Computers in Biology and Medicine 144 (2022) 105357

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Uncertainty-aware skin cancer detection: The element of doubt

Pegah Tabarisaadi [a,*], Abbas Khosravi [a], Saeid Nahavandi [b,c]

[a] *Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Waurn Ponds, VIC, 3216, Australia*
[b] *Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Geelong, Australia*
[c] *Harvard Paulson School of Engineering and Applied Sciences, Harvard University, Allston, MA, 02134, USA*

ARTICLE INFO

Keywords:
Uncertainty quantification
Skin cancer
Monte Carlo (MC) dropout
Bayesian ensembling
Spectral-normalized Neural Gaussian Process (SNGP)

ABSTRACT

Artificial intelligence (AI)-based medical diagnosis has received huge attention due to its potential to improve and accelerate the decision-making process at the patient level in a range of healthcare settings. Despite the recent signs of progress in this field, reliable quantification and proper communication of predictive uncertainties have been fully or partially overlooked in the existing literature on AI applications for medical diagnosis. This paper studies the automatic diagnosis of skin cancer using dermatologist spot images. Three different uncertainty-aware training algorithms (MC dropout, Bayesian Ensembling, and Spectral Normalized Neural Gaussian Process) are utilized to detect skin cancer. The performances of the three above-mentioned algorithms are compared from different perspectives. In addition, some images from the Cifar10 dataset are applied as the out-of-domain data and the performances of the algorithms are evaluated and compared for images that are far from the training samples. The accuracy, uncertainty accuracy, uncertainty accuracy for out-of-domain distribution samples, and the uncertainties of the predictions are reported in all cases and compared.

1. Introduction

Melanoma and non-melanoma are considered as two main types of skin cancer. Melanoma is the 19th most common cancer worldwide. In 2018 nearly 300,000 new cases of melanoma were reported. On the other hand, non-melanoma is the 5th common cancer globally and in 2018 more than 1 million cases were diagnosed [1]. It is projected that new cases of melanoma and non-melanoma will increase to 450,000 and 2 million cases per year over the next 20 years respectively [2].

For skin cancer, the chance of survival is highly related to the stage that it is diagnosed. For instance, in the United States, the five-year survival for localized melanoma is reported at 98.4%, while for metastatic melanoma it is reported at 22.5% [3]. The mortality rate for non-melanoma skin cancer is low with 0.64 death per 100,000 cases [2]. So a timely diagnosis is critical for a successful treatment.

Analyzing the images of dermatologist spots are manually done by medical experts and it will take approximately a long time. As it is forecasted that the infection rate will increase in the coming years, new alternatives are highly desirable and can disburden the healthcare community.

In recent years deep learning (DL) algorithms have been extensively applied in different fields and their capabilities are proven. Different DL algorithms were applied in order to automatically diagnose diseases such as Covid-19 [4–6], brain tumor [7], coronary diseases [8,9], MRI images [10], cardiology [11], different types of cancer [12–15]. The studies on applying DL algorithms to medical image analysis reveal their potential in medical fields. The remarkable achievement of DL algorithms in medical fields makes them an efficient candidate for automatic diagnosing of skin cancer.

1.1. Literature survey

Recently many studies focused on classifying skin cancer images using DL. In Ref. [16], a deep convolutional neural network (CNN) is applied for automatically diagnosing skin cancer cases. In another study, thirteen CNNs are proposed as the ensemble of classifiers for detecting skin cancer [17]. In Ref. [18] deep learning and sound analysis algorithms are utilized to increase the accuracy of skin cancer classification. Particle swarm optimization and deep learning models are employed in Ref. [19] for skin cancer detection. In Ref. [20], Monte Carlo (MC) dropout, Ensemble MC and Deep Ensemble are utilized to quantify the uncertainty of the predicted outputs.

While many types of research focused on improving the accuracy of the predictions, the uncertainty quantification (UQ) of the predictions in

[*] Corresponding author.
E-mail address: abbas.khosravi@deakin.edu.au (P. Tabarisaadi).

https://doi.org/10.1016/j.compbiomed.2022.105357
Received 17 October 2021; Received in revised form 25 February 2022; Accepted 25 February 2022
Available online 2 March 2022

most of them remains unnoticed. The UQ of neural network predictions is of critical importance in real-world applications such as healthcare. Considering the uncertainty associated with the predictions, we may decide to ask an expert to investigate the samples that were labeled as uncertain by the model manually. The uncertainty contains epistemic and aleatoric uncertainties. The epistemic uncertainty that is mainly attributed to the data and knowledge deficiency, can be reduced by the collection of more quality data. Epistemic uncertainty illustrates how much the predictions on new samples can be trusted. The aleatoric uncertainty is attributed to the data's inherent randomness [21]. Another factor that is needed for a proper uncertainty-aware algorithm is being distanced aware. It means that the algorithm distinguishes when facing data that are far from those it was trained on.

The necessity of estimating the uncertainties associated with neural network predictions leads to the advent of different uncertainty-aware training algorithms. In Ref. [22] a bayesian approximation is proposed that is based on applying dropout before all the weight layers of a NN. In the MC dropout method, the test accuracy will not be sacrificed and the computing cost is reasonably low.

