ($\tau = 1.4828538$). **(c)**: Ensemble (mean) prediction. **(d)**: Mean prediction. **(e)**: Median prediction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

achieved a ROC-AUC score of 0.975 with single predictions under onset 1 scenario. TTAUG with T=128 increased the performance up to 0.982. The improved performance on the IDRiD data is likely due to the quality differences between the respective datasets with less low quality images. While the calibration of network outputs was not as good as it was with the Kaggle DR data (Fig. 8 in supplements), uncertainty estimates using the entropy of the median of the TTAUG distribution of predictive probabilities was sufficient to detect incorrect predictions (Fig. 7a). As the correct predictions on the IDRiD images had mostly low uncertainty and the incorrect ones mostly high uncertainty, decision referral allows to yield almost perfect classification performance with only 40% of the data referred (Fig. 7b). Interestingly, the uncertainty computed from the temperature scaled single predictive probability fitted to the Kaggle DR validation data did not generalize well to the IDRiD images (see Discussion for generalization). Ensemble predictions were more robust to data shift than temperature scaled predictions and achieved better decision referral (Fig. 7b), despite their comparable calibration errors on the IDRiD images. The ensembling approach yielded almost as good referral curves as our approach despite its worse calibration (Fig. 7b).

Further, we evaluated whether images with high uncertainty were also more difficult for human experts to grade. To this end, we measured the agreement between the reference labels provided with the dataset and expert annotations provided by four ophthalmologists as well as among the ophthalmologists themselves (Fig. 8). First, we selected 65 images covering a range of uncertainties (see Methods and Fig. 6 in supplements), which we divided into two groups of low and high uncertainty based on the entropies of single, mean or median predictions. All images were graded by the ophthalmologists according to the DR severity scale and the grades were dichotomized with respect to the onset level 1. Then, they were compared to the reference labels in the dataset as well as among the ophthalmologists w.r.t. the same disease detection task.

Interestingly, we found that the agreement of judgements by the human experts and the reference labels was higher for the low uncertainty images, with a difference in Cohen's d of 0.23 ± 0.05, 0.61 ± 0.05 and 0.55 ± 0.05 (mean estimate based on ANOVA with SE) between low and high uncertainty cases for single, mean and median uncertainty measures, respectively (Table 2 and Table 3 in supplements, Fig. 8a). In addition, the agreement was significantly higher in the low uncertainty condition for the mean and median of the TTAUG-distribution of predictive probabilities than the single predictions (ANOVA, post-hoc test with Tukey correction of multiple comparisons, $p = 0.0003$ and $p = 0.0018$, respectively; Table 4 in supplements). Furthermore, we measured the agreement between the four ophthalmologists (Fig. 8b) and found that the inter-annotator agreement on the low uncertainty cases was high, whereas agreement between experts was lower when the uncertainty was high with a difference in Cohen's d of 0.27 ± 0.09, 0.61 ± 0.05 and 0.55 ± 0.05 (Table 5 and Table 6 in supplements) for the three measures respectively. Again, the agreement between experts in the low uncertainty condition was significantly higher for the mean and the median of the TTAUG-distribution than for the single predictive entropy (ANOVA, post-hoc test with Tukey correction of multiple comparisons, $p < 0.0001$ and $p = 0.0019$, respectively; Table 7 in supplements). These findings suggest that the TTAUG-based uncertainty measure derived here not only provides better calibration and more efficient referral, but reflects also bet-

