Uncertainty Estimation in Deep Neural Networks for Dermoscopic Image Classification

Marc Combalia *
Hospital Clínic de Barcelona
Barcelona, Spain
mcombalia@clinic.cat

Ferran Hueto
Massachusetts Institute of Technology
Boston, EEUU
ferranhueto@gmail.com

Susana Puig
Hospital Clínic de Barcelona
Barcelona, Spain
spuig@clinic.cat

Josep Malvehy
Hospital Clínic de Barcelona
Barcelona, Spain
jmalvehy@clinic.cat

Verónica Vilaplana *
Universitat Politècnica de Catalunya
Barcelona, Spain
veronica.vilaplana@upc.edu

Abstract

The high performance of machine learning algorithms for the task of skin lesion classification has been shown over the past few years. However, real-world implementations are still scarce. One of the reasons could be that most methods do not quantify the uncertainty in the predictions and are not able to detect data that is anomalous or significantly different from that used in training, which may lead to a lack of confidence in the automated diagnosis or errors in the interpretation of results. In this work, we explore the use of uncertainty estimation techniques and metrics for deep neural networks based on Monte-Carlo sampling and apply them to the problem of skin lesion classification on data from ISIC Challenges 2018 and 2019. Our results show that uncertainty metrics can be successfully used to detect difficult and out-of-distribution samples.

1. Introduction

Machine learning and specifically deep learning, have dramatically improved the state-of-the-art in many areas of research, including computer vision, speech recognition, and natural language processing [27]. These advances are now seeing an application in the medical field, where deep neural networks are being used for a wide range of different purposes [29], including tumor segmentation [3], diabetic retinopathy detection [18], and cancer classification from histological tissue images [15].

Skin lesion classification by deep neural networks into different cancer sub-types has also experienced significant progress in the last years. A breakthrough moment occurred when a convolutional neural network (CNN) was trained on a dataset of 129450 images of skin lesions of different diseases in [13]. The neural network achieved the same accuracy as expert dermatologists on two binary classification cases: keratinocyte carcinomas versus benign seborrheic keratoses and malignant melanomas versus benign nevi. Since then, many other deep learning models have been proposed for the same purpose [5, 38, 16]. A key player in the evolution of the field is the International Skin Imaging Collaboration (ISIC) [1]. This expert consortium has been developing digital imaging standards for skin cancer imaging and has created a public archive containing the most extensive publicly available collection of quality controlled dermoscopic images of skin lesions. Moreover, since 2016, ISIC organizes yearly artificial intelligence challenges presenting problems in lesion segmentation and lesion classification, promoting the growth of automated diagnostic systems for skin cancer [6, 2, 7, 31].

Despite the rapid acceleration of deep learning research in healthcare, with potential applications being demonstrated across various domains, there are currently limited examples of these techniques being successfully deployed into clinical practice [22]. The challenges and limitations for the deployment of such systems into real-world envi-

ronments are related to several factors: ethics and regulatory aspects, data availability and variability, and technology issues intrinsic to machine learning solutions. Among the last ones, essential factors include dataset shift, fitting of confounders, interpretability or explainability of decisions, generalization to different populations, and the development of reliable measures of model confidence.

Most deep learning-based solutions produce deterministic outputs and do not quantify or control the uncertainty in the prediction, which may lead to a lack of confidence in the automated diagnosis and errors in the interpretation of results.

Usually, performance is given in terms of global metrics related to the models' discriminative power like sensitivity, specificity, AUC, or ROC curves. However, it is crucial to know how sure or confident the model is about a prediction, especially in the clinical practice where diagnostic errors are especially relevant, and there are always difficult cases that may require closer examination or a second expert opinion [28].

Uncertainty can not only be used to determine which samples are hard to classify, thus requiring further inspection by an expert but also to detect samples that deviate from the data used for training the model. When training and test data distributions differ, the network can still provide (arbitrary) predictions with high confidence. This problem is known as the out-of-distribution problem. Identifying when a model is being used on a domain different from the training domain is also a critical issue in medical applications [36, 4].

Obtaining reliable uncertainty estimates of neural network predictions is a long-standing challenge. In this work, we explore the use of uncertainty estimation techniques and metrics for deep neural networks based on Monte-Carlo sampling and apply them to the problem of skin lesion classification in the context of the ISIC Challenges 2018 and 2019. We measure the uncertainty associated with the predictions of the classifier and study the use of uncertainty in two cases: (i) identification of difficult cases where uncertainty is related to errors in the classifier predictions, and (ii) detection of out-of-distribution samples, where we do not assume the availability of samples or the distribution of the anomalous class.

2. Related Work

While there are many sources of uncertainty, they are generally characterized as epistemic or aleatoric [11].

Epistemic uncertainty or model uncertainty captures the uncertainty in the model parameters. It arises from the lack of sufficient data to train the model to infer the underlying data-generating function correctly. Therefore, epistemic uncertainty is inversely proportional to the density of training examples and could be reduced by collecting and training on more data.