The parameters of Bayesian neural networks (BNNs) that were proposed in Ref. [23], are modeled with a probability distribution. As BNNs face infeasibility in large-scale problems, the Bayesian Ensembling proposed in Ref. [24] to address this problem. Authors in Ref. [25] proposed Spectral-normalized Neural Gaussian Process (SNGP). In the SNGP algorithm, a weight normalization step is added during training, and the output layer is replaced with a Gaussian Process. Low computing cost and the increased distance aware ability are two main factors that make the SNGP a suitable candidate for real-world applications.

This paper studies the automatic diagnosis of the skin cancer dataset. Mc dropout, Bayesian Ensembling, and SNGP algorithms are applied to the dataset separately. The uncertainties associated with the predictions are reported and the results are compared from different perspectives. The Cifar10 dataset is chosen as the out-of-domain distribution (OOD) data and the algorithms are compared in this regard. This can be considered as the main contribution of this paper. To the best of the author's knowledge, it has not been referenced in the literature yet.

The rest of this paper is organized as follows. Section II illustrates the uncertainty-aware algorithms utilized in this study. Section III, describes the uncertainty estimation, and some metrics for evaluation of uncertainty-aware algorithms is discussed. Simulation results and experiments of applying MC dropout, BE, and SNGP on the skin cancer dataset is reported in 4. And finally, section V concludes the paper.

2. Predictive uncertainty evaluation

2.1. Monte Carlo dropout

Dropout is a technique that was originally proposed in Ref. [26] to prevent overfitting. The main idea is omitting each neuron with the probability p during each training iteration that is equivalent to averaging over all possible networks during training [27]. In Ref. [22], dropout is interpreted as a Bayesian approximation. It is proved that applying the dropout layer before each weight layer in a neural network with arbitrary depth and non-linearities is equivalent to a variational approximation of a deep Gaussian Process.

The predictive mean and variance of the model can be written as follows [22]:

$$E[y] \approx \frac{1}{T}\sum\nolimits_{t=1}^{T} \widehat{y}_t(x) \tag{1}$$

$$Var_q(y) \approx \tau^{-1} I_D + \frac{1}{T}\sum\nolimits_{t=1}^{T} \widehat{y}_t(x)^T \widehat{y}_t(x) - E[y]^T E[y] \tag{2}$$

In the above equations, x is the input, $\widehat{y}_t(x)$ is the corresponding output and τ is a constant value defined based on the model structure and T is the number of stochastic forward passes.

The MC dropout technique received lots of attention and has been applied for many applications. One of the main advantages of this technique in comparison to its peers is that the training procedure remains unchanged, while other UQ methods impose considerable complexity on the training algorithm.

2.2. Bayesian Ensembling

In Bayesian neural networks (BNNs) that were first proposed in Ref. [23], the NN parameters are modeled by probability distributions that are computed with the Bayes rule. While BNNs work ideally for small-scale problems they face infeasibility for large-scale problems.

As an alternative, the Ensembling method is applied as a UQ technique. The variances of predictions of different NNs trained from different initialization are interpreted as the uncertainties of the predictions. Although the Ensembling performs efficiently, the main criticism is not being Bayesian.

The authors in Ref. [24] bridge the idea of BNN and Ensembling and propose the Bayesian Ensembling technique. Consider a NN with H hidden layer, parameters θ trained with N data points and the predicted outputs are $\widehat{y}$. Considering prior as a normal distribution (equation (3)).

$$P(\theta) = N(\mu_{prior}, \Sigma_{prior}) \tag{3}$$

the MAP parameter estimate is derived as follows:

$$\theta_{MAP} = argmax_\theta log(P_D(D|\theta)) - \frac{1}{2}\left\| \sum\nolimits_{prior}^{-1/2}(\theta - \mu_{prior}) \right\|_2^2 \tag{4}$$

Considering $\mu_{prior} = 0$, equation (4) is the $L2$ regularization. Now μ_{prior} is replaced with $\theta_{anc} = N(\mu_{prior}, \sum_{prior})$. For classification the loss function is defined as follows:

$$Loss = -\frac{1}{N}\sum\nolimits_{n=1}^{N}\sum\nolimits_{c=1}^{C} y_{n,c} log \widehat{y}_{n,c,j} + \frac{1}{N}\left\| \Gamma^{\frac{1}{2}}(\theta_j - \theta_{anc,j}) \right\|_2^2 \tag{5}$$

In equation (5), $y_{n,c}$ is the class probability produced by the classifier for sample n and class c and $diag(\Gamma)_i = 1/2\sigma^2_{prior_i}$.

2.3. Spectral-normalized Neural Gaussian Process

Authors in Ref. [25] focus on proposing a method that can effectively quantify the distance between the test samples and the training domain. The so-called Spectral-normalized Neural Gaussian Process (SNGP) method is proposed. In the SNGP method, a weight normalization step is added to the training algorithm that improves the ability of distance awareness. In addition, the Gaussian Process is utilized as the output layer. On the other hand, since the SNGP only needs one deep neural network, the computation cost is also relatively low. The low computation cost and high uncertainty quality estimation make the SNGP a suitable algorithm in practical cases.