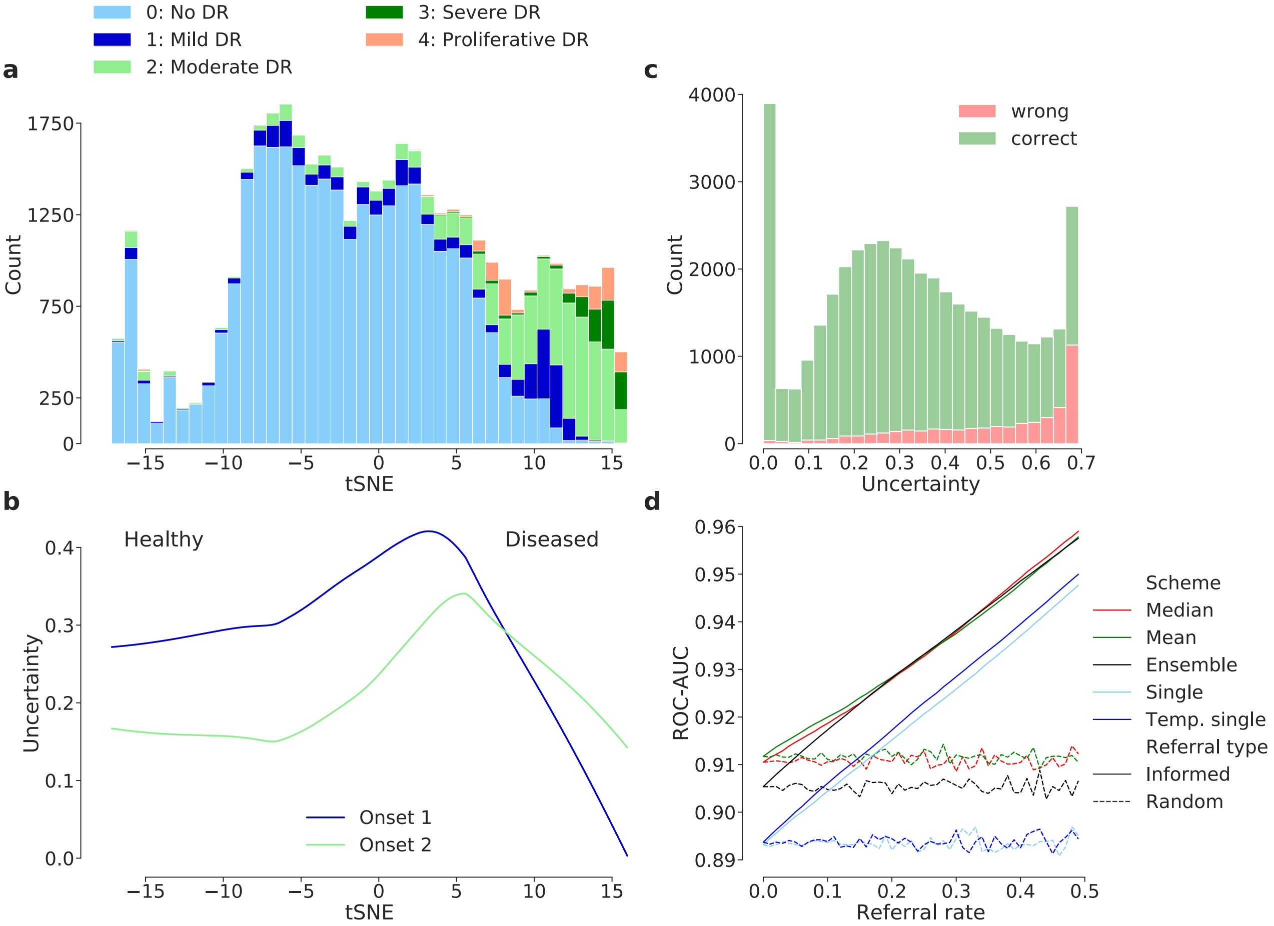

Fig. 6. Interpretation of the network's feature representation and predictive uncertainty for diagnosis, given the test examples. See Fig. 3 in supplements for the onset 2 counterparts of (c) and (d). **(a)** 1D tSNE map via a perplexity of 1000 and 512 features from the penultimate layer. **(b)** Entropy of the median prediction along the disease continuum. Each image is associated with two uncertainty measures under two disease detection scenarios. We fit locally-weighted regression curves to summarize the relations between the tSNE coordinates and predictive uncertainty of examples. **(c)** Distributions of median prediction entropy for all test images grouped by classification results w.r.t the median predictions under onset 1 scenario. **(d)** Improvement in performance via uncertainty-informed decision referral, given different measures of uncertainty under onset 1 scenario. To observe the impact of the number of augmentations *T* on uncertainty and decision referral, see Fig. 4b in supplements. For additional referral plots w.r.t. uncertainty measures derived from the entropy of the entire predictive distribution, see Fig. 5 in supplements. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