On the other hand, aleatoric uncertainty is described by the noise in the observations; it is the input-dependent uncertainty. This type of uncertainty arises due to hidden variables or measurement errors and cannot be explained away by capturing more data [23, 36].

Uncertainty in neural networks can be modeled with a Bayesian analysis. Uncertainties are formalized as probability distributions over the model parameters (for epistemic uncertainty) or model inputs (for aleatoric uncertainty) [23].

Epistemic uncertainty is modeled by placing a prior distribution over the model's weights W, and capturing how much these weights vary given some data. Bayesian neural networks (BNN) provide a principled mathematical framework for this kind of uncertainty. Denoting by $f^W(x)$ the output of the network, Bayesian inference is used to compute a posterior over the weights $p(W|X,Y)$ for given training data (X,Y), where X and Y are training samples and labels. From this posterior, $p(y|w,x)$, the predictive distribution of a test sample x is obtained and used to quantify the predictive uncertainty.

However, except in trivial cases, exact Bayesian inference is computationally intractable for neural networks, and therefore several approximations have been proposed, like Laplace approximation, Markov chain Monte Carlo methods, variational Bayesian methods, or assumed density filtering (see [25] and references therein).

A second group of works uses Monte Carlo sampling as an alternative way to generate an ensemble of predictions and estimate uncertainty. Predictions can be generated by using an ensemble of differently trained neural networks as proposed by Lakshminarayanan et al. [25], or by dropout at test time. This last method proposed by Gal [14] is prevalent in practice due to its simplicity. It consists of training a model with dropout at every layer and then executing dropout at test time to sample from the approximate posterior.

Other strategies for epistemic uncertainty are related to its use for out-of-distribution detection [36]. They consist of creating a dataset with both anomalies (or 'negative samples') and in-distribution data. These negative examples resemble realistic input configurations that lay outside the data distribution. Then, a binary classifier is trained to distinguish between original training points and negative examples and may be used to measure epistemic uncertainty. Another approach consists of adding a 'none-of-the-above' category to the original classification problem. Examples of this strategy include noise-contrastive priors [20], and GANs [17].

The previous approaches represent the posterior $p(Y|X)$ by sampling but do not model aleatoric uncertainty. One strategy to achieve this goal is to assume that the conditional distribution of a target variable given the input is Gaussian.

This is the approach followed by Kendall [23], which trains a model that approximates the posterior distribution and produces two outputs, a predictive mean and a predictive variance. The method is combined with Monte-Carlo sampling to estimate both aleatoric and epistemic uncertainties at the same time. One limitation of this technique is that it requires to adapt the network architecture and loss function, which hinders the application to already trained models.

Another approach for estimating the aleatoric uncertainty is based on data augmentation. Data augmentation has been typically used to obtain additional training samples by applying transformations (flipping, cropping, rotating, scaling, elastic deformations) to the dataset samples. Ayhan [34] and Wang [41] proposed using data augmentation at test time as a means to estimate aleatoric uncertainty, by generating several augmented examples per test case, obtaining their predictions and using them to estimate uncertainty.

Estimation of epistemic and aleatoric uncertainty has also been investigated for medical image classification and segmentation. Laves et al. [26] modify a ResNet architecture to produce posterior distributions using a variational model and Monte Carlo dropout at test time and estimate uncertainty using the variance of the predictions in a problem of OCT scan classification. They show that cases incorrectly classified correlate with higher uncertainty. Leibig [28] also uses Monte Carlo dropout at test time for diabetic retinopathy detection from fundus images and shows that uncertainty informed decision referral can improve diagnostic performance. Ayhan and Berens [34] use data augmentation at test time to capture aleatoric uncertainty in a diabetic retinopathy detection dataset. They use it to classify healthy vs. disease samples showing that wrongly classified samples have higher uncertainty than correct predictions, and the AUC increases as more uncertain samples are removed from the test set. Regarding medical image segmentation, Wang [41] uses Monte Carlo dropout and test data augmentation uncertainty estimates to help to reduce overconfident incorrect predictions in 3D brain tumor and 2D fetal brain segmentation. Also, for brain segmentation, Eaton-Rosen [12] uses the combined strategy proposed in [23] to quantify uncertainty, convert voxel-wise uncertainty into volumetric uncertainty and calibrate the accuracy and reliability of confidence intervals of derived measurements. Nair [32] trains a CNN for multiple sclerosis lesion detection, and segmentation augmented to provide different voxel-based uncertainty measures based on Monte Carlo dropout and analyzes the performance of voxel-based segmentation and lesion-level detection by choosing operating points based on the uncertainty. Jungo [21] reports the results of evaluating several voxelwise uncertainty measures with respect to their reliability and limitations for brain tumor segmentation and skin lesion segmentation.