Consider x, y, $P^*(y|x)$ respectively as input, output, and the data generating distribution that is the mixture of in domain (IND) and OOD distribution that is defined as equation (6).

$$P^*(y|x) = P^*(y, x \in \chi_{IND}|x) + P^*(y, x \notin \chi_{IND}|x) \tag{6}$$

The IND is a set where the input data $D = \{y_i, x_i\}_{i=1}^{N}$ are extracted from. Equation (6) can be written as follows:

$$\begin{aligned} P^*(y|x) = P^*(y|x, x \in \chi_{IND}) \times P^*(x \in \chi_{IND}) + \\ P^*(y|x, x \notin \chi_{IND}) \times P^*(x \notin \chi_{IND}) \end{aligned} \tag{7}$$

The predictive distribution can also be illustrate as:

$$\begin{aligned} P(y|x) = P(y|x, x \in \chi_{IND}) \times P^*(x \in \chi_{IND}) + \\ P_{uniform}(y|x, x \notin \chi_{IND}) \times P^*(x \notin \chi_{IND}) \end{aligned} \tag{8}$$

If the input data is from the training data domain (IND), the model is trusted and if the input is far from the training domain (OOD) a uniform prediction is used (8).

3. Uncertainty quantification

Consider an uncertainty-aware neural network that was already trained based on the training samples $D_{train} = \{X_{train}, Y_{train}\}$. Receiving the test samples $D_{test} = \{X_{test}, Y_{test}\}$, where $X_{test} = \{x_1, x_2, \ldots, x_n\}$ and $Y_{test} = \{y_1, y_2, \ldots, y_n\}$ illustrate the inputs and test labels respectively and n is the total number of test samples. In an uncertainty-aware NN the output of the model is presented as a probability distribution. The corresponding output for the $i'th$ input can be considered as $p_i(\hat{y}_i|x_i,\Theta)$. The entropy of the predictions $(\mathbb{H}_i)$ most of the time is interpreted as the uncertainty of the predictions.

$$\mathbb{H}_i(\hat{y}_i|x_i,\Theta) = -\sum_c p_i(\hat{y} = c|x_i,\Theta) logp(\hat{y}_i = c|x_i,\Theta) \tag{9}$$

In equation (9), c is the class that y can take. Θ denotes the model's parameters and, $\mathbb{H}_i(.)$ represents the uncertainty predicted for y_i.

The NN's prediction $\hat{Y} = \{\hat{y_1}, \hat{y_2}, \ldots, \hat{y_n}\}$ can be classified in two groups:

- correctly classified when $\hat{y}_i = y_i$
- incorrectly classified when $\hat{y}_i \neq y_i$

On the other hand, choosing a threshold value $\tau \in [0, 1]$, the test samples can be categorized into two groups:

- certainly classified when $\mathbb{H}_i(.) < \tau$
- uncertainly classified when $\mathbb{H}_i(.) \geq \tau$

So the test samples can be classified into four groups as follows:

- Correct and Certain (CC)
- Correct and Uncertain (CU)
- Incorrect and Certain (IC)
- Incorrect and Uncertain (IU)

Based on the above mentioned, the CC, CU, IC, and IU can be calculated as follows:

$$CC = \hat{y}_i = y_i \ \& \ \mathbb{H}_i(y_i|x_i,\Theta) < \tau \tag{10}$$

$$CU = \hat{y}_i = y_i \ \& \ \mathbb{H}_i(y_i|x_i,\Theta) \geq \tau \tag{11}$$

$$IC = \hat{y}_i \neq y_i \ \& \ \mathbb{H}_i(y_i|x_i,\Theta) < \tau \tag{12}$$

$$IU = \hat{y}_i \neq y_i \ \& \ \mathbb{H}_i(y_i|x_i,\Theta) \geq \tau \tag{13}$$

The number of samples in CC, CU, IC, and IU are illustrated by N_{CC}, N_{CU}, N_{IC} and N_{IU} respectively.

$$N_{CC} = \sum_i \mathbb{1}(\hat{y}_i = y_i \ \& \ \mathbb{H}_i(y_i|x_i,\Theta) < \tau) \tag{14}$$

$$N_{CU} = \sum_i \mathbb{1}(\hat{y}_i = y_i \ \& \ \mathbb{H}_i(y_i|x_i,\Theta) \geq \tau) \tag{15}$$

$$N_{IC} = \sum_i \mathbb{1}(\hat{y}_i \neq y_i \ \& \ \mathbb{H}_i(y_i|x_i,\Theta) < \tau) \tag{16}$$

$$N_{IU} = \sum_i \mathbb{1}(\hat{y}_i \neq y_i \ \& \ \mathbb{H}_i(y_i|x_i,\Theta) \geq \tau) \tag{17}$$

$$N_{CC} + N_{CU} + N_{IC} + N_{IU} = n \tag{18}$$

Ideally, it is desirable that an uncertainty-aware algorithm predicts all the labels certainly and correctly (CC). The second preference is the incorrect predictions that the model is uncertain about them (IU). After that, the correct and uncertain samples are considered as our third preference, which means the model correctly classified the samples but it is not sure about it. The less desirable category is incorrect and certain (IC), which means the model confidently makes wrong predictions. Based on the four groups presented, several metrics are defined to evaluate the performances of uncertainty-aware algorithms.