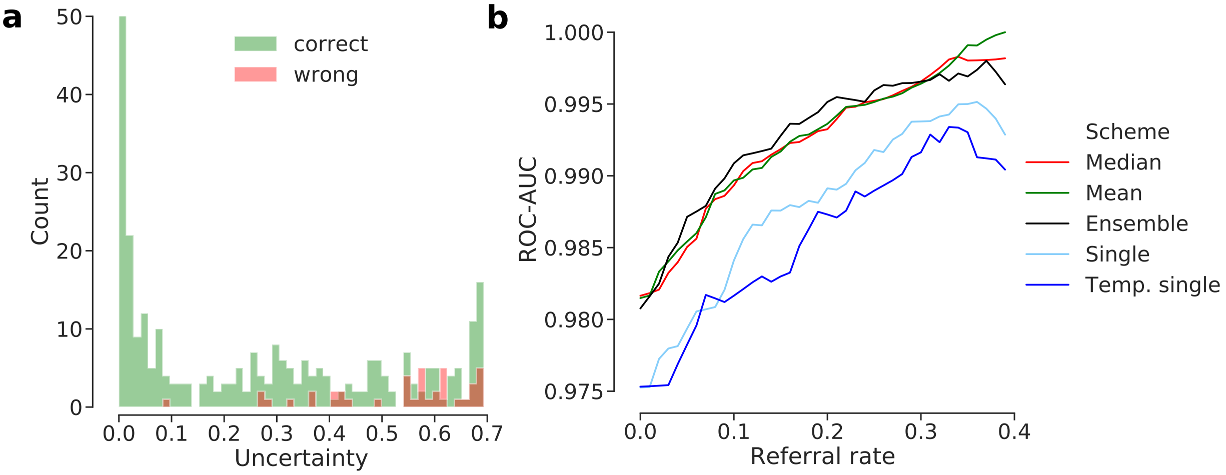

Fig. 7. Generalization of the predictive performance and uncertainty estimates to the unseen IDRiD data under the onset 1 scenario. See Fig. 9 in supplements for onset 2. **(a)** Distributions of median prediction entropy for all images from the whole IDRiD data grouped by classification results w.r.t the median predictions. Counts are clipped at 50 for better visualization of the small bins. **(b)** Decision referral on IDRiD images. Random referral is omitted for clarity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

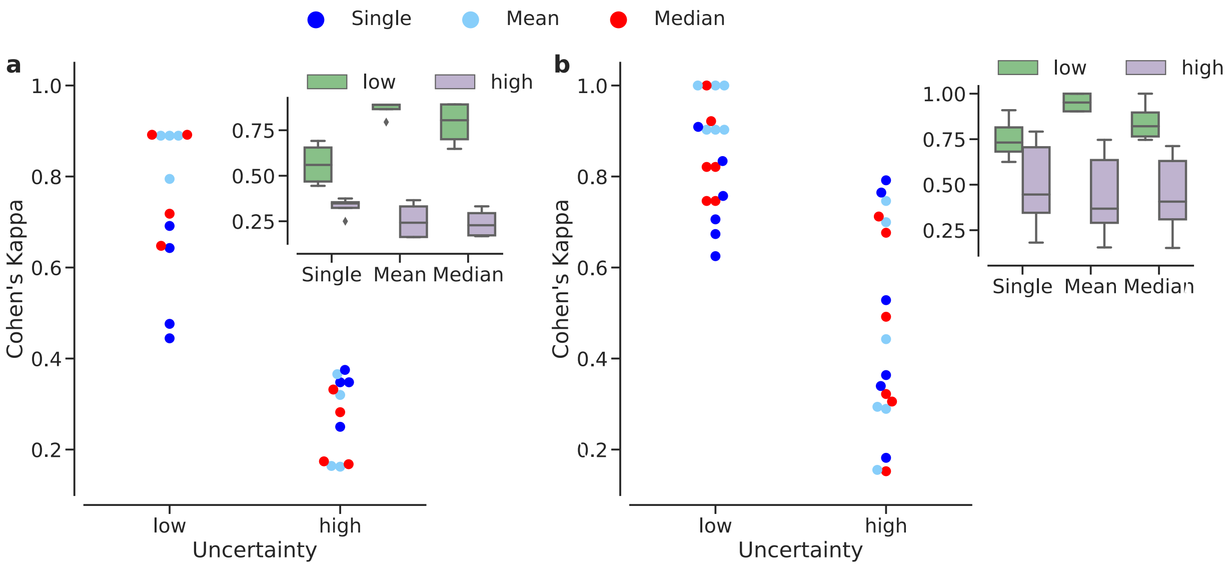

Fig. 8. Agreement between annotators under low or high uncertainty from the onset 1 scenario, given 65 images selected from the Kaggle DR test partition. Uncertainty groups are determined w.r.t. the predictive uncertainty due to single (blue), mean (light blue) or median (red) predictions. Insets show the box plots for each group within each prediction type. **(a)** The agreement between the reference labels and our experts' decisions. **(b)** The agreement between experts themselves. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