3. Materials and Methods

3.1. Datasets

3.1.1 ISIC 2018 Challenge and Dataset

The ISIC Challenge 2018 comprises three different tasks: skin lesion segmentation, skin lesion attribute detection, and skin lesion classification. In our work, we focus on the third task, which consists of the implementation of machine learning algorithms to classify dermoscopic images of skin tumors into seven different diagnostic categories: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis/Bowen's disease, benign keratosis, dermatofibroma, and vascular lesions. The ISIC Challenge 2018 dataset is based on the HAM10000 database [19]. 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.

3.1.2 ISIC 2019 Challenge and Dataset

In the ISIC Challenge 2019, the participants are asked to build classifiers for the diagnosis of skin tumors from dermoscopic images. The challenge is divided into two tasks. In the first task, the participants are asked to classify the lesions using only the information available in the images, while in the second task, the participants can use extra metadata (age, sex, and anatomic location of the lesion) to come up with the diagnosis. The training dataset of the ISIC Challenge 2019 consists of 25331 dermoscopic images of 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 [39], and also adds images from the BCN20000 dataset [8] and the MSK dataset [7]. The BCN20000 dataset is considered to be remarkably complex since it includes uncurated images from day to day clinical practice [8].

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 training split (Unknown category). The participants' task is to identify these images as being out-of-distribution and to label them accordingly in their submission file, together with the rest of the classes.

3.2. Uncertainty methods

To model the epistemic uncertainty, we follow the Monte-Carlo Test Time Dropout method proposed by Gal

[14]. We train the network using random dropout (with a dropout rate of 0.2). To capture model uncertainty for a given image x we use dropout at test time with the same probability and perform multiple predictions $\{\mathbf{p}_t = p(y|x, X, Y, W^t)\}_{t=1...T}$, where each prediction is a vector of softmax scores for the C classes.

In order to account for aleatoric uncertainty in the predictions, we use Test Data Augmentation [34, 41]. We forward the images several times through the neural network with random data augmentation configurations during test time. For each image x we obtain $\{x_t\}_{t=1...T}$ versions by data augmentation, and forward the images to obtain a set of predictions $\{\mathbf{p}_t = p(y|x_t, X, Y, W)\}_{t=1...T}$.

Finally, we combine the two approaches to estimate both epistemic and aleatoric uncertainty: for each image x, we generate T augmented examples and forward each example once using dropout at test time. We obtain a set of predictions $\{\mathbf{p}_t = p(y|x_t, X, Y, W^t)\}_{t=1...T}$.

3.3. Uncertainty measures

We use three different metrics to quantify the predictive uncertainty: entropy, variance, and the Bhattacharyya coefficient between distributions.

First, for an image x we obtain T different predictions $\mathbf{p}_t(x)$ (either by Monte-Carlo Dropout, Test Augmentation of the combination of the two strategies), where each prediction is a vector of softmax scores for the C classes. Then we compute the average prediction score for the T samples:

$$\mathbf{p}_T(x) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{p}_t(x) \tag{1}$$

Predictive Entropy: The entropy is a measure interpreted as the average level of information or uncertainty inherent in the possible outcomes of a random variable [35].

$$H(\mathbf{p}_T(x)) = -\sum_{c=1}^{C} p_T(x)[c]\log(p_T(x)[c]) \tag{2}$$

where $p_T(x)[c]$ is the c-th element of the vector $\mathbf{p}_T(x)$, the average prediction score for class c.

Prediction Variance: We compute the variance of the T predictions for each class, and then the mean-variance over all classes

$$\sigma^2(\mathbf{p}_T(x)) = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{p}_t(x) - \mathbf{p}_T(X))^2 \tag{3}$$

Note that now all the operations are element-wise.

Bhattacharyya Coefficient (BC): The Bhattacharyya coefficient is a measure of the amount of overlap between two statistical samples or populations. As proposed in [40], we compute the normalized BC between the two classes with higher predictive mean.

$$BC(h_{c1}, h_{c2})(x) = \sum_{n=1}^{N}\sqrt{h_{c1}[n] * h_{c2}[n]} \tag{4}$$

where h_{c1} and h_{c2} are the N-bin histograms of the two classes with higher predictive mean $P_T(x)[c1]$ and $P_T(x)[c2]$.

4. Experiments and Results

We conduct a series of experiments to study whether the uncertainty metrics described in section 3 can be used to identify difficult to classify and out-of-distribution samples. All the classifiers used in the experiments are based in the Efficient-Net-B0 [37] architecture. During training, each RGB input image is resized to (224, 224) and augmented by performing the following random operations: 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 brightness (10%), saturation (10%), contrast (10%) and hue (3%), horizontal and vertical flips. We use Adam optimization [24] with a base learning rate of 0.001 and Cosine Annealing Warm Restarts [30] to modify the learning rate during training and we use early stopping to select the set of weights with best validation performance. To account for the severe class imbalance present in the datasets, we use weighted sampling [33] to construct a uniform class distribution in the training batches.