Table 1
Uncertainty confusion matrix.

		Correctness	
		Correct	Incorrect
Confidence	Certain	CC	IC
	Uncertain	CU	IU

3.1. Uncertainty confusion matrix

The uncertainty confusion matrix that is inspired by the concept of the confusion matrix is illustrated in Table 1. It is desired that the uncertainty-aware algorithm predict all the samples in CC and IU (green cells).

3.2. Correct certain ratio

The correct-certain ratio R_{CC} is defined as the ratio of correct and certain predictions to all certain predictions. As it is interesting that the model predicts confidently only for true predictions, the best value for R_{CC} is one.

$$R_{CC} = P(Correct|Certain) = \frac{P(Correct, Certain)}{P(Certain)} = \frac{N_{CC}}{N_{CC} + N_{IC}} \tag{19}$$

3.3. Incorrect uncertain ratio

The incorrect-uncertain ratio R_{IU} is defined as the ratio of incorrect and uncertain predictions to all uncertain predictions. As it is interesting that the model estimates high uncertainty for incorrect predictions, the best value for R_{IU} is one.

$$R_{IU} = P(Incorrect|Uncertain) \frac{P(Incorrect, Uncertain)}{P(Uncertain)} = \frac{N_{IU}}{N_{IU} + N_{CU}} \tag{20}$$

3.4. Uncertainty accuracy

Uncertainty accuracy (UA) is defined as the number of correct and certain and incorrect and uncertain predictions (favorable results) to all the samples. The best value for UA is one.

$$UA = \frac{N_{CC} + N_{IU}}{N_{CC} + N_{IU} + N_{CU} + N_{IC}} \tag{21}$$

4. Experiment

The skin cancer dataset 2017 ISIC challenge [28] contains three types of dermatologist spot images including melanoma, nevus, and seborrheic keratoses Fig. 1. Melanoma is cancerous and most lethal while nevus and seborrheic keratoses are categorized as benign (not cancerous). The dataset contains 2000 images for training, 600 images for testing, and 150 images for validation tasks, and all the images are normalized and transformed.

This study applies three different uncertainty-aware algorithms including Monte Carlo dropout, Bayesian Ensembling, and SNGP on the skin cancer dataset to automatically diagnose melanoma from benign images. The results are reported and compared from different

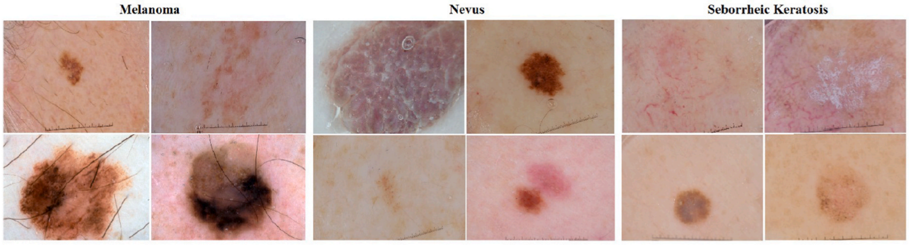


Fig. 1. A few examples of dermatologist spot images used in this study. Automatic diagnosis of melanoma from nevus and seborrheic keratoses is the main purpose of the used algorithms.

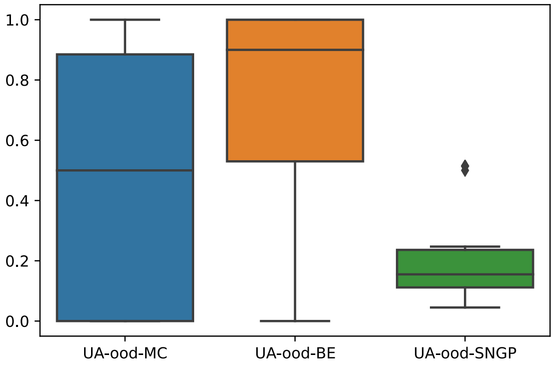


Fig. 2. The uncertainty accuracy values of OOD samples for Monte Carlo dropout, SNGP and Bayesian Ensembling for 30 runs are presented as boxplots.

perspectives.

In medical image analysis, most of the time large enough datasets are not available. Transfer learning is applied as a technique to address this issue. In transfer learning, a CNN that was already trained on a large and general dataset (most of the time a completely different dataset) is applied to extract the features. The output of the pre-trained model is used as the input of a different uncertainty-aware algorithm. It is worth mentioning that the threshold value is chosen at $\tau = 0.6$ in all algorithms.