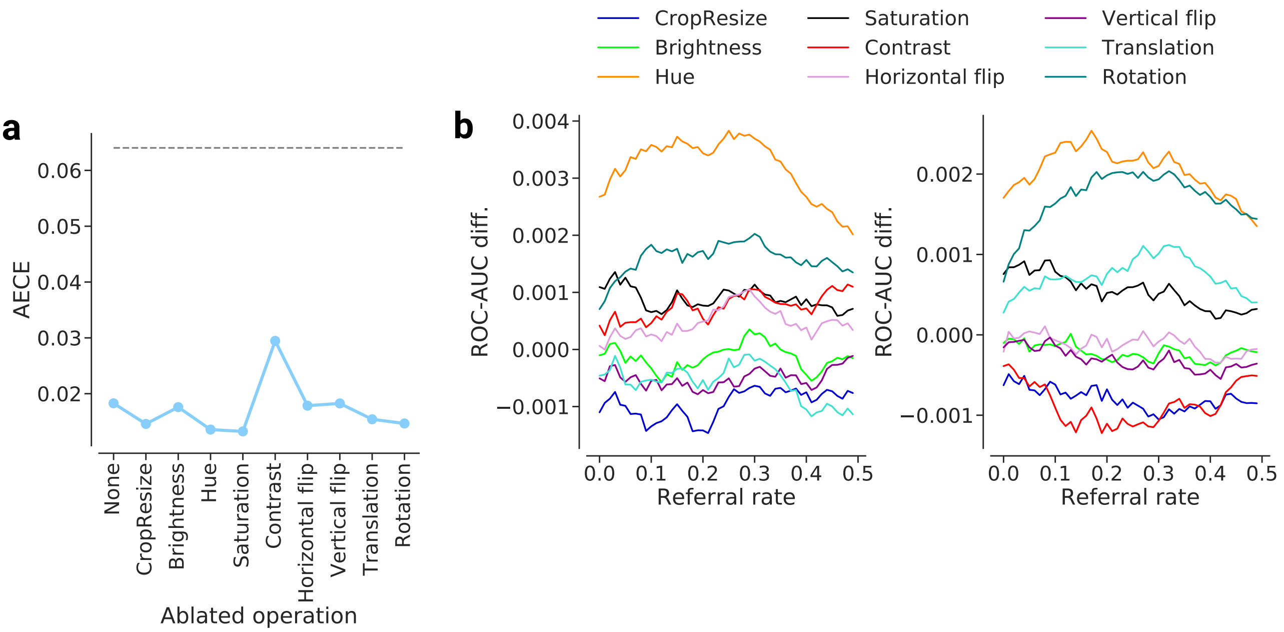

Fig. 9. Impact of data augmentation operations on calibration and uncertainty, based on the Kaggle DR test images and $T = 128$. **(a)**: Calibration error w.r.t. the ablation of operations from the data augmentation pipeline. Dashed line indicates the expected calibration error induced by the single point predictions. *None* corresponds to the fully equipped pipeline with no ablation. **(b)**: Difference in the decision referral performance after the ablation of operations from the complete data augmentation pipeline. *left*: onset 1, *right*: onset 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ter the inherent difficulties of trained ophthalmologists in grading a set of fundus images.

Finally, we conducted an ablation study to investigate the effects of data augmentation operations on the calibration quality and decision referral. The absence of the contrast adjustment led to an increase in the expected calibration error (Fig. 9a), whereas the removal of brightness, hue and saturation adjustments did not hinder calibration. The ablations of the geometric transformations surprisingly did not hamper calibration, either. Given that the convolutional neural networks are robust to translation, its inclusion in the augmentation pipeline could be seen redundant in the first place. However, the ablation of rotation did not disrupt the calibration quality, either, even though our network has no means for rotation invariance. Probably, the specified range of rotation was too narrow to yield a significant impact here. Likewise, vertical and horizontal flips also had almost no effect on calibration. In regards to decision referral (Fig. 9b), the most important operations were the contrast adjustment and zooming via cropping and resizing. Their removal led to highest drop in referral performance, especially for the onset 2 scenario (Fig. 9b, *right*). In the absence of hue and saturation adjustments as well as rotation and translation, decision referral performance slightly improved.

4. Discussion

Here, we used a state-of-the-art DNN for detecting DR from two well-known datasets with fundus images and evaluated a simple technique to quantify the predictive uncertainty of DNNs. We demonstrated its applicability and benefits for this disease detection scenario with a focus on diagnostic uncertainty and a *human-in-the-loop* mindset. Our network meets the requirements set forth for the detection of sight-threatening DR (Leibig et al., 2017; Younis et al., 2003) and our uncertainty measures are useful for inferring the DR cases that are difficult for the network to classify. Furthermore, we showed that images with high uncertainty are also difficult to grade for physicians with high disagreement between them. This indicates that these images are genuinely hard to classify, a crucial prerequisite for making such uncertainty estimates useful in practice. Also note that the proposed approach can be straightforwardly generalized to multiclass outputs since the softmax outputs can be taken as discrete probability distributions and one can compute the entropy to measure how concentrated the probability mass is on one of the classes.

Implementing artificial intelligence in medicine goes well beyond engineering challenges (Grote and Berens, 2019). To this end, a network must not only generalize well to new data sets in terms of predictive accuracy, but also with regards to the calibration of its predictive probabilities, such that they reflect the average probability to make a wrong decision. While our network met the first requirement, it did not naturally meet the second. Temperature scaling helped with calibration but it aggravated the referral performance in comparison to the overconfident single point predictions on the IDRiD data, which suggests that the achieved calibration is not able to support decision referral on out-of-distribution data. This is in line with recent work reporting that post-hoc calibration on the validation data via temperature scaling fails under mild-to-high distributional shifts (Ovadia et al., 2019). We attribute the poor performance by temperature scaling under data shift to the parametric nature of the method. In contrast, our TTAUG approach is non-parametric and adapts to data more flexibly. As a result, it offered better calibrated outputs and improved decision referral. The ensemble approach has been reported to be robust to out-of-distribution examples (Lakshminarayanan et al., 2017; Fort et al., 2019). In line, while not being quite as well calibrated as our method, it showed as good referral performance as our approach. We typically evaluated referral rates up to 40–50%, despite the fact that these are unrealistic in practice, since the goal of automated screening is to refer as few as possible patients to physicians. These high referral rates are still useful for investigating the limits of such measures of uncertainty. In a clinically realistic scenario with a low referral rate, of e.g., 10%, the median of the TTAUG predictive probabilities typically provided the best referral performance, especially considering the onset 2 scenario. Potentially, a large ensemble could be more accurate and better calibrated than a small one with only 3 networks. It could further benefit from our TTAUG strategy for space exploration. However, such attempts would significantly complicate the process in terms of both computation and practicality. The fact that our measures readily generalize well to new data highlights their utility.

In our study, fundus photographs of the central retina were evaluated for the presence (and severity) or absence of DR. In this setting, uncertainty in human grading could be caused by different circumstances, such as poor image quality, the restricted field of view, and unimodal imaging. First, if poor image quality results from technical errors, image acquisition and grading may be repeated within a given time frame. However, when both the DNN and experts are uncertain, the solution is not just to take a nicer picture, since we ensured that the images evaluated for this comparison are of sufficient quality. Rather, such images genuinely seem to form cases on the fuzzy border between disease stages, as also indicated by the tSNE analysis. The agreement in uncertainty between humans and DNNs highlights the necessity to very carefully evaluate such images with multiple human experts, and evaluate whether treatment strategies differ between the potential outcomes of the uncertain decision. Alternatively, experts could have been able to classify these images correctly with high agreement, akin to the ability of humans to classify adversarial examples correctly. In this case, principled solutions, e.g., that of Meinke and Hein (2019), probably along with a reject option (Geifman and El-Yaniv, 2019), integrated into network architectures would be needed in order to increase the robustness of DNNs. If poor image quality results from media opacities (e.g. cataract, vitreous haemorrhage), the patient needs to be referred to an ophthalmologist for further examinations and possible therapeutic actions. Second, in this particular study, the grading was performed by evaluating (just) the fundus image of the posterior pole. DR grading, however, typically takes into account a larger field of view. Capturing a larger field of view, or, if this is not an option, referring the patient to an ophthalmologist for further examinations, will thus likely increase the certainty of the assessment. Third, when judging for the presence and severity of DR, clinicians rely on not only fundus examination but also additional assessments (mainly visual acuity testing, slit lamp examination, funduscopy of not just the posterior pole but also the mid-peripheral and peripheral retina, fluorescein angiography, and/or optical coherence tomography). Thus, referring the patient to an ophthalmologist for further examinations again will likely increase the certainty of the grading. Concretely, these clinical aspects highlight the importance of obtaining a complete diagnostic picture. In this respect, keeping the human in the loop for automated screening via a realistic referral rate is a crucial step towards improved diagnosis through the cooperation of man and AI systems.

Our intuitive and data-driven TTAUG approach yielded a general purpose solution for obtaining well-calibrated, accurate and generalizing uncertainty estimates without requiring substantial changes to the standard training or inference pipelines. In fact, data augmentation is frequently used in training of diagnostic DNNs. Given the variety of transformation operations readily available in many deep learning frameworks or via third party packages, one can flexibly design new data augmentation pipelines and tailor them to the need of new domains and tasks, regardless of the architectural design choices or regularization methods. Of course, this comes at the computational cost of needing to compute several forward passes through the network for each augmented image. In practice, we found the augmentation operations with regards to contrast adjustments and zooming (via cropping and resizing) to be most important. While these operations are also crucial to the visual inspection by clinicians, data augmentation policies could be ideally learned from data in order to find the most suitable combination for a given task using frameworks such as AutoAugment (Cubuk et al., 2019). Such optimal pipelines may achieve better performance with smaller referral rates and improve the utility of DNNs' predictive uncertainty.

4.1. Outlook

This study demonstrates the effectiveness of TTAUG in capturing the predictive uncertainty of DNNs and its concordance with the uncertainty in clinicians' decisions on a well-known diagnostic task. However, the procedure only works on the outputs of a network which itself is not capable of exploiting the estimated uncertainty. By assuming the predictive uncertainty of DNNs as a proxy for interobserver variability and integrating it into training, we may emulate the transfer of human grader variability to DNNs and facilitate the cooperation of man and machine in the form of assisted reading (Sayres et al., 2019). The selective classification framework (Geifman and El-Yaniv, 2017; 2019) may provide a starting point for this. This approach could be extended to visualization (Arcadu et al., 2019; Carson Lam et al., 2018; Dai et al., 2018; Khojasteh et al., 2018; Quellec et al., 2017) efforts in the hope of improving the interpretability of DNN decisions in medical settings.

Author contributions statement

MSA and PB designed research; MSA performed research; LK, GA, WI and FZ provided medical advice and graded images; PB supervised research; MSA and PB wrote the paper with input from all authors.

Declaration of Competing Interest

None.

Acknowledgements

We thank Christian Leibig for many useful discussions, helpful suggestions as well as providing the preprocessed Kaggle DR images. This research was supported by the German Federal Ministry of Science and Education (Bundesministerium für Bildung und Forschung) (BMBF, 01GQ1601 and 01IS18039A) and the German Science Foundation (Deutsche Forschungsgemeinschaft) (BE5601/4-1 and EXC 2064, project number 390727645).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2020.101724.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems.

Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N., Folk, J.C., 2018. Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digital Med. 1 (1), 39.

Arcadu, F., Benmansour, F., Maunz, A., Michon, J., Haskova, Z., McClintock, D., Adamis, A.P., Willis, J.R., Prunotto, M., 2019. Deep learning predicts oct measures of diabetic macular thickening from color fundus photographs. Invest. Ophthalmol. Vis. Sci. 60 (4), 852–857.

Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D.P., Shetty, S., 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med..

Ayhan, M.S., Berens, P., 2018. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In: Proceedings of the International Conference on Medical Imaging with Deep Learning.

Bhise, V., Rajan, S.S., Sittig, D.F., Morgan, R.O., Chaudhary, P., Singh, H., 2018. Defining and measuring diagnostic uncertainty in medicine: a systematic review. J. Gen. Intern. Med. 33 (1), 103–115.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Carson Lam, D.Y., Guo, M., Lindsey, T., 2018. Automated detection of diabetic retinopathy using deep learning. In: AMIA Summits on Translational Science Proceedings, 2017, p. 147.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.

Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V., 2019. Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 113–123.

Dai, L., Fang, R., Li, H., Hou, X., Sheng, B., Wu, Q., Jia, W., 2018. Clinical report guided retinal microaneurysm detection with multi-sieving deep learning. IEEE Trans. Med. Imaging 37 (5), 1149–1161.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 24 (9), 1342.

Ding, Y., Liu, J., Xiong, J., Shi, Y., 2019. Evaluation of neural network uncertainty estimation with application to resource-constrained platforms. arXiv:1903.02050.

Dong, X., Kang, G., Zhan, K., Yang, Y., 2017. Eraserelu: a simple way to ease the training of deep convolution neural networks. arXiv:1709.07634.

El-Yaniv, R., Wiener, Y., 2010. On the foundations of noise-free selective classification. J. Mach. Learn. Res. 11 (May), 1605–1641.

Elmore, J.G., Longton, G.M., Carney, P.A., Geller, B.M., Onega, T., Tosteson, A.N.A., Nelson, H.D., Pepe, M.S., Allison, K.H., Schnitt, S.J., O'Malley, F.P., Weaver, D.L., 2015. Diagnostic concordance among pathologists interpreting breast biopsy specimensdiagnostic concordance in interpreting breast biopsiesdiagnostic concordance in interpreting breast biopsies. JAMA 313 (11), 1122–1132.

Elmore, J.G., Wells, C.K., Lee, C.H., Howard, D.H., Feinstein, A.R., 1994. Variability in radiologists' interpretations of mammograms. N. Engl. J. Med. 331 (22), 1493–1499.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. Nat. Med. 25 (1), 24.

Eulenberg, P., Köhler, N., Blasi, T., Filby, A., Carpenter, A.E., Rees, P., Theis, F.J., Wolf, F.A., 2017. Reconstructing cell cycle and disease progression using deep learning. Nat. Commun. 8 (1), 463.

FDA News,. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm. Accessed: 2019-03-21.

Fort, S., Hu, H., Lakshminarayanan, B., 2019. Deep ensembles: a loss landscape perspective. arXiv:1912.02757.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: international Conference on Machine Learning, pp. 1050–1059.

Geifman, Y., El-Yaniv, R., 2017. Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems, pp. 4878–4887.

Geifman, Y., El-Yaniv, R., 2019. SelectiveNet: a deep neural network with an integrated reject option. arXiv:1901.09192.

Goodfellow, I. J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K., 2017. Accurate, large minibatch sgd: training imagenet in 1 hour. arXiv:1706.02677.

Grote, T., Berens, P., 2019. On the ethics of algorithmic decision-making in healthcare. J. Med. Ethics.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316 (22), 2402–2410.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 1321–1330.

Haenssle, H., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A., et al., 2018. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann. Oncol. 29 (8), 1836–1842.

Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., Ng, A.Y., 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat. Med. 25 (1), 65.

Hasler, D., Suesstrunk, S.E., 2003. Measuring colorfulness in natural images. In: Human Vision and Electronic Imaging VIII, 5007. International Society for Optics and Photonics, pp. 87–96.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. In: European Conference on Computer Vision. Springer, pp. 630–645.

Hendrycks, D., Gimpel, K., 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Proceedings of International Conference on Learning Representations.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Int. Council of Ophth.,. The international council of ophthalmology (ICO) guidelines for diabetic eye care. http://www.icoph.org/downloads/ICOGuidelinesforDiabeticEyeCare.pdf. Accessed: 2019-05-28.

Ioffe, S., 2017. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In: Advances in Neural Information Processing Systems, pp. 1942–1950.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456.

Iqbal, H., 2018. PlotNeuralNet. Accessed: 2019-05-20.

Kaggle.com, 2015. Kaggle competition on diabetic retinopathy detection. https://www.kaggle.com/c/diabetic-retinopathy-detection. Accessed: 2019-07-07.

Kanagasingam, Y., Xiao, D., Vignarajan, J., Preetham, A., Tay-Kearney, M.-L., Mehrotra, A., 2018. Evaluation of artificial intelligence–based grading of diabetic retinopathy in primary care. JAMA Netw. Open 1 (5). e182665–e182665

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems, pp. 5580–5590.

Khojasteh, P., Aliahmad, B., Kumar, D.K., 2018. Fundus images analysis using deep features for detection of exudates, hemorrhages and microaneurysms. BMC Ophthalmol. 18 (1), 288.

Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., Yu, Y., Langlotz, C.P., Ball, R.L., Montine, T.J., et al., 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digital Med. 3 (1), 1–8.

Kobak, D., Berens, P., 2019. The art of using t-SNE for single-cell transcriptomics. Nat. Commun. 10 (1), 1–14.

Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G.S., Peng, L., Webster, D.R., 2018. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 125 (8), 1264–1272.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, pp. 6405–6416.

Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S., 2017. Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. 7 (1), 17816.

Liang, S., Li, Y., Srikant, R., 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In: Proceedings of International Conference on Learning Representations.

Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S., Kluger, Y., 2019. Fast interpolation-based t-SNE for improved visualization of single-cell rna-seq data. Nat. Methods 16 (3), 243.

Loshchilov, I., Hutter, F., 2016. Sgdr: stochastic gradient descent with warm restarts. arXiv:1608.03983.

Luo, P., Wang, X., Shao, W., Peng, Z., 2019. Towards understanding regularization in batch normalization. In: International Conference on Learning Representations.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605.

Malinin, A., Gales, M., 2018. Predictive uncertainty estimation via prior networks. In: Advances in Neural Information Processing Systems, pp. 7047–7058.

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.C., Darzi, A., et al., 2020. International evaluation of an ai system for breast cancer screening. Nature 577 (7788), 89–94.

Meinke, A., Hein, M., 2019. Towards neural networks that provably know when they don't know. arXiv:1909.12180.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.

Neal, R.M., 2012. Bayesian Learning for Neural Networks, 118. Springer Science & Business Media.

Niculescu-Mizil, A., Caruana, R., 2005. Predicting good probabilities with supervised learning. In: Proceedings of the 22Nd International Conference on Machine Learning. ACM, New York, NY, USA, pp. 625–632.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. arXiv:1906.02530.

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F., 2018. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. Data 3 (3), 25.

Quellec, G., Charrière, K., Boudi, Y., Cochener, B., Lamard, M., 2017. Deep image mining for diabetic retinopathy screening. Med. Image Anal. 39, 178–193.

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., et al., 2019. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology 126 (4), 552–564.

Smith, L., Gal, Y., 2018. Understanding measures of uncertainty for adversarial example detection. arXiv:1803.08533.

Sussman, E.J., Tsiaras, W.G., Soper, K.A., 1982. Diagnosis of diabetic eye disease. JAMA 247 (23), 3231–3234.

Sutskever, I., Martens, J., Dahl, G.E., Hinton, G.E., 2013. On the importance of initialization and momentum in deep learning.. ICML (3) 28 (1139–1147), 5.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence.

Teye, M., Azizpour, H., Smith, K., 2018. Bayesian uncertainty estimation for batch normalized deep networks. In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. PMLR, Stockholmsmäsan, Stockholm Sweden, pp. 4907–4916.

Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 25 (1), 44.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., Schön, T., 2019. Evaluating model calibration in classification. In: Chaudhuri, K., Sugiyama, M. (Eds.), Proceedings of Machine Learning Research. PMLR, pp. 3459–3467.

Verbraak, F.D., Abramoff, M.D., Bausch, G.C., Klaver, C., Nijpels, G., Schlingemann, R.O., van der Heijden, A.A., 2019. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. Diabetes Care.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing.

Wu, L., Fernandez-Loaiza, P., Sauma, J., Hernandez-Bogantes, E., Masis, M., 2013. Classification of diabetic retinopathy and diabetic macular edema. World J. Diabetes 4 (6), 290.

Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Younis, N., Broadbent, D.M., Harding, S.P., Vora, J.P., 2003. Incidence of sight-threatening retinopathy in type 1 diabetes in a systematic screening programme. Diabetic Med. 20 (9), 758–765.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2017. Understanding deep learning requires rethinking generalization. In: International Conference on Learning Representations.