Uncertainty Estimation in Deep Neural Networks for Dermoscopic Image Classification

Marc Combalia, Ferran Hueto, Susana Puig, Josep Malvehy, Veronica Vilaplana

CLÍNIC
BARCELONA
Hospital Universitari

UPC

Introduction

Neural Networks in HealthCare

- High performance of AI in HealthCare
- Real World Implementations are still scarce…
- Why? One of the reasons…
  - **Uncertainty**: current neural networks produce point estimates, and don’t give any measure of confidence of the prediction.

Uncertainty

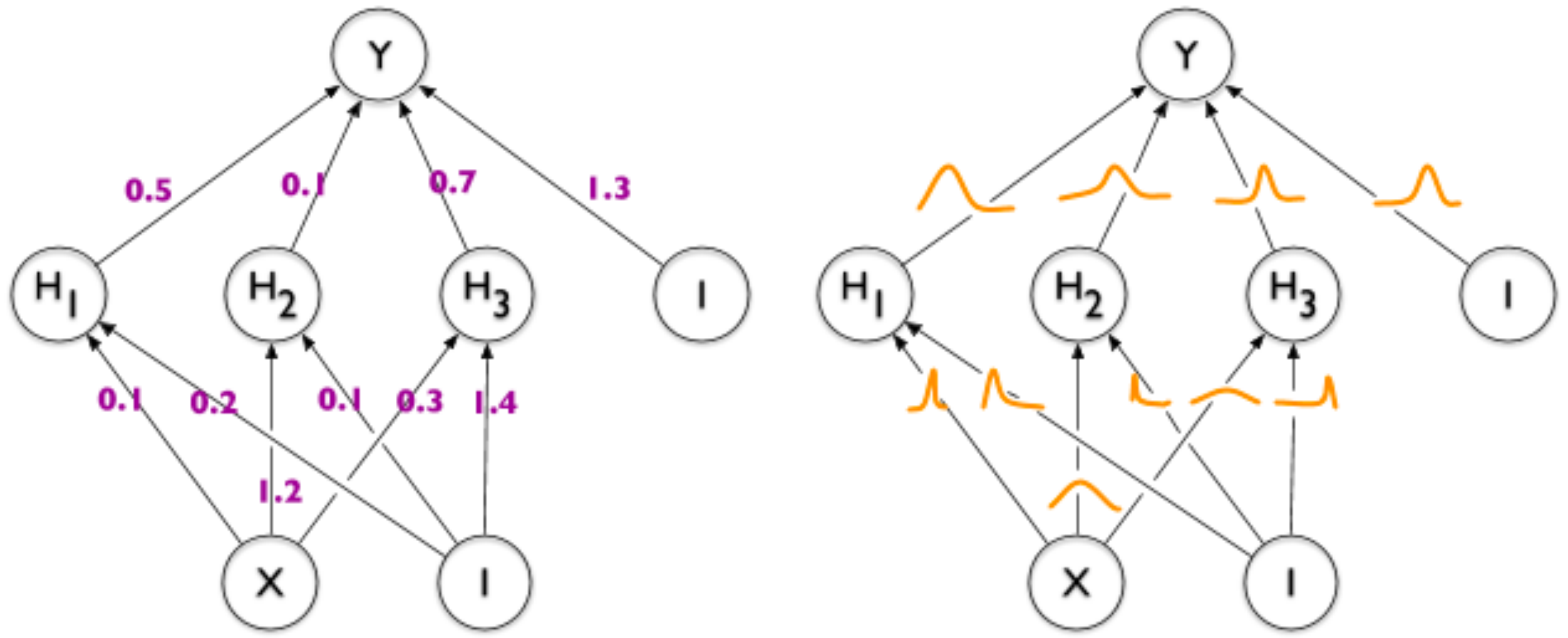
- **Epistemic Uncertainty**
  - Epistemic uncertainty or model uncertainty captures the uncertainty in the model parameters.
- **Aleatoric Uncertainty**
  - Aleatoric uncertainty is described by the noise in the observations; it is the input-dependent un- certainty.

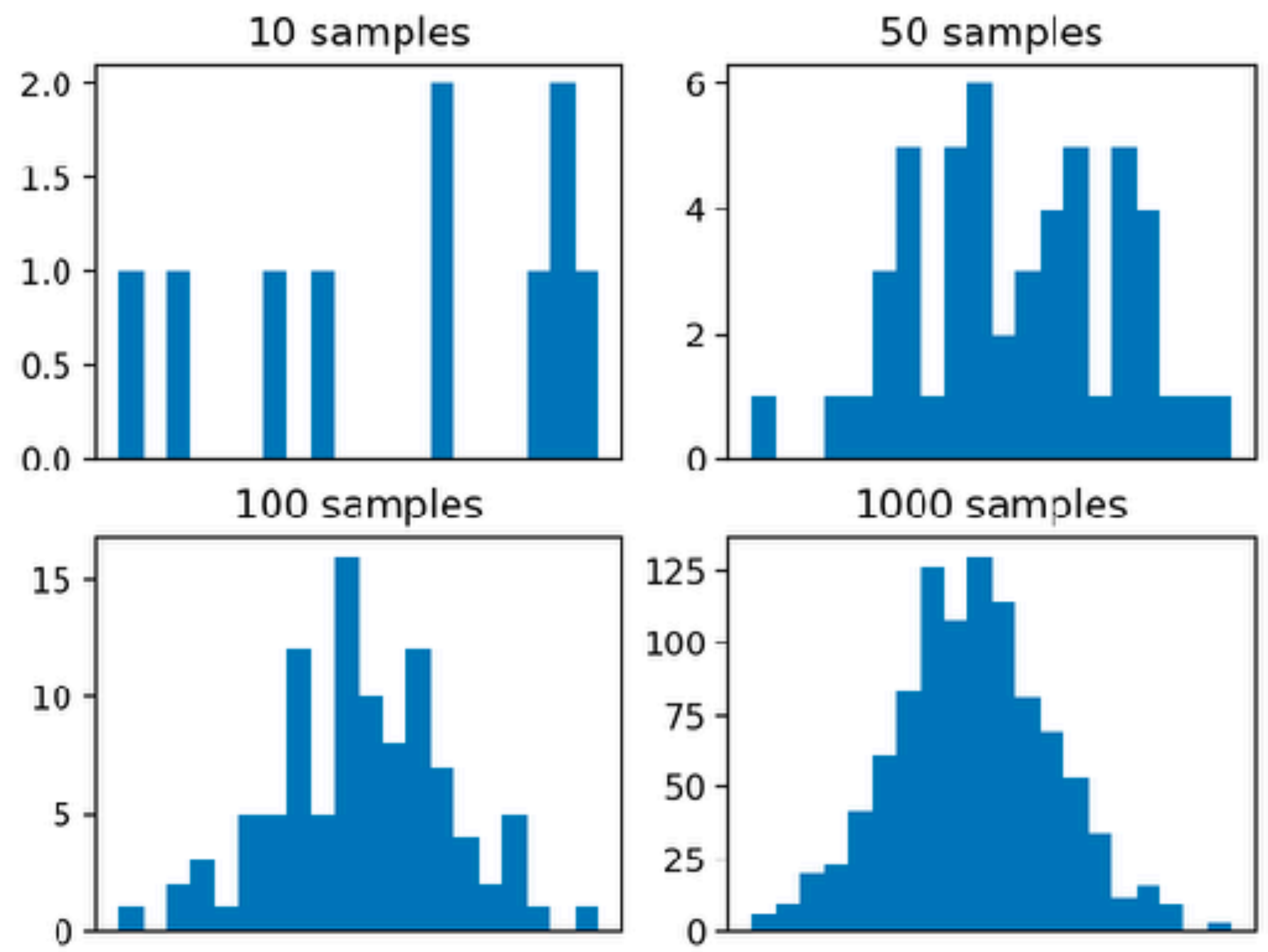
Methods

Bayesian Neural Networks

- Uncertainties are formalized as probability distributions over the model parameters (for epistemic uncertainty) or model inputs (for aleatoric uncertainty)