4.1. Monte Carlo dropout

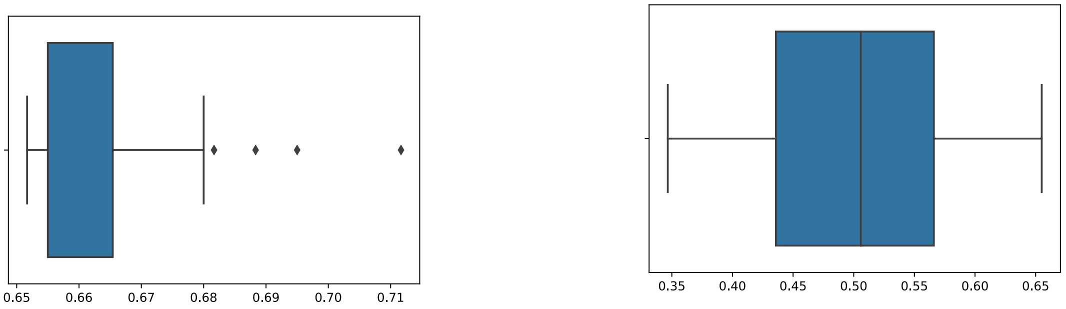
In order to apply MC dropout for skin cancer classification, transfer learning is applied instead of training the model from scratch. VGG16 that is also called OxfordNet [29] is applied as the pre-trained model. The extracted features of VGG16 are used as the input of the MC dropout model. The number of hidden layers is considered as 50 and the training number of epochs is chosen at 20. To be able to compare the algorithms more efficiently, all algorithms are run 30 times separately and the results are illustrated as a boxplot. The accuracy of the MC dropout algorithm for 30 runs is reported in Fig. 3a. It can be concluded that around 93% of the accuracy values are reported between 0.65 and 0.68. The uncertainty accuracy for the MC dropout is also reported in Fig. 3b. It allows the conclusion that around 75% of results are reported around 0.45 and 0.65.

To evaluate the MC dropout algorithm for OOD samples, the Cifar10 dataset is applied as the OOD data and the uncertainties are estimated for IND (test samples) and OOD samples. Fig. 8a illustrates the results of uncertainties estimation. One can assume that the MC dropout algorithm predicts relatively higher uncertainties when it faces OOD samples in comparison to the test samples which is a favorable feature in an uncertainty-aware algorithm which means that the algorithm successfully distinguishes the distance between the test samples and OOD data.

4.2. Bayesian Ensembling

The VGG16 is applied and the extracted features are fed to a Bayesian Ensembling classifier with 64 hidden layers and 8 ensembles for 100 epochs. The accuracy of 30 runs is illustrated in Fig. 4a. 75% of the reported results are between 0.35 and 0.65. The uncertainty accuracy of the BE algorithm for 30 runs is also reported in Fig. 4b. It can be estimated that around 93% of the reported results for uncertainty accuracy are between 0.35 and 0.75.

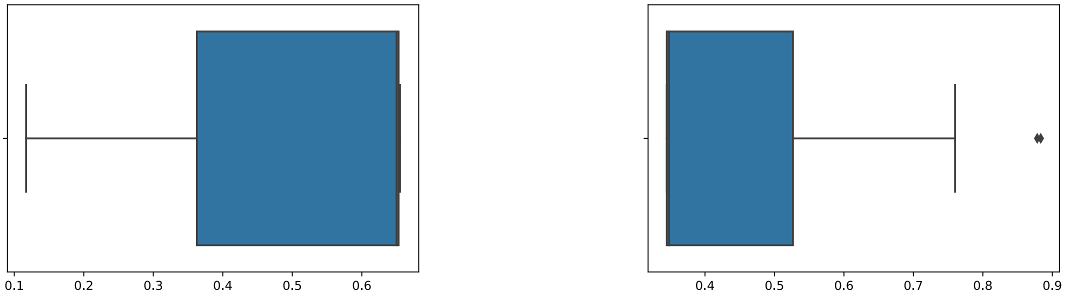
Applying Cifar10 samples as OOD data, the uncertainties quantified by the model for the test samples (IND) and OOD are reported in Fig. 8b. It is worth mentioning that the BE predicts relatively higher uncertainties for OOD data in comparison to IND samples. The uncertainties for the test samples are reported between 1 and 7 and almost



(a) The accuracy values of the MC dropout for 30 runs.

(b) The uncertainty accuracy values of the MC dropout for 30 runs.

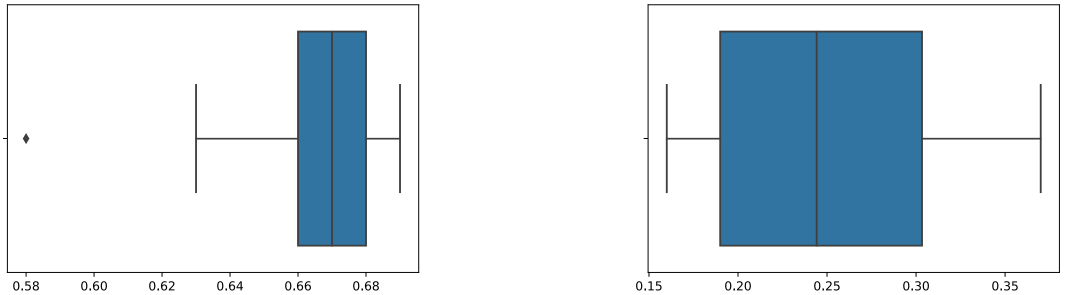
Fig. 3. The accuracy and uncertainty accuracy values of MC dropout for the skin cancer dataset are reported as boxplots. (a) The accuracy values of the MC dropout for 30 runs. (b) The uncertainty accuracy values of the MC dropout for 30 runs.