4.1. Experiment 1 - Uncertainty as a measure of confidence

In this set of experiments, we aim to determine if the proposed uncertainty metrics can be related to errors in the prediction from the classifier. To do so, we train two classifiers for the problem of skin lesion classification in the ISIC Challenge 2018 and 2019 datasets, respectively. We divide the datasets into train (64 %), validation (16 %) and test (20 %) splits. During inference, we forward each image $T = 100$ times through the neural network using Test Augmentation, Test Time Dropout, and both uncertainty techniques simultaneously.

We obtain predictions for the classifiers by averaging the output softmax vector over the T iterations and compute the uncertainty of each prediction based on the methods described in section 3.

We report the balanced accuracy in the test set for each inference configuration in table 1. Figure 1 shows the distribution of the uncertainty metrics, stratified by the correctness of the predictions of the classifier in the test set. Similarly to [40], we aim to understand how the uncertainty metrics relate to the accuracy of the predictions of the classifier. To do so, we compute the evolution in balanced ac-

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.

	Train	Validation	Test
Inlier	6261	1546	1951
Skin Outliers	0	0	257
Imagenet Outliers	0	0	12264

Table 2. Inlier / Outlier distribution of samplings in training, validation and test splits.

curacy as we iteratively discard the most uncertain samples from the test dataset. Figure 2 show these results for the ISIC Challenge 2018 and 2019 datasets.

4.2. Experiment 2 - Uncertainty to detect out-of-distribution Samples

In this set of experiments, 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.

4.2.1 ISIC Challenge 2018

We create a controlled experiment where the classifier is used with diagnostic categories not present in the training dataset. To do so, we move a subset of classes from the training set to the test set, train the network with the reduced training set, and compute the uncertainty metrics for samples grouped into in-distribution and out-of-distribution cases.

More concretely, we move the Dermatofibroma and Vascular Lesion categories from the training dataset to the test set of the ISIC Challenge 2018 dataset. We choose these two classes for being the ones with the lowest representation in the training split to minimize the impact on the classifier performance. We also add 12264 images from the Imagenet dataset [10] to the test split, to compare the magnitude of uncertainty metrics for images from a completely different domain. Table 2 shows sample distribution across the train, validation, and test splits.

After training, we compute the uncertainty metrics and report them in figure 4, stratified by in-distribution, and out-of-distribution categories. We also compute the AUC for each uncertainty metrics when used as predictor for out-of-distribution detection (excluding samples from ImageNet),

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).

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.

and report the results in table 3.

4.2.2 ISIC Challenge 2019

We train a classifier on the images from the ISIC Challenge 2019. We split the training dataset into the train (80 %), and validation (20 %) splits (used for early stopping). We obtain the predictions and uncertainty metrics averaging over $T = 100$ iterations during inference in the official test dataset downloaded from the ISIC Live Challenge 2019 platform.

Figure 5 shows the histograms of each uncertainty metric for the validation and testing splits. In order to obtain a threshold to label the samples as out-of-distribution, we use the intersection between the probability density function for the validation and test splits, based on the estimation by the Parzen–Rosenblatt window technique [9] (the higher threshold is selected if there was more than one intersection point). We then upload the results to the ISIC Live Challenge platform, and report the balanced accuracy results and AUC for the Unknown category in table 4.

5. Discussion

The results from our first set of experiments show that uncertainty metrics can be used to explain the confidence of the classifier. Figure 4 shows that all the uncertainty metrics reported in this publication were higher for the predic-