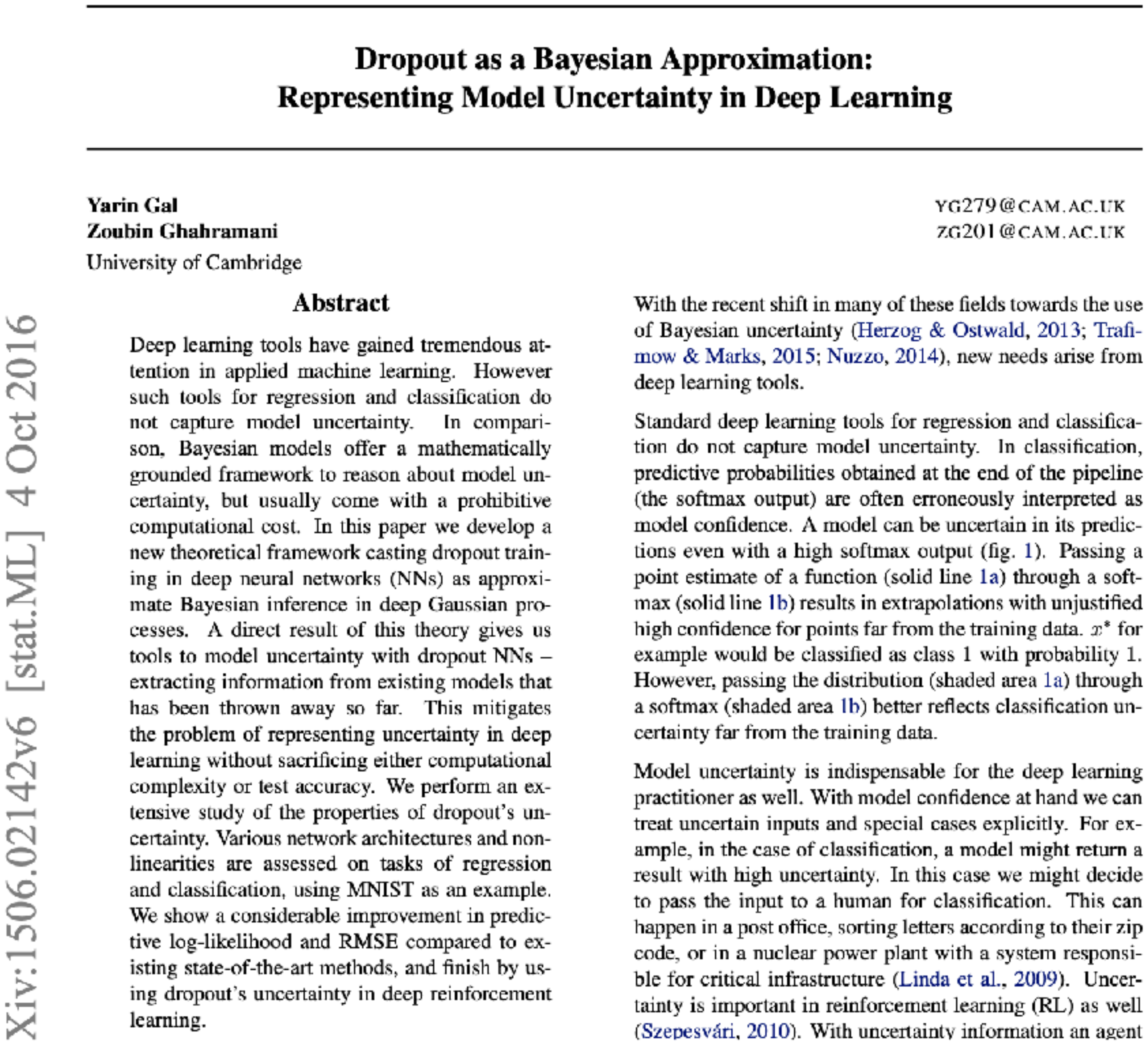
- But how can we estimate the probability distributions?
  - **MONTECARLO SAMPLING**





Epistemic Uncertainty Estimation

- MonteCarlo Dropout
  - When you want to estimate using MonteCarlo dropout, you sample using a “helper” distribution (generally bayesian / uniform…)
  - MonteCarlo Dropout can be seen as sampling the parameters of the NN with a Binomial Distribution.



Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning

Yarin Gal YG279@CAM.AC.UK
Zoubin Ghahramani ZG201@CAM.AC.UK
University of Cambridge

Xiv:1506.02142v6 [stat.ML] 4 Oct 2016

Abstract

Deep learning tools have gained tremendous attention in applied machine learning. However such tools for regression and classification do not capture model uncertainty. In comparison, Bayesian models offer a mathematically grounded framework to reason about model uncertainty, but usually come with a prohibitive computational cost. In this paper we develop a new theoretical framework casting dropout training in deep neural networks (NNs) as approximate Bayesian inference in deep Gaussian processes. A direct result of this theory gives us tools to model uncertainty with dropout NNs – extracting information from existing models that has been thrown away so far. This mitigates the problem of representing uncertainty in deep learning without sacrificing either computational complexity or test accuracy. We perform an extensive study of the properties of dropout's uncertainty. Various network architectures and non-linearities are assessed on tasks of regression and classification, using MNIST as an example. We show a considerable improvement in predictive log-likelihood and RMSE compared to existing state-of-the-art methods, and finish by using dropout's uncertainty in deep reinforcement learning.

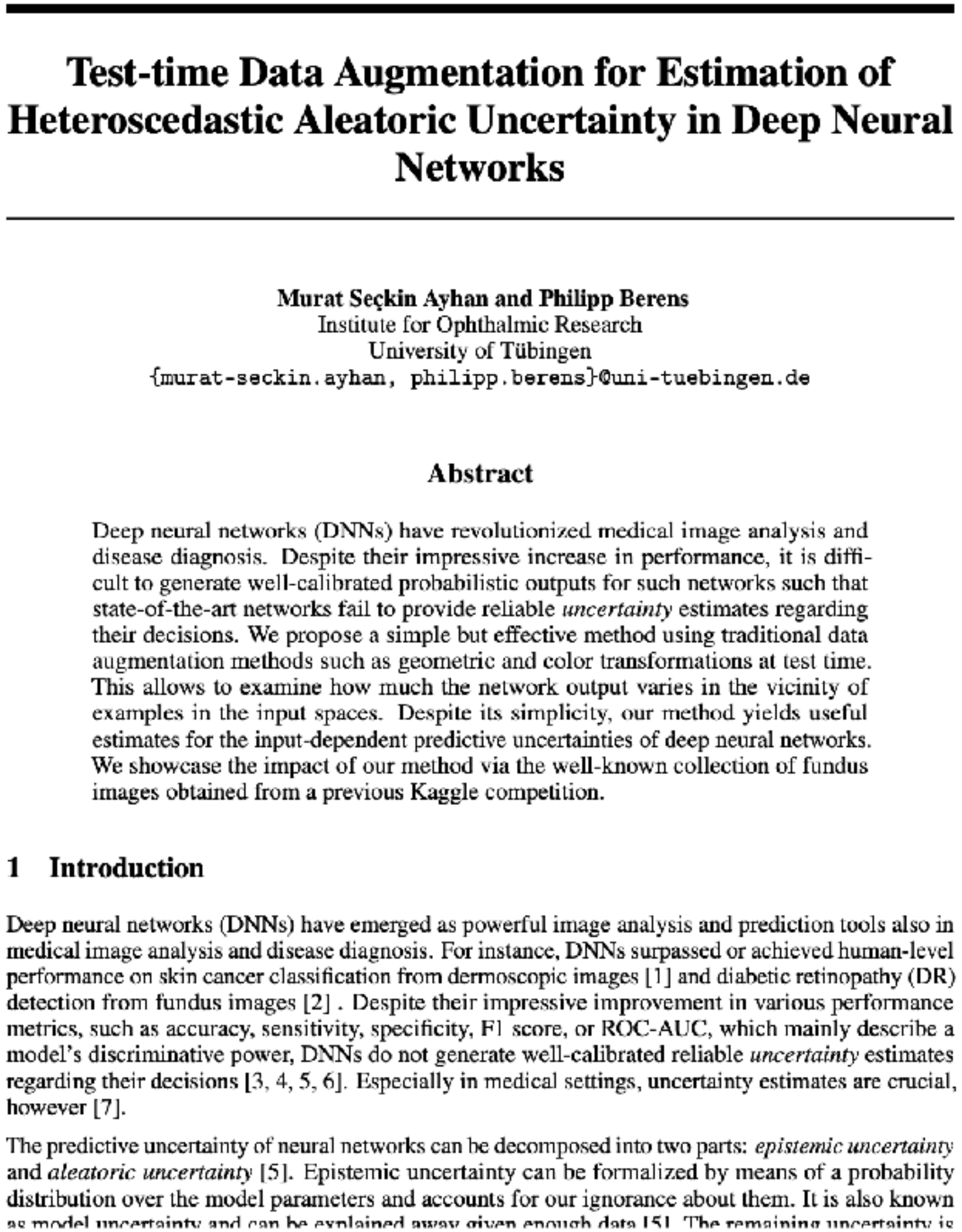
With the recent shift in many of these fields towards the use of Bayesian uncertainty (Herzog & Ostwald, 2013; Trafimow & Marks, 2015; Nuzzo, 2014), new needs arise from deep learning tools.

Standard deep learning tools for regression and classification do not capture model uncertainty. In classification, predictive probabilities obtained at the end of the pipeline (the softmax output) are often erroneously interpreted as model confidence. A model can be uncertain in its predictions even with a high softmax output (fig. 1). Passing a point estimate of a function (solid line 1a) through a softmax (solid line 1b) results in extrapolations with unjustified high confidence for points far from the training data. x^* for example would be classified as class 1 with probability 1. However, passing the distribution (shaded area 1a) through a softmax (shaded area 1b) better reflects classification uncertainty far from the training data.

Model uncertainty is indispensable for the deep learning practitioner as well. With model confidence at hand we can treat uncertain inputs and special cases explicitly. For example, in the case of classification, a model might return a result with high uncertainty. In this case we might decide to pass the input to a human for classification. This can happen in a post office, sorting letters according to their zip code, or in a nuclear power plant with a system responsible for critical infrastructure (Linda et al., 2009). Uncertainty is important in reinforcement learning (RL) as well (Szepesvári, 2010). With uncertainty information an agent

Aleatoric Uncertainty Estimation

- MonteCarlo Sampling.. but from capturing parameters
  - We already know that! **Data Augmentation!**
  - Sampling the data with a priori random distribution over capturing parameters (rotation, translation, color, …)



Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks

Murat Seçkin Ayhan and Philipp Berens
Institute for Ophthalmic Research
University of Tübingen
{murat-seckin.ayhan, philipp.berens}@uni-tuebingen.de

Abstract

Deep neural networks (DNNs) have revolutionized medical image analysis and disease diagnosis. Despite their impressive increase in performance, it is difficult to generate well-calibrated probabilistic outputs for such networks such that state-of-the-art networks fail to provide reliable *uncertainty* estimates regarding their decisions. We propose a simple but effective method using traditional data augmentation methods such as geometric and color transformations at test time. This allows to examine how much the network output varies in the vicinity of examples in the input spaces. Despite its simplicity, our method yields useful estimates for the input-dependent predictive uncertainties of deep neural networks. We showcase the impact of our method via the well-known collection of fundus images obtained from a previous Kaggle competition.

1 Introduction

Deep neural networks (DNNs) have emerged as powerful image analysis and prediction tools also in medical image analysis and disease diagnosis. For instance, DNNs surpassed or achieved human-level performance on skin cancer classification from dermoscopic images [1] and diabetic retinopathy (DR) detection from fundus images [2] . Despite their impressive improvement in various performance metrics, such as accuracy, sensitivity, specificity, F1 score, or ROC-AUC, which mainly describe a model's discriminative power, DNNs do not generate well-calibrated reliable *uncertainty* estimates regarding their decisions [3, 4, 5, 6]. Especially in medical settings, uncertainty estimates are crucial, however [7].

The predictive uncertainty of neural networks can be decomposed into two parts: *epistemic uncertainty* and *aleatoric uncertainty* [5]. 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 [5]. The remaining uncertainty is

Uncertainty Aggretation Metrics

- Prediction Entropy

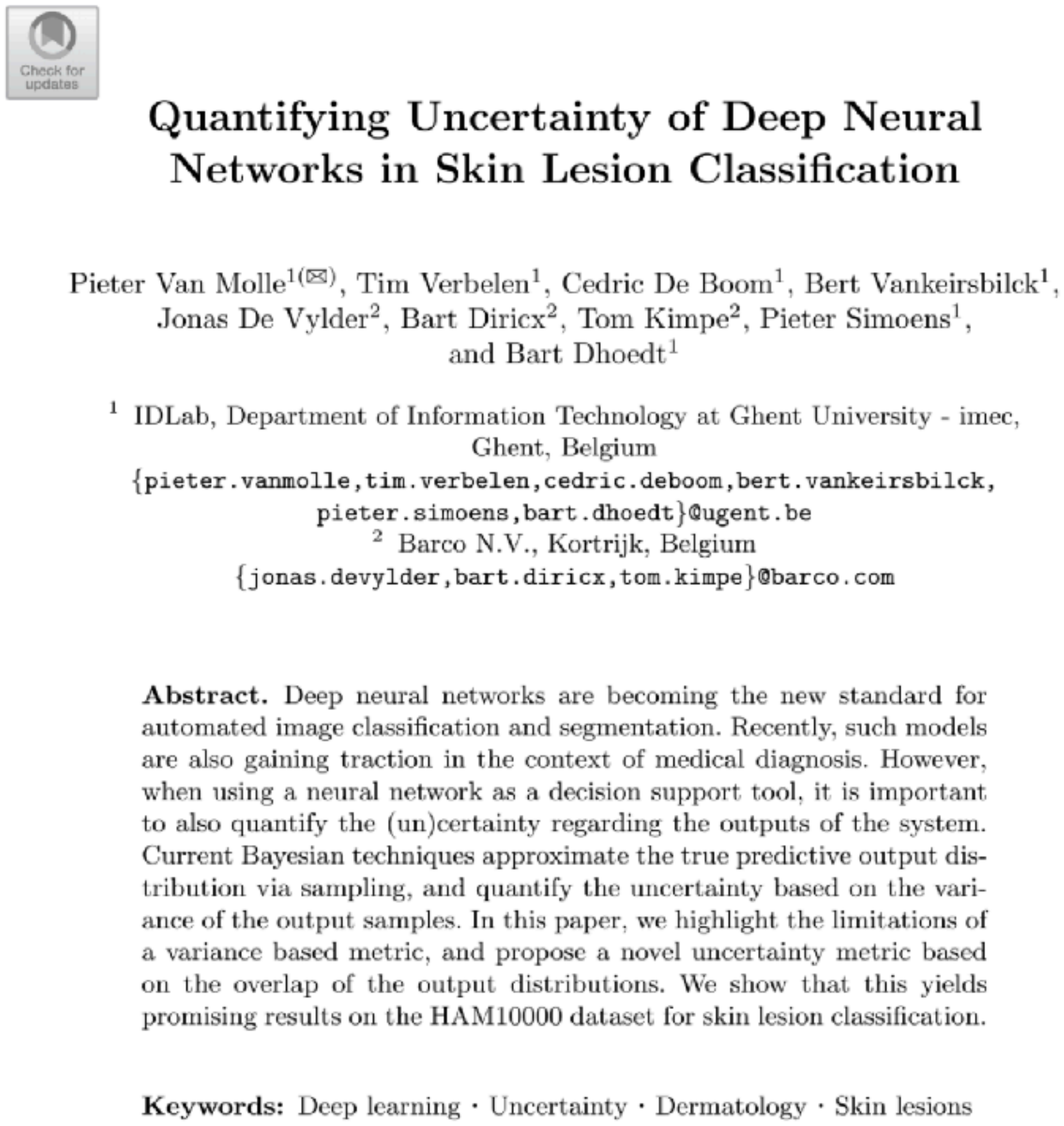
$$H(\mathbf{p}_T(x)) = -\sum_{c=1}^{C} p_T(x)[c]\log(p_T(x)[c])$$

- Prediction Variance

$$\sigma^2(\mathbf{p}_T(x)) = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{p}_t(x) - \mathbf{p}_T(X))^2$$

- Bhattacharyya Coefficient

$$BC(h_{c1}, h_{c2})(x) = \sum_{n=1}^{N}\sqrt{h_{c1}[n] * h_{c2}[n]}$$



Quantifying Uncertainty of Deep Neural Networks in Skin Lesion Classification

Pieter Van Molle[1(✉)], Tim Verbelen[1], Cedric De Boom[1], Bert Vankeirsbilck[1],
Jonas De Vylder[2], Bart Diricx[2], Tom Kimpe[2], Pieter Simoens[1],
and Bart Dhoedt[1]

[1] IDLab, Department of Information Technology at Ghent University - imec, Ghent, Belgium
{pieter.vanmolle,tim.verbelen,cedric.deboom,bert.vankeirsbilck, pieter.simoens,bart.dhoedt}@ugent.be
[2] Barco N.V., Kortrijk, Belgium
{jonas.devylder,bart.diricx,tom.kimpe}@barco.com

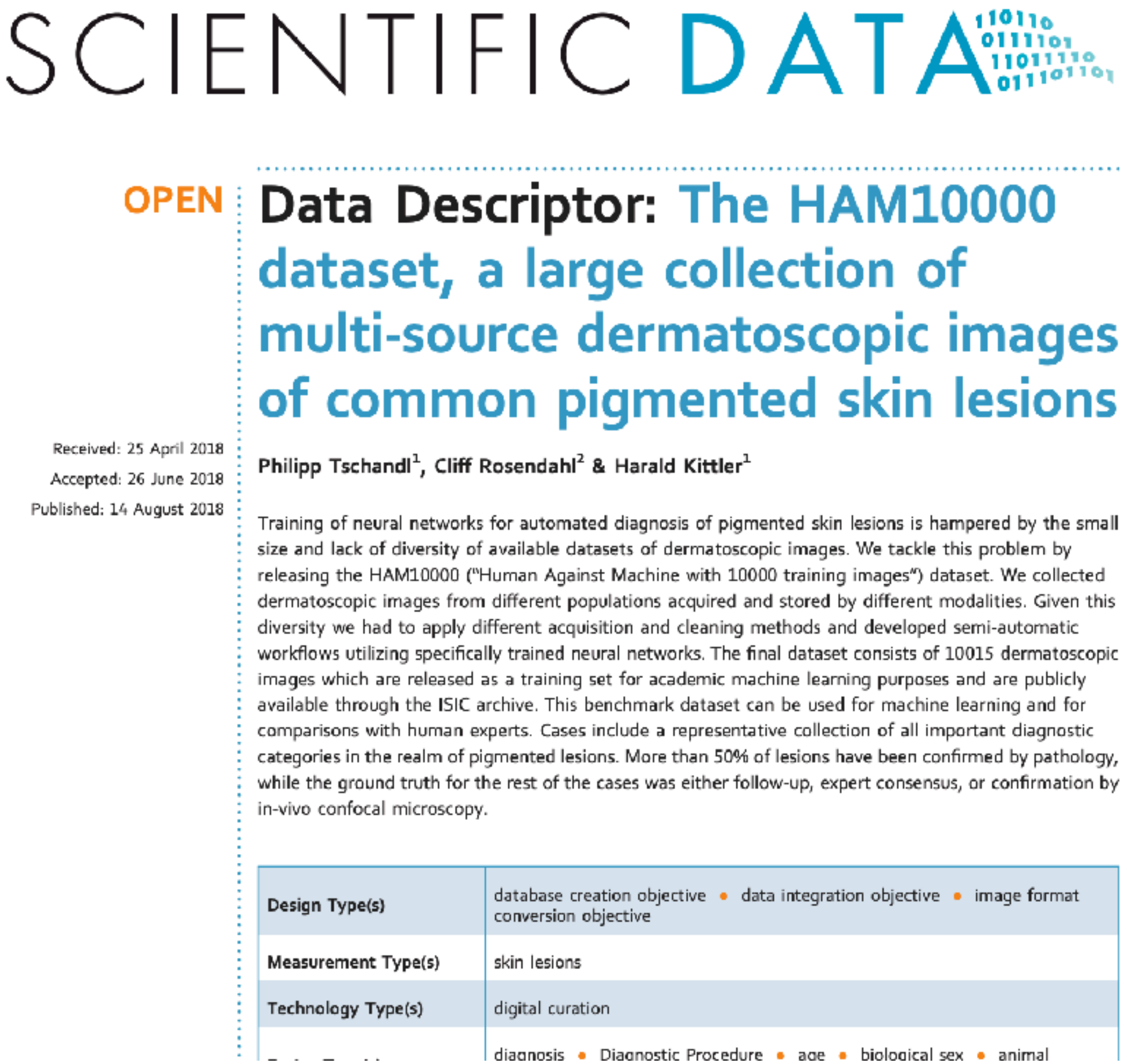
Abstract. Deep neural networks are becoming the new standard for automated image classification and segmentation. Recently, such models are also gaining traction in the context of medical diagnosis. However, when using a neural network as a decision support tool, it is important to also quantify the (un)certainty regarding the outputs of the system. Current Bayesian techniques approximate the true predictive output distribution via sampling, and quantify the uncertainty based on the variance of the output samples. In this paper, we highlight the limitations of a variance based metric, and propose a novel uncertainty metric based on the overlap of the output distributions. We show that this yields promising results on the HAM10000 dataset for skin lesion classification.

Keywords: Deep learning · Uncertainty · Dermatology · Skin lesions

Materials

ISIC Challenge 2018

- This dataset is composed of 10,015 dermoscopic images corresponding to 7,470 skin lesions.
- Each image is paired with its corresponding label indicating the lesion diagnosis and other metadata surrounding the lesion and the patient.
- The test dataset of the ISIC 2018 Challenge contains 1512 images that the participants are asked to classify in their submission file.



SCIENTIFIC DATA

OPEN

Received: 25 April 2018
Accepted: 26 June 2018
Published: 14 August 2018

Data Descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions

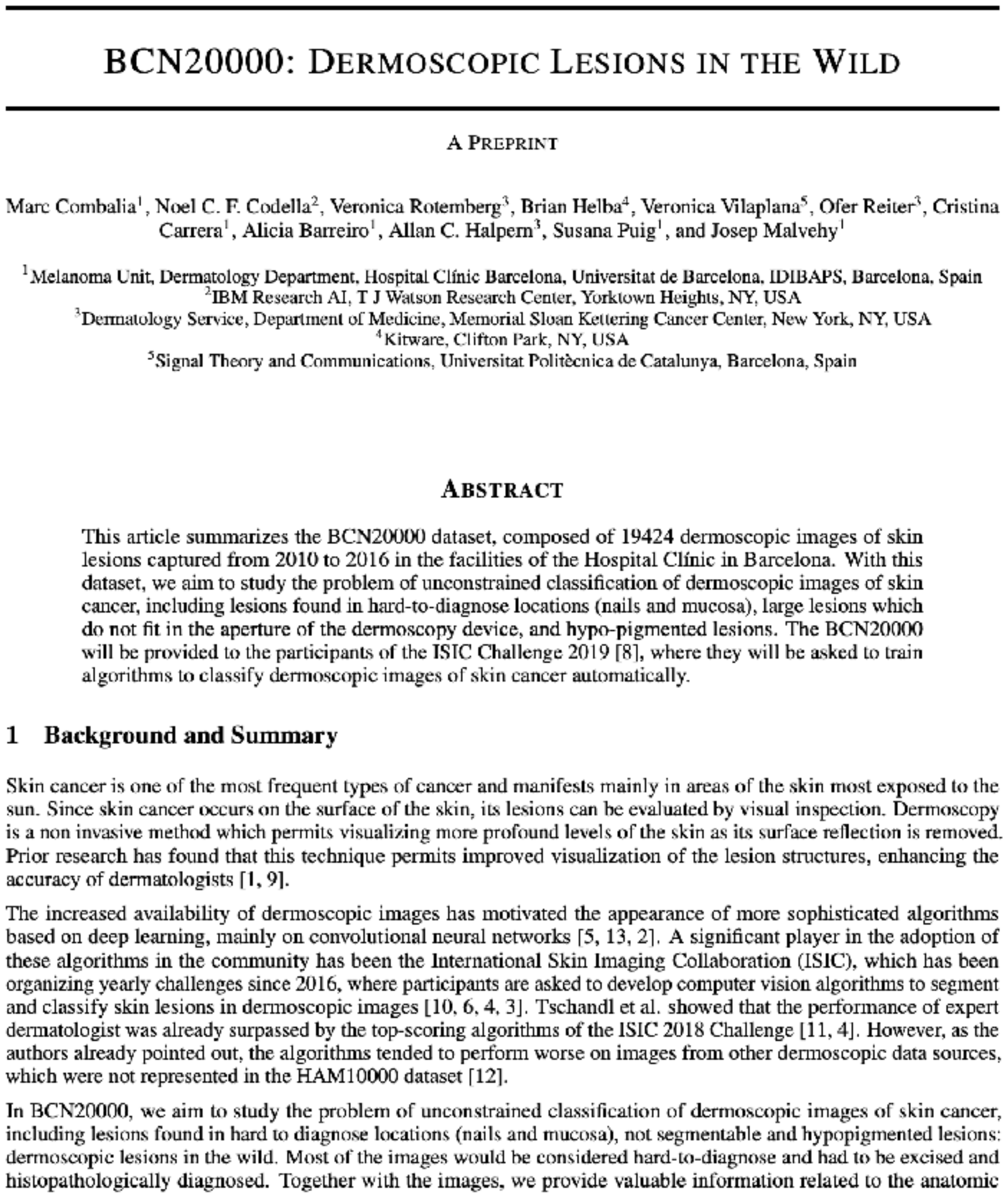
Philipp Tschandl[1], Cliff Rosendahl[2] & Harald Kittler[1]

Training of neural networks for automated diagnosis of pigmented skin lesions is hampered by the small size and lack of diversity of available datasets of dermatoscopic images. We tackle this problem by releasing the HAM10000 ("Human Against Machine with 10000 training images") dataset. We collected dermatoscopic images from different populations acquired and stored by different modalities. Given this diversity we had to apply different acquisition and cleaning methods and developed semi-automatic workflows utilizing specifically trained neural networks. The final dataset consists of 10015 dermatoscopic images which are released as a training set for academic machine learning purposes and are publicly available through the ISIC archive. This benchmark dataset can be used for machine learning and for comparisons with human experts. Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions. More than 50% of lesions have been confirmed by pathology, while the ground truth for the rest of the cases was either follow-up, expert consensus, or confirmation by in-vivo confocal microscopy.

Design Type(s)	database creation objective • data integration objective • image format conversion objective
Measurement Type(s)	skin lesions
Technology Type(s)	digital curation
	diagnosis • Diagnostic Procedure • age • biological sex • animal

ISIC Challenge 2019

- The training dataset of the ISIC Challenge 2019 consists of 25331 dermoscopic images.
- Eight diagnostic categories: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma.
- This dataset includes all the images from the HAM10000 dataset, and also adds images from the BCN20000 dataset and the MSK dataset.
- The BCN20000 dataset is considered to be remarkably complex since it includes uncurated images from day to day clinical practice.
- The test dataset from the ISIC Challenge 2019 consists of 8238 images and includes a set of images that are not contained in the diagnostic categories provided in the train- ing split (Unknown category)



BCN20000: Dermoscopic Lesions in the Wild

A Preprint

Marc Combalia[1], Noel C. F. Codella[2], Veronica Rotemberg[3], Brian Helba[4], Veronica Vilaplana[5], Ofer Reiter[3], Cristina Carrera[1], Alicia Barreiro[1], Allan C. Halpern[3], Susana Puig[1], and Josep Malvehy[1]

[1]Melanoma Unit, Dermatology Department, Hospital Clínic Barcelona, Universitat de Barcelona, IDIBAPS, Barcelona, Spain
[2]IBM Research AI, T J Watson Research Center, Yorktown Heights, NY, USA
[3]Dermatology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[4]Kitware, Clifton Park, NY, USA
[5]Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, Spain

Abstract

This article summarizes the BCN20000 dataset, composed of 19424 dermoscopic images of skin lesions captured from 2010 to 2016 in the facilities of the Hospital Clínic in Barcelona. With this dataset, we aim to study the problem of unconstrained classification of dermoscopic images of skin cancer, including lesions found in hard-to-diagnose locations (nails and mucosa), large lesions which do not fit in the aperture of the dermoscopy device, and hypo-pigmented lesions. The BCN20000 will be provided to the participants of the ISIC Challenge 2019 [8], where they will be asked to train algorithms to classify dermoscopic images of skin cancer automatically.

1 Background and Summary

Skin cancer is one of the most frequent types of cancer and manifests mainly in areas of the skin most exposed to the sun. Since skin cancer occurs on the surface of the skin, its lesions can be evaluated by visual inspection. Dermoscopy is a non invasive method which permits visualizing more profound levels of the skin as its surface reflection is removed. Prior research has found that this technique permits improved visualization of the lesion structures, enhancing the accuracy of dermatologists [1, 9].

The increased availability of dermoscopic images has motivated the appearance of more sophisticated algorithms based on deep learning, mainly on convolutional neural networks [5, 13, 2]. A significant player in the adoption of these algorithms in the community has been the International Skin Imaging Collaboration (ISIC), which has been organizing yearly challenges since 2016, where participants are asked to develop computer vision algorithms to segment and classify skin lesions in dermoscopic images [10, 6, 4, 3]. Tschandl et al. showed that the performance of expert dermatologist was already surpassed by the top-scoring algorithms of the ISIC 2018 Challenge [11, 4]. However, as the authors already pointed out, the algorithms tended to perform worse on images from other dermoscopic data sources, which were not represented in the HAM10000 dataset [12].

In BCN20000, we aim to study the problem of unconstrained classification of dermoscopic images of skin cancer, including lesions found in hard to diagnose locations (nails and mucosa), not segmentable and hypopigmented lesions: dermoscopic lesions in the wild. Most of the images would be considered hard-to-diagnose and had to be excised and histopathologically diagnosed. Together with the images, we provide valuable information related to the anatomic

Experiments

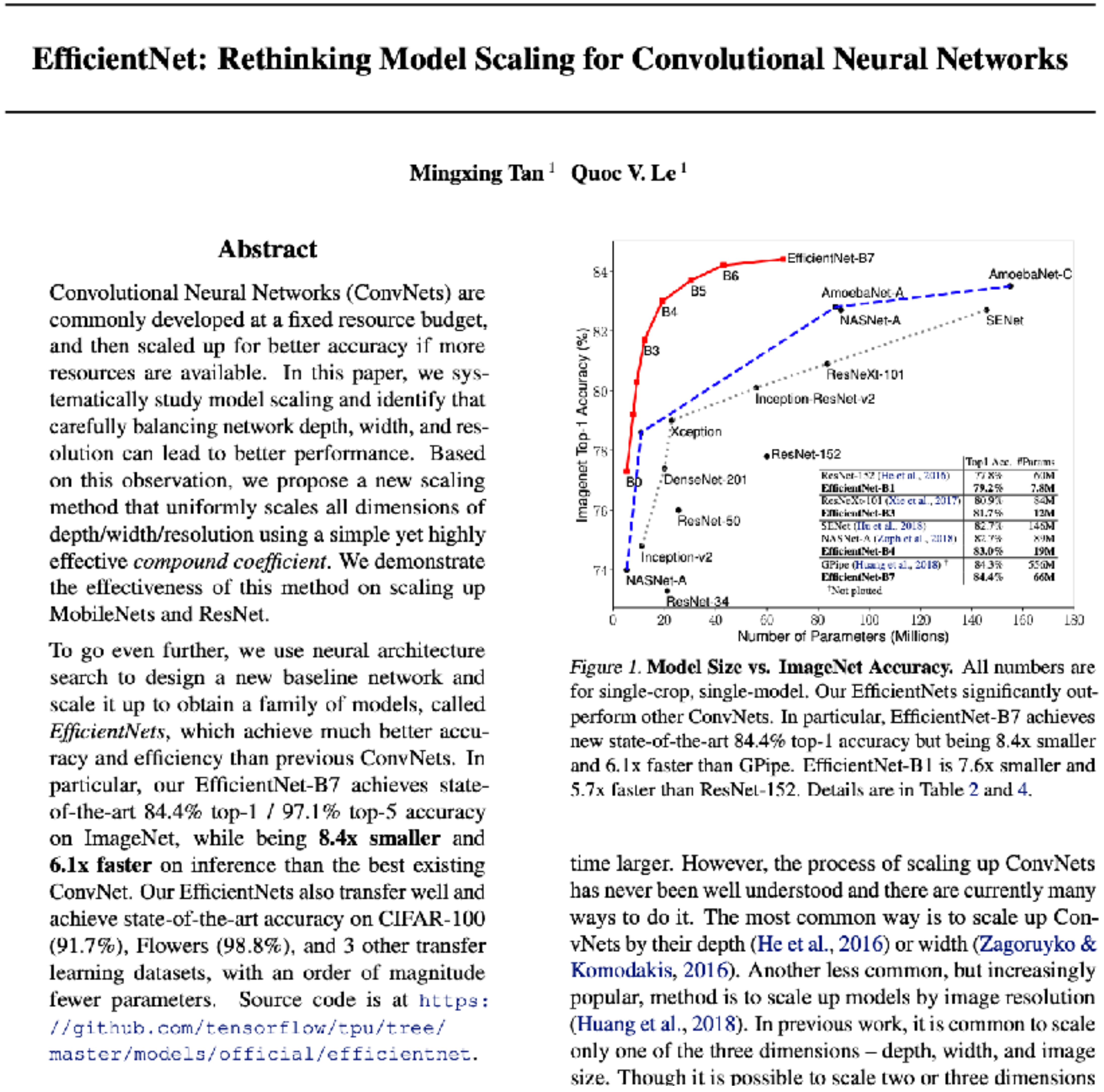
Base Architecture

- Efficient-Net-B0 architecture.

- Training Data Augmentation: rotations within a range of 180 degrees, resized crops with scales 0.4 to 0.6 and ratio of 0.9 to 1.1, color jitters including bright- ness (10%), saturation (10%), contrast (10%) and hue (3%), horizontal and vertical flips.

- We use Adam optimization with a base learning rate of 0.001 and Cosine Annealing Warm Restarts

- To account for the severe class imbalance present in the datasets, we use weighted sampling to construct a uniform class distribution in the training batches.



EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

Mingxing Tan [1] **Quoc V. Le** [1]

Abstract

Convolutional Neural Networks (ConvNets) are commonly developed at a fixed resource budget, and then scaled up for better accuracy if more resources are available. In this paper, we systematically study model scaling and identify that carefully balancing network depth, width, and resolution can lead to better performance. Based on this observation, we propose a new scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective *compound coefficient*. We demonstrate the effectiveness of this method on scaling up MobileNets and ResNet.

To go even further, we use neural architecture search to design a new baseline network and scale it up to obtain a family of models, called *EfficientNets*, which achieve much better accuracy and efficiency than previous ConvNets. In particular, our EfficientNet-B7 achieves state-of-the-art 84.4% top-1 / 97.1% top-5 accuracy on ImageNet, while being **8.4x smaller** and **6.1x faster** on inference than the best existing ConvNet. Our EfficientNets also transfer well and achieve state-of-the-art accuracy on CIFAR-100 (91.7%), Flowers (98.8%), and 3 other transfer learning datasets, with an order of magnitude fewer parameters. Source code is at https://github.com/tensorflow/tpu/tree/master/models/official/efficientnet.

	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.2%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.7%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**83.0%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.4%**	**66M**

†Not plotted

Figure 1. **Model Size vs. ImageNet Accuracy.** All numbers are for single-crop, single-model. Our EfficientNets significantly outperform other ConvNets. In particular, EfficientNet-B7 achieves new state-of-the-art 84.4% top-1 accuracy but being 8.4x smaller and 6.1x faster than GPipe. EfficientNet-B1 is 7.6x smaller and 5.7x faster than ResNet-152. Details are in Table 2 and 4.

time larger. However, the process of scaling up ConvNets has never been well understood and there are currently many ways to do it. The most common way is to scale up ConvNets by their depth (He et al., 2016) or width (Zagoruyko & Komodakis, 2016). Another less common, but increasingly popular, method is to scale up models by image resolution (Huang et al., 2018). In previous work, it is common to scale only one of the three dimensions – depth, width, and image size. Though it is possible to scale two or three dimensions

Experiment Set 1

- We aim to determine if the proposed uncertainty metrics can be related to errors in the prediction from the classifier.

- We train two classifiers for the problem of skin lesion classification in the ISIC Challenge 2018 and 2019 datasets, respectively

- During inference, we forward each image T = 100 times through the neural network using Test Augmentation, Test Time Dropout, and both uncertainty tech- niques simultaneously

Experiment Set 2

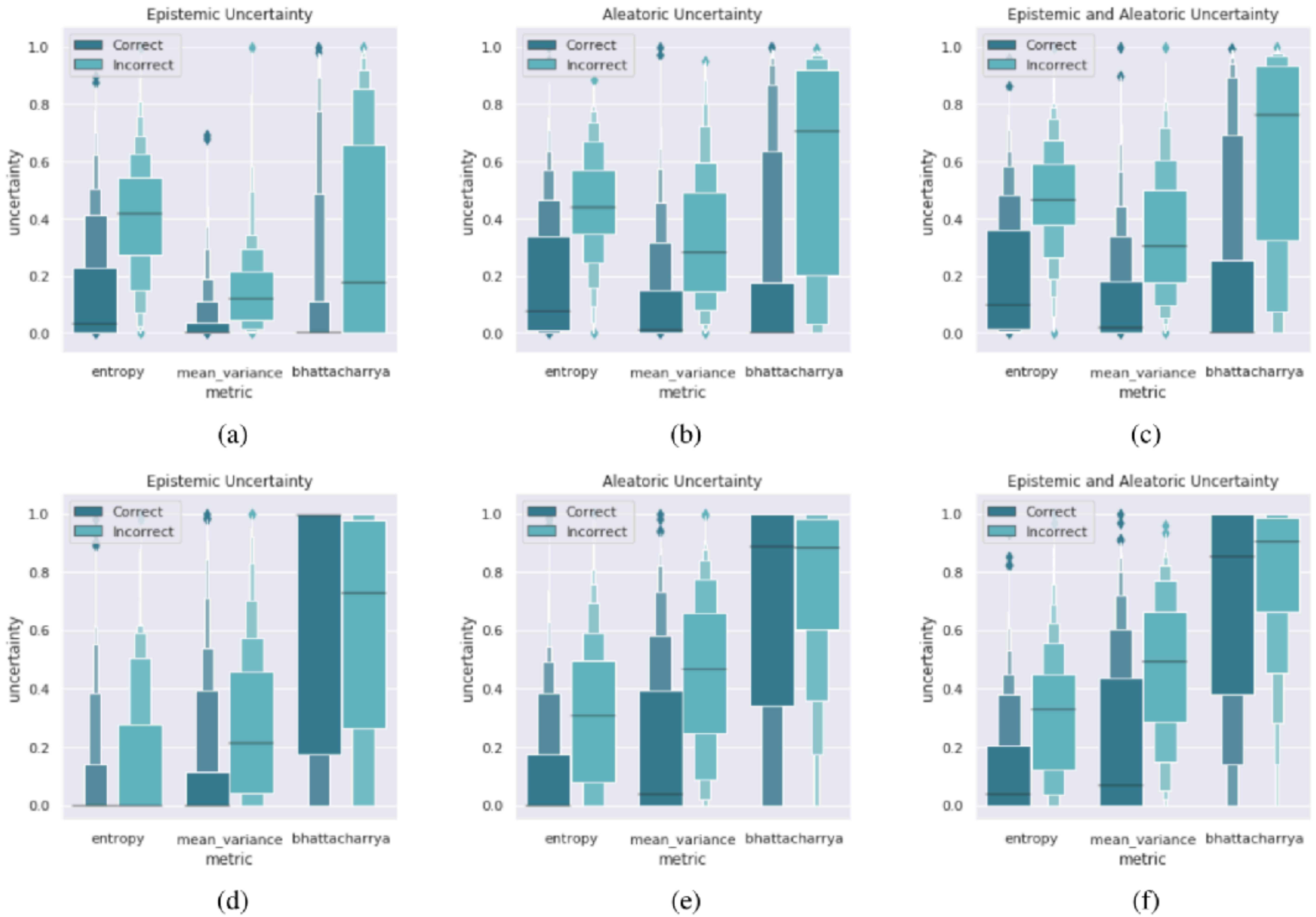
- We aim to determine if we can use the uncertainty metrics presented in section 3 to detect out-of-distribution samples, that is, samples from diagnostic categories that are not present in the training set.

- ISIC Challenge 2018: we move a subset of classes from the training set to the test set, train the network with the re- duced training set.

- ISIC Challenge 2019 as is.

Experiment Set 1

Results Experiment Set 1 (I)



(a) (b) (c)

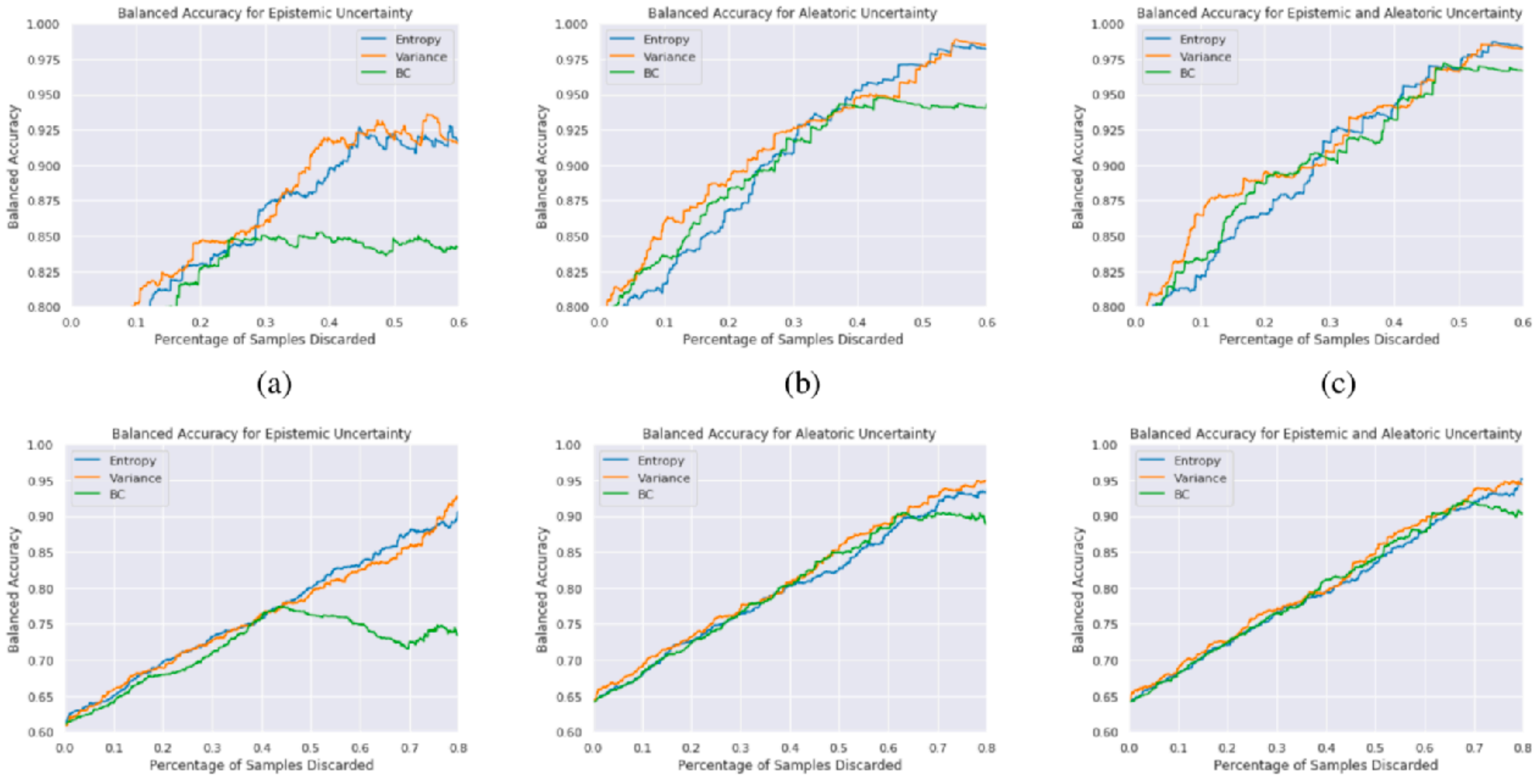
(d) (e) (f)

Figure 1. Experiment 1. Uncertainty metrics as a function of correct or incorrect predictions in the ISIC Challenge 2018 dataset.

ISIC Challenge	2018	2019
No sampling	0.74	0.61
Monte Carlo Dropout	0.73	0.61
Test Augmentation	**0.76**	**0.64**
Both	**0.76**	**0.64**

Table 1. Balanced accuracy for the trained classifier for different inference sampling techniques.

Results Experiment Set 1(II)



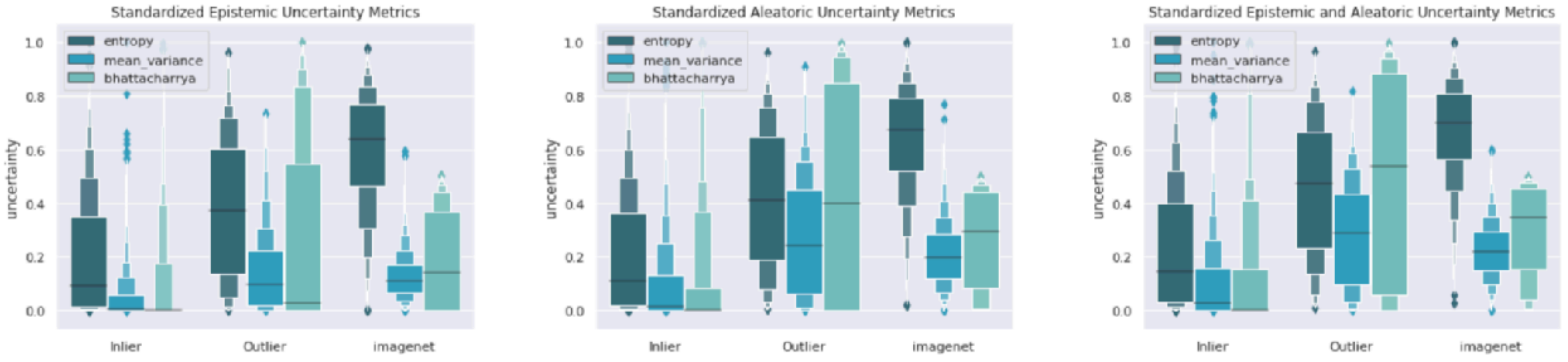
(a) (b) (c)

(d) (e) (f)

Figure 2. Experiment 1. Evolution of balanced accuracy as the most uncertain samples are removed from the test dataset for Monte Carlo dropout (a, d), Test Augmentation (b, e) and the combined method (c, f) for the ISIC Challenge 2018 (a, b, c) and ISIC Challenge 2019 (d, e, f) datasets.

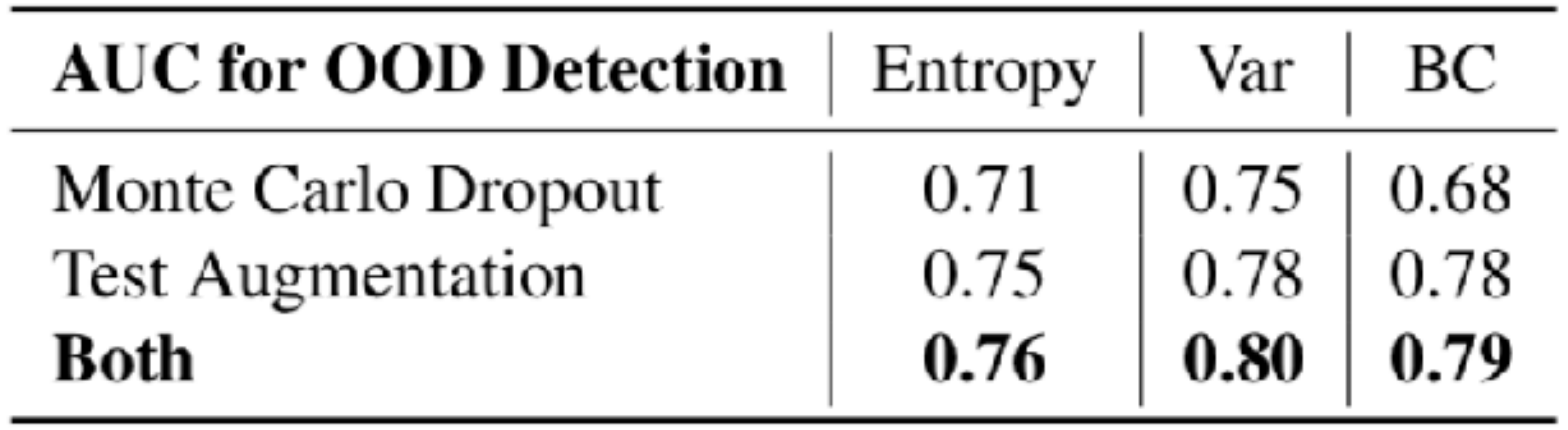
Experiment Set 2

Results Experiment Set 2 - ISIC Challenge 2018



(a) (b) (b)

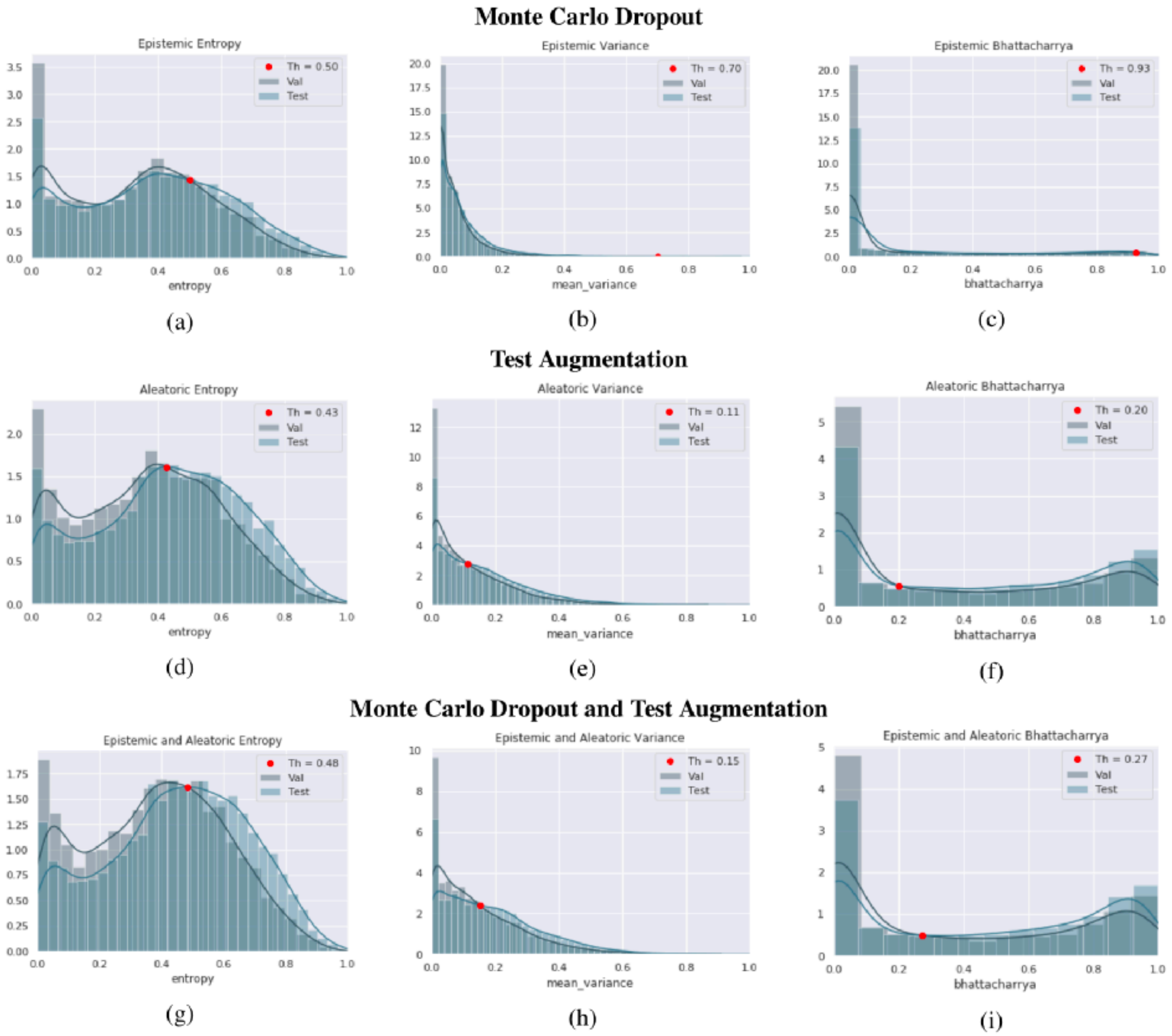
Figure 4. Experiment 2. Uncertainty metrics used to detect out-of-distribution samples in the ISIC Challenge 2018 dataset. (a) shows epistemic uncertainty estimation metrics based on the Monte Carlo Dropout method and (b) shows aleatoric uncertainty metrics based on the Test Augmentation method.



AUC for OOD Detection	Entropy	Var	BC
Monte Carlo Dropout	0.71	0.75	0.68
Test Augmentation	0.75	0.78	0.78
Both	**0.76**	**0.80**	**0.79**

Table 3. AUC of uncertainty metrics when used as predictors for out-of-distribution detection in the ISIC Challenge 2018 dataset (excluding samples from ImageNet).

Results Experiment Set 2 - ISIC Challenge 2019

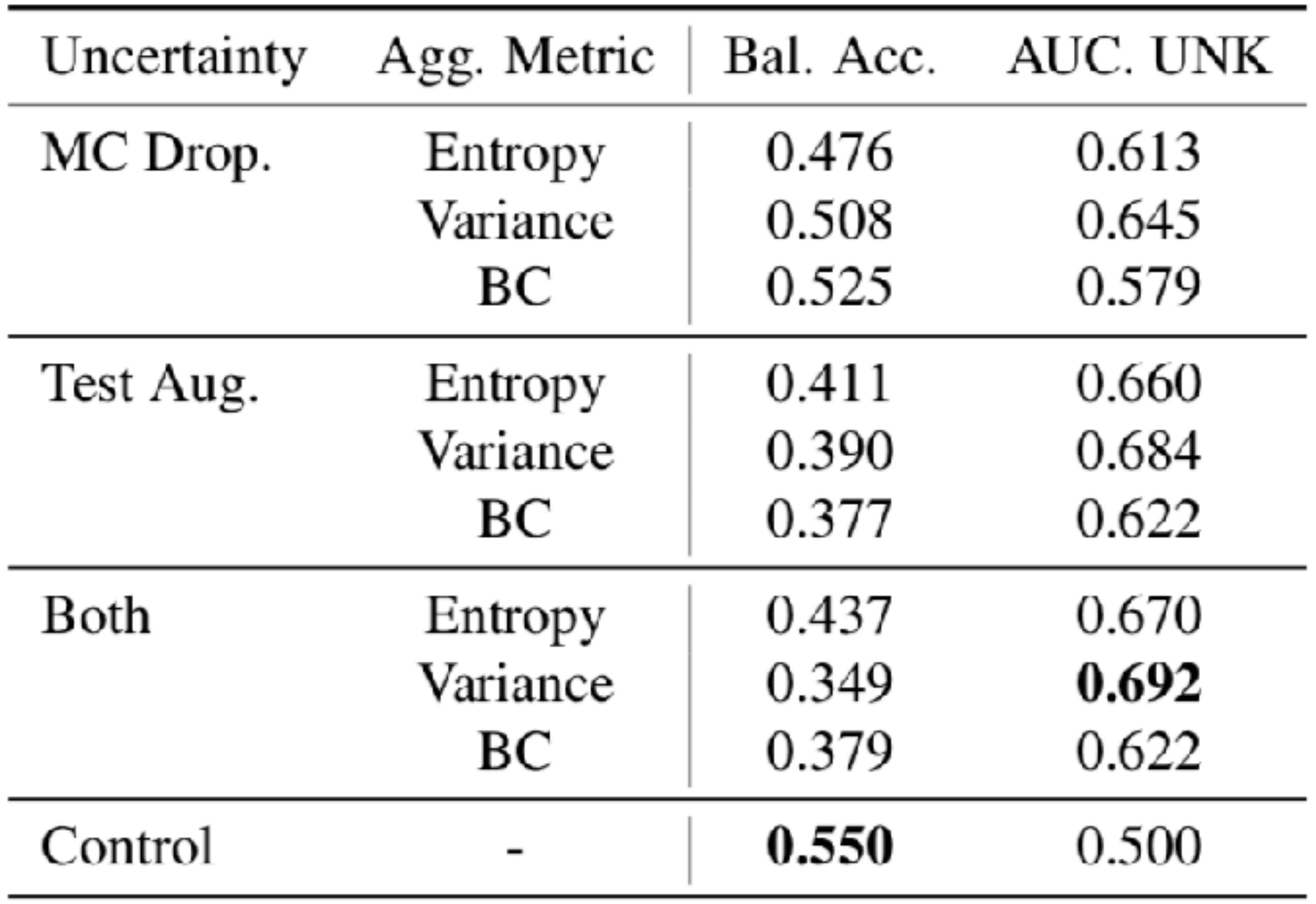


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5. Standardized histograms for the uncertainty metrics of the ISIC Challenge 2019 validation and test splits, with their corresponding probability density function estimations and selected thresholds.



Uncertainty	Agg. Metric	Bal. Acc.	AUC. UNK
MC Drop.	Entropy	0.476	0.613
	Variance	0.508	0.645
	BC	0.525	0.579
Test Aug.	Entropy	0.411	0.660
	Variance	0.390	0.684
	BC	0.377	0.622
Both	Entropy	0.437	0.670
	Variance	0.349	**0.692**
	BC	0.379	0.622
Control	-	**0.550**	0.500

Table 4. Balanced accuracy and AUC for out-of-distribution category in the live leaderboard from the ISIC Challenge 2019.

Conclusions

- Uncertainty metics are predictive of sample error
- Uncertainty metrics are predictive of out of distribution
- Selecting a threshold for OOD is hard without exemplar samples