(a) The accuracy values of the BE for 30 runs.

(b) The uncertainty accuracy values of the BE for 30 runs.

Fig. 4. The accuracy and uncertainty accuracy values of BE for the skin cancer dataset are reported as boxplots. (a) The accuracy values of the BE for 30 runs. (b) The uncertainty accuracy values of the BE for 30 runs.



(a) The accuracy values of the SNGP for 30 runs.

(b) The uncertainty accuracy values of the SNGP for 30 runs.

Fig. 5. The accuracy and uncertainty accuracy values of SNGP for the skin cancer dataset are reported as boxplots. (a) The accuracy values of the SNGP for 30 runs. (b) The uncertainty accuracy values of the SNGP for 30 runs.

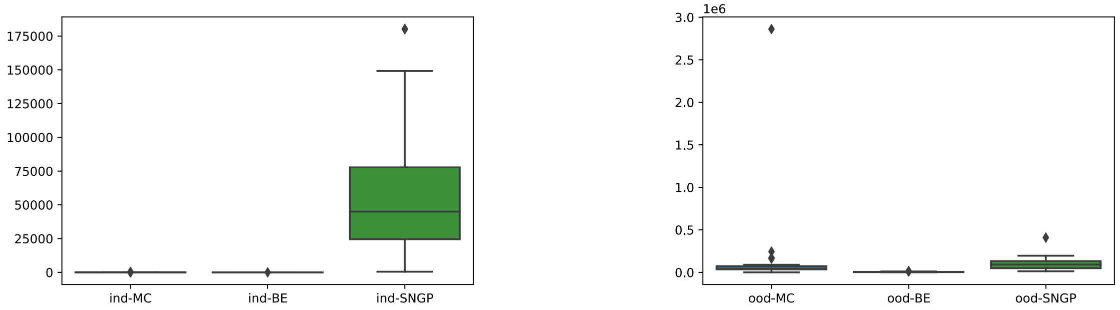
75% of the reported uncertainties are between 2 and 4. On the other hand, the uncertainties associated with OOD are reported between 1000 and 11000 and more than 75% of the reported uncertainties are between 2000 and 6000. It leads to the conclusion that the Bayesian Ensembling successfully detects the OOD data.

4.3. *Spectral-normalized Neural Gaussian Process*

For the third scenario, the SNGP is added to VGG16. The model is trained for 20 epochs. The accuracy of the predictions for 30 runs is reported in Fig. 5a. It can be concluded that 75% of the results are reported between 0.66 and 0.69 which is considered the best accuracy reported among its peers.

The uncertainty accuracy is also illustrated in Fig. 5b. It is worth mentioning that 75% of the results are reported between 0.2 and 0.37. One can conclude that the SNGP algorithm does not reach good performance in uncertainty estimation for the skin cancer dataset.

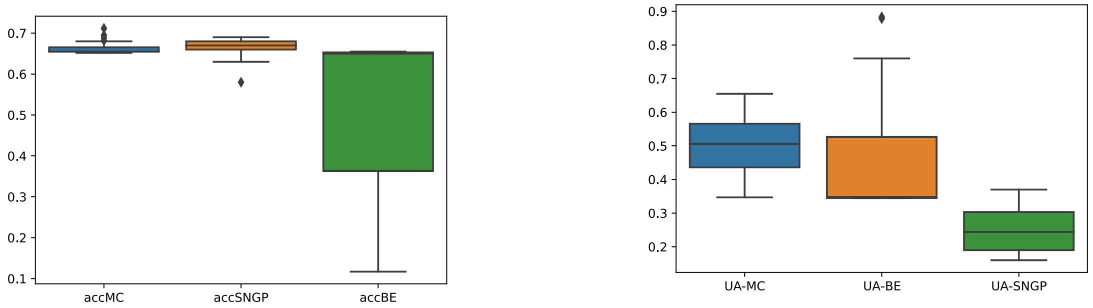
Considering Cifar10 as OOD, the uncertainties reported for OOD and IND data are compared in Fig. 8c. It can be concluded that most of the time higher uncertainties are reported for OOD data than the test samples. More than 75% of the predicted uncertainties are between 25000 and 150000 and for OOD data around 75% of the estimated



(a) The uncertainty values reported for IND.

(b) The uncertainty values reported for OOD.