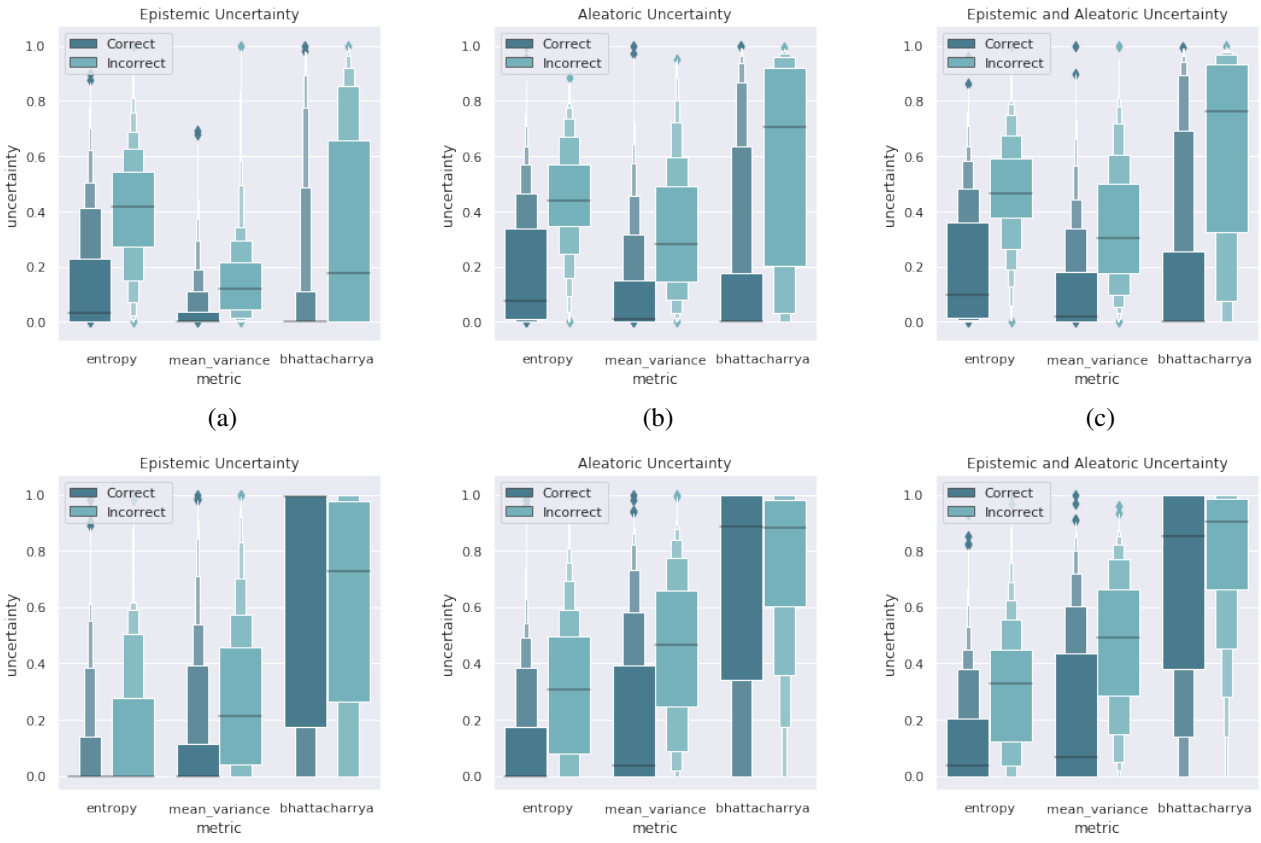


Figure 1. Experiment 1. Uncertainty metrics as a function of correct or incorrect predictions in the ISIC Challenge 2018 dataset.

tion errors at the output of the classifier. In figure 2, we can observe how the balanced accuracy of the classifier improved as the most uncertain samples were removed from the test set. Both of these results suggest that uncertainty metrics can be useful to identify samples where the classifier is prone to produce incorrect predictions. The combination of Monte Carlo Dropout and Test Augmentation was the most efficient technique at identifying error-prone samples, as seen in table 3. Moreover, the results in table 1 confirm that these techniques can also be used to boost the accuracy of the trained classifier during inference.

The second set of experiments validate that the uncertainty metrics can also be useful to identify out-of-distribution samples during inference. Figure 4 reveals that both Monte Carlo Dropout and Test Augmentation uncertainty metrics are higher for the out-of-distribution samples of the test dataset. These metrics were not only efficient at identifying samples from the Imagenet dataset but were also valuable for detecting diagnostic categories from skin lesions not present in the training split. The results in table 3 support this conclusion: the AUC for the metrics described was high when used as predictors for out-of-distribution detection. Once again, the combination of both methods was the most efficient, achieving an AUC of 0.80.

The uncertainty estimation metrics were also useful to detect out-of-distribution samples in the ISIC Challenge 2019 dataset. The higher AUC amongst the techniques presented for out-of-distribution detection was 0.692 for the combination of Monte Carlo Dropout and Test Augmentation (table 4). However, selecting an appropriate threshold to detect the out-of-distribution category was found to be challenging without representative samples from the Unknown class in the validation set, and the balanced accuracy of the classifier was negatively impacted.

6. Conclusions

In this study, we use Monte Carlo dropout and Test Augmentation uncertainty estimation techniques to identify hard samples at the output of a classifier. We also report the use of these techniques to detect out-of-distribution samples. The uncertainty metrics showed a high AUC for the problem of out-of-distribution detection. However, selecting an appropriate threshold to label them accordingly proved to be a difficult task without the representation of such categories in the validation set. We present our re-

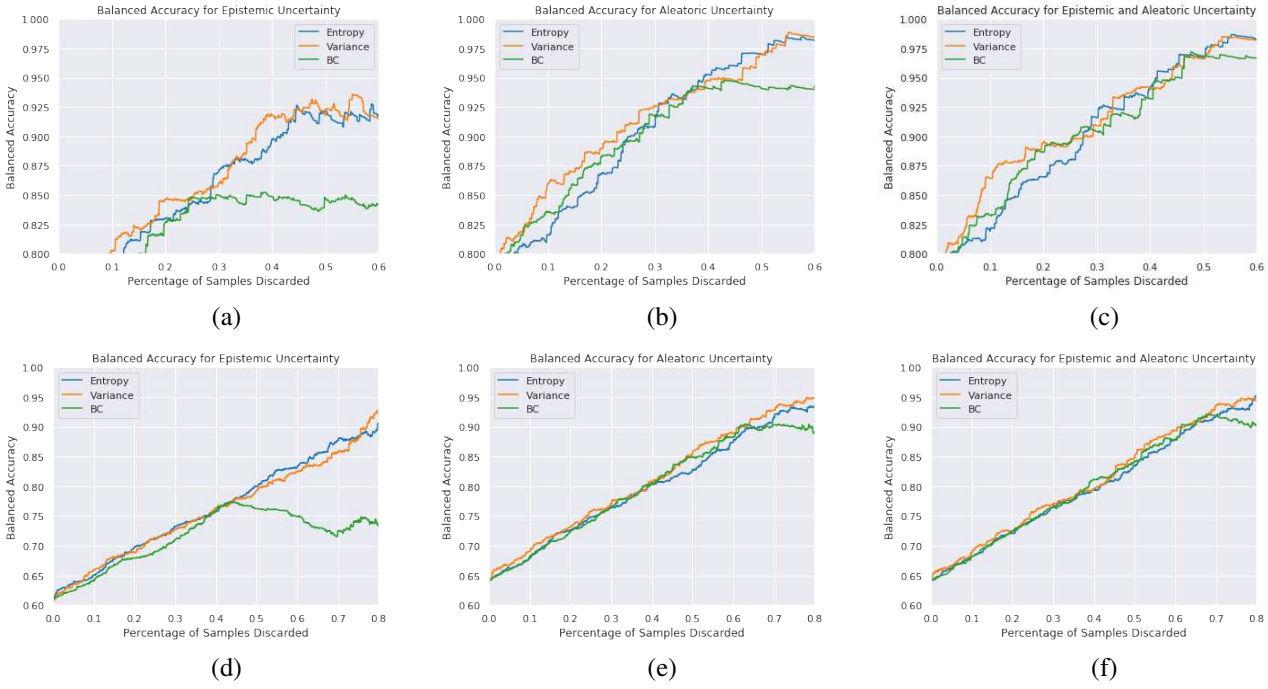


Figure 2. Experiment 1. Evolution of balanced accuracy as the most uncertain samples are removed from the test dataset for Monte Carlo dropout (a, d), Test Augmentation (b, e) and the combined method (c, f) for the ISIC Challenge 2018 (a, b, c) and ISIC Challenge 2019 (d, e, f) datasets.

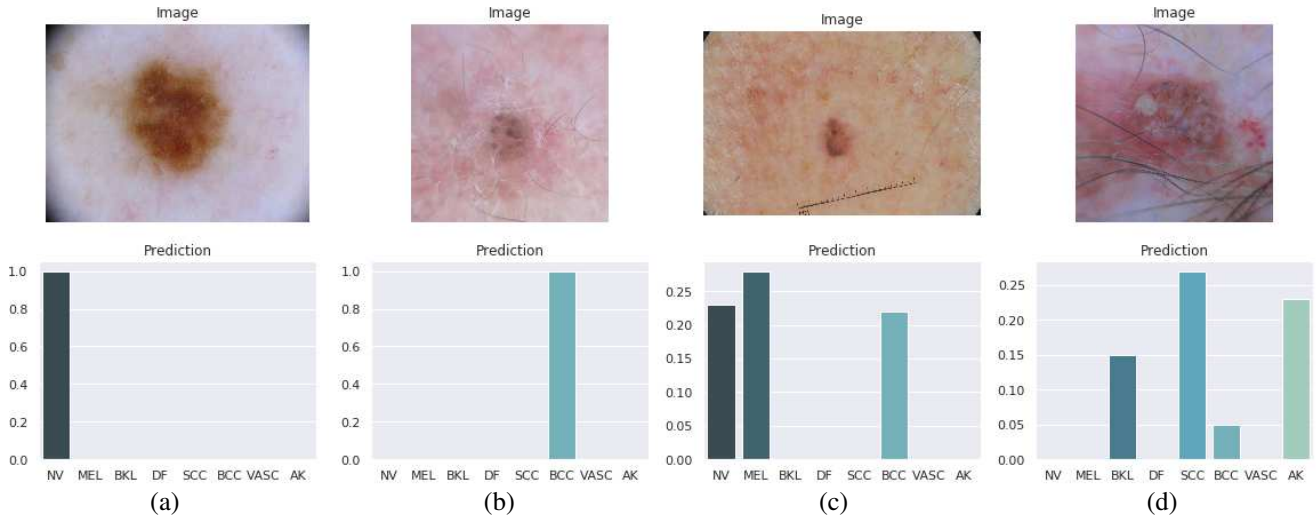


Figure 3. Examples of the most (c and d) and least (a and b) uncertain samples from the ISIC Challenge 2019 dataset (test augmentation, variance), with their corresponding softmax output from the classifier.

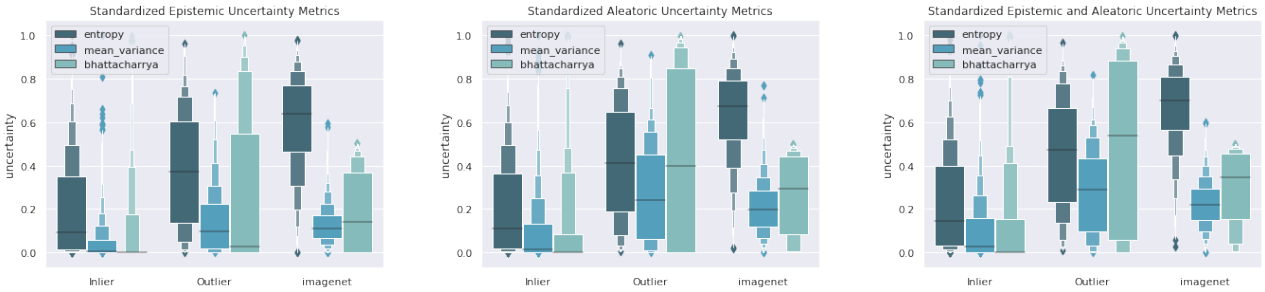
sults for the problem of skin lesion classification from dermoscopic images with the ISIC Challenge 2018 and 2019 datasets.

References

[1] The international skin imaging collaboration. https://www.isic-archive.com/. Accessed: 2020-02-21. 1

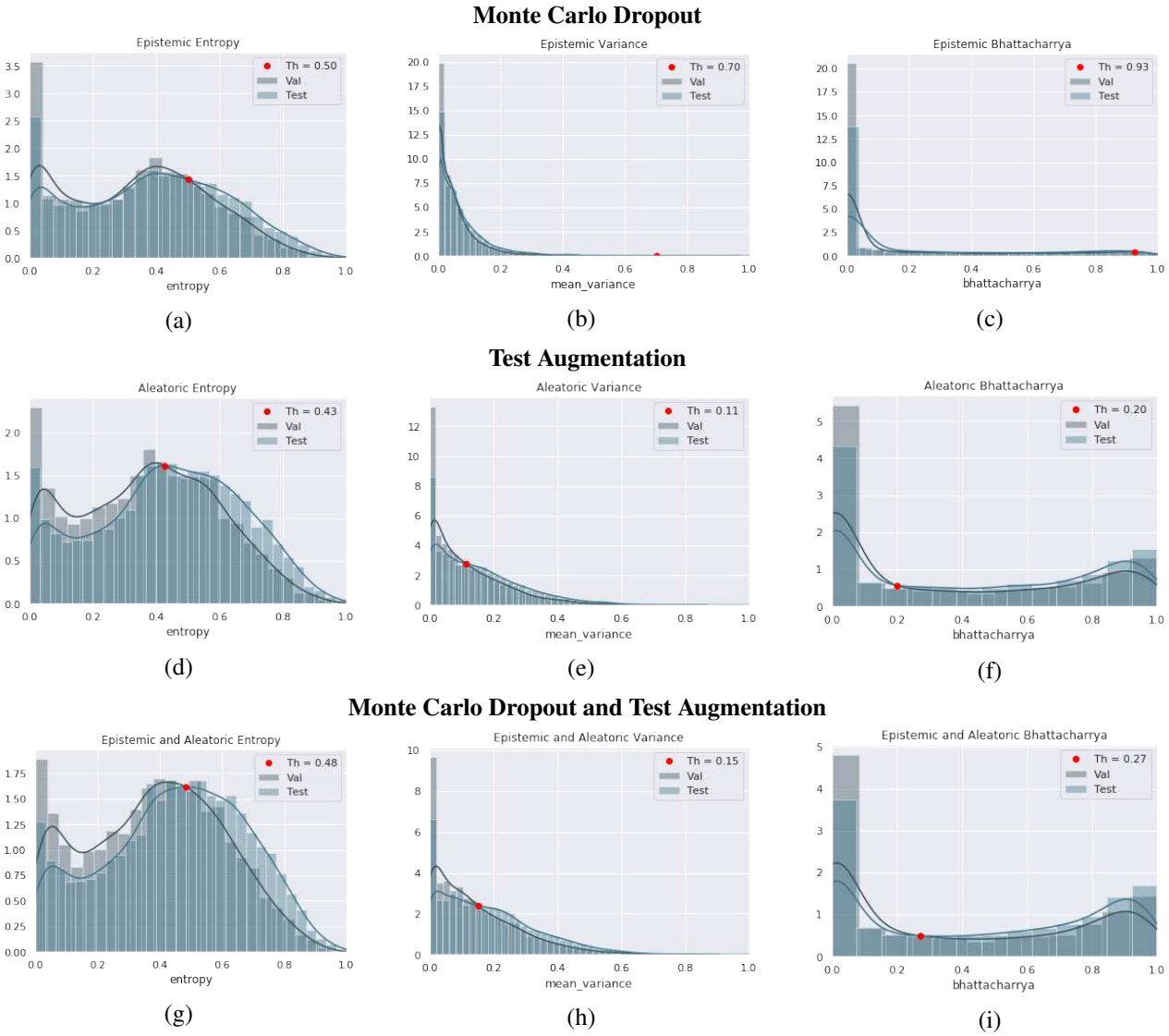
[2] Isic challenge 2019. https://challenge2019.isic-archive.com/. Accessed: 2020-02-21. 1

[3] Spyridon Bakas, Mauricio Reyes, András Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, Marcel Prastawa, Esther Alberts, and et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *CoRR*, abs/1811.02629, 2018. 1



(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.



(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.

[4] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019. 2

[5] Titus J Brinker, Achim Hekler, Alexander H Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Stefan Fröhling, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154, 2019. 1

[6] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 1

[7] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 1, 3

[8] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Allan C Halpern, Susana Puig, and Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019. 3

[9] Richard A Davis, Keh-Shin Lii, and Dimitris N Politis. Remarks on some nonparametric estimates of a density function. In *Selected Works of Murray Rosenblatt*, pages 95–100. Springer, 2011. 5

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 5

[11] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009. 2

[12] Zach Eaton-Rosen, Felix Bragman, Sotirios Bisdas, Sébastien Ourselin, and M Jorge Cardoso. Towards Safe Deep Learning: Accurately Quantifying Biomarker Uncertainty in Neural Network Predictions. In Alejandro F Frangi, Julia A Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 691–699, Cham, 2018. Springer International Publishing. 3

[13] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. 1

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *arXiv preprint arXiv:1506.02142*, 2015. 2, 4

[15] Daniel S Gareau, Joel Correa da Rosa, Sarah Yagerman, John A Carucci, Nicholas Gulati, Ferran Hueto, Jennifer L DeFazio, Mayte Suárez-Fariñas, Ashfaq Marghoob, and James G Krueger. Digital imaging biomarkers feed machine learning for melanoma screening. *Experimental dermatology*, 26(7):615–618, July 2017. 1

[16] Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. *arXiv preprint arXiv:1910.03910*, 2019. 1

[17] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[18] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip Q Nelson, Jessica Mega, and Dale Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 2016. 1

[19] Holger A Haenssle, Christine Fink, R Schneiderbauer, Ferdinand Toberer, Timo Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, Luc Thomas, A Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018. 3

[20] Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. Noise Contrastive Priors for Functional Uncertainty. 2018. 2

[21] Alain Jungo and Mauricio Reyes. Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation. In Dinggang Shen, Tianming Liu, Terry M Peters, Lawrence H Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 48–56, Cham, 2019. Springer International Publishing. 3

[22] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):1–9, 2019. 1

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 2, 3

[24] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations*, May 2015. San Diego, CA. 4

[25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 2017-December(Nips):6403–6414, 2017. 2

[26] Max-Heinrich Laves, Tobias Ortmaier, and Sontje Ihler. Uncertainty quantification in computer-aided diagnosis: Make

your model say "i don't know" for ambiguous cases. 06 2019. 3

[27] Y. LeCun, Y. Bengio, and G. Hinton. Deep Learning. *Nature*, 521:436–444, May 2015. 1

[28] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1):17816, dec 2017. 2, 3

[29] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[30] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 4

[31] Michael A Marchetti, Noel CF Codella, Stephen W Dusza, David A Gutman, Brian Helba, Aadi Kalloo, Nabin Mishra, Cristina Carrera, M Emre Celebi, Jennifer L DeFazio, et al. Results of the 2016 international skin imaging collaboration isbi challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *Journal of the American Academy of Dermatology*, 78(2):270, 2018. 1

[32] Tanya Nair, Doina Precup, Douglas L. Arnold, and Tal Arbel. Exploring Uncertainty Measures in Deep Networks for Multiple Sclerosis Lesion Detection and Segmentation. aug 2018. 3

[33] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018. 4

[34] Murat Seçkin Ayhan and Philipp Berens. Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks. In *1st Conference on Medical Imaging with Deep Learning, Amsterdam, The Netherlands*, 2018. 3, 4

[35] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001. 4

[36] Natasa Tagasovska and David Lopez-Paz. Single-Model Uncertainties for Deep Learning. (NeurIPS), 2018. 2

[37] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 4

[38] Philipp Tschandl, Giuseppe Argenziano, Majid Razmara, and Jordan Yap. Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features. *British Journal of Dermatology*, 181(1):155–165, 2019. 1

[39] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018. 3

[40] Pieter Van Molle, Tim Verbelen, Cedric De Boom, Bert Vankeirsbilck, Jonas De Vylder, Bart Diricx, Tom Kimpe, Pieter Simoens, and Bart Dhoedt. Quantifying uncertainty of deep neural networks in skin lesion classification. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*, pages 52–61. Springer, 2019. 4

[41] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019. 3, 4