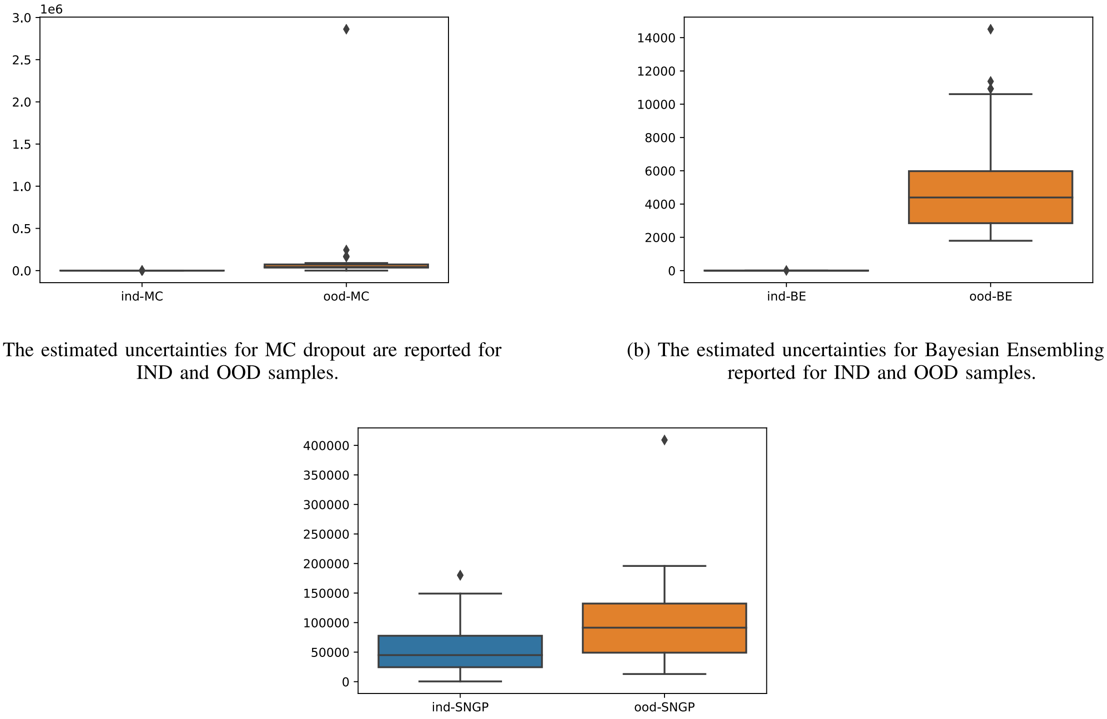
Fig. 6. The uncertainty values of IND and OOD samples for Monte Carlo dropout, SNGP and Bayesian Ensembling for 30 runs. (a) The uncertainty values reported for IND. (b) The uncertainty values reported for OOD.



(a) Comparison of the accuracy reported for all the algorithms.

(b) Comparison of the uncertainty accuracy reported for all the algorithms.

Fig. 7. The accuracy and uncertainty accuracy of Monte Carlo dropout, SNGP and Bayesian Ensembling for the skin cancer dataset for 30 runs. (a) Comparison of the accuracy reported for all the algorithms. (b) Comparison of the uncertainty accuracy reported for all the algorithms.



(a) The estimated uncertainties for MC dropout are reported for IND and OOD samples.

(b) The estimated uncertainties for Bayesian Ensembling are reported for IND and OOD samples.

(c) The estimated uncertainties for SNGP are reported for IND and OOD samples.

Fig. 8. Comparision of the uncertainties reported for IND and OOD data for all the algorithms. (a) The estimated uncertainties for MC dropout are reported for IND and OOD samples. (b) The estimated uncertainties for Bayesian Ensembling are reported for IND and OOD samples. (c) The estimated uncertainties for SNGP are reported for IND and OOD samples.

uncertainties are reported between 50000 and 200000.

Fig. 7a compares the accuracy of MC dropout, BE, and SNGP algorithms. It can be concluded that the SNGP reaches the best results in accuracy in comparison to its peers. The MC dropout also reaches competitive results. It seems that BE sometimes fails to reach competitive accuracy.

Fig. 7b compares the uncertainty accuracy reported for the three applied algorithms. The best result is related to MC dropout. The BE also achieves good results. While, the SNGP algorithm reaches good results for accuracy, for uncertainty accuracy, it fails to have a good performance. Considering the uncertainty accuracy and accuracy simultaneously, it can be concluded that the MC dropout is the best algorithm among the three uncertainty-aware algorithms used for diagnosing the skin cancer dataset.

Fig. 6a compares the estimated uncertainties for the test samples for all the applied algorithms. The SNGP algorithm obviously predicts higher uncertainties than BE and MC dropout. So it can be concluded that the SNGP algorithm is the most conservative predictor among the others.

Fig. 6b also compares the uncertainties reported by the three algorithms for OOD samples. It seems that the SNGP also predicts higher uncertainties for OOD data than its peers. So the SNGP is also the most conservative algorithm when it faces OOD data.

To compare the performance of the algorithms for OOD data the uncertainty accuracy is a reliable metric. Fig. 2 compares the uncertainty accuracy of OOD for the three algorithms. It can be concluded that the best result is related to BE and the worst result is related to MC dropout. It is worth mentioning that also the MC dropout is the best in predicting accuracy and uncertainty accuracy, for OOD the two other algorithms gain better results for uncertainty accuracy.

In summary, the best accuracy for the skin cancer classification is reported for the SNGP algorithm. The accuracy values for the MC dropout are also competitive. However, the BE sometimes fails to reach good results. Considering the reported uncertainty accuracy values, the MC dropout has the best performance and the SNGP significantly have low qualified uncertainty estimation. It can be concluded that the MC dropout is the best uncertainty-aware algorithm for the skin cancer dataset among its peers.

For an uncertainty-aware algorithm, it is desirable that the algorithm predicts significantly higher uncertainty values for OOD in comparison to the test samples (IND). It can be interpreted as the algorithm being able to distinguish the data that are different from those used during training. The Cifar10 is chosen as OOD since it is completely different from the training set and it is worth checking the performance of the algorithms when facing such irrelevant data. It seems that all the algorithms are able to identify OOD data. It should be mentioned that also the SNGP generally predicts higher uncertainty values for OOD than IND data, sometimes it fails. For OOD data the MC dropout sometimes fails to reach good results for UA and the SNGP also has low performance. It leads to the conclusion that the Bayesian Ensembling reaches the best in class UA for OOD data.

The appropriate uncertainty-aware algorithm for automatic diagnosing of skin cancer can be chosen based on considering the situation, priorities, and demands.

5. Conclusion

This paper addresses the automatic diagnosing of skin cancer dermatologist images. The main purpose is to not only accurately classify images but also measure and communicate the lack of confidence in predictions generated by deep neural networks. In Clinical trials of artificial intelligence, the uncertainty quantification of the predictions is of high importance. Predictions with a high level of uncertainty can then be flagged as unreliable and a second opinion from an expert can be requested for them. Three effective uncertainty-aware algorithms including Monte Carlo dropout, Bayesian Ensembling, and spectral-normalized neural Gaussian process are applied separately. The three algorithms are evaluated in uncertainty estimation. In addition, the performances of the algorithms for quantifying the distance between the test samples and the training domain are evaluated and compared. Obtained results indicate that the SNGP reaches the best in class accuracy however it performs poorly in uncertainty accuracy. On the other hand, the Bayesian Ensembling reaches acceptable results for uncertainty accuracy but does not have good performance on accuracy. It can be concluded that the Monte Carlo dropout has the best performance in estimating the accuracy and uncertainty accuracy simultaneously. It is worth mentioning that for OOD data, Bayesian Ensembling has better performance in estimating the uncertainty accuracy in general. The most suitable algorithm for a specific scenario can be chosen by considering the priorities and conditions of the automatic diagnosing systems.

Acknowledgment

This research was partially supported by the Australian Research Council's Discovery Projects funding scheme (project DP190102181 and DP210101465).

References

[1] wcrf, "skin-cancer-statistics. https://www.wcrf.org/dietandcancer/skin-cancer-statistics/, 2021.
[2] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, F. Bray, Global Cancer Observatory: Cancer Today, International Agency for Research on Cancer, Lyon, France, 2018.
[3] N.C. Institute, Surveillance, Epidemiology, and End Results (SEER) Program, 2018.
[4] P. Tabarisaadi, A. Khosravi, S. Nahavandi, A deep bayesian ensembling framework for COVID-19 detection using chest CT images, in: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2020, pp. 1584–1589.
[5] A. Shamsi, H. Asgharnezhad, S.S. Jokandan, A. Khosravi, P.M. Kebria, D. Nahavandi, S. Nahavandi, D. Srinivasan, An uncertainty-aware transfer learning-based framework for COVID-19 diagnosis, IEEE Transact. Neural Networks Learn. Syst. 32 (4) (2021) 1408–1417.
[6] M. Abdar, S. Salari, S. Qahremani, H.-K. Lam, F. Karray, S. Hussain, A. Khosravi, U. R. Acharya, S. Nahavandi, UncertaintyFuseNet: Robust Uncertainty-Aware Hierarchical Feature Fusion with Ensemble Monte Carlo Dropout for COVID-19 Detection, 2021 arXiv preprint arXiv:2105.08590.
[7] S. Bauer, R. Wiest, L.-P. Nolte, M. Reyes, A survey of MRI-based medical image analysis for brain tumor studies, Phys. Med. Biol. 58 (13) (2013) R97.
[8] R. Alizadehsani, A. Khosravi, M. Roshanzamir, M. Abdar, N. Sarrafzadegan, D. Shafie, F. Khozeimeh, A. Shoeibi, S. Nahavandi, M. Panahiazar, et al., Coronary artery disease detection using artificial intelligence techniques: a survey of trends, geographical differences and diagnostic features 1991-2020, Comput. Biol. Med. (2020) 104095.
[9] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P.M. Kebria, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, U.R. Acharya, Machine learning-based coronary artery disease diagnosis: a comprehensive review, Comput. Biol. Med. 111 (2019) 103346.
[10] M.A. Mazurowski, M. Buda, A. Saha, M.R. Bashir, Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI, J. Magn. Reson. Imag. 49 (4) (2019) 939–954.
[11] P. Bizopoulos, D. Koutsouris, Deep learning in cardiology, IEEE Rev. Biomed. Eng. 12 (2018) 168–193.
[12] Z. Alyafeai, L. Ghouti, A fully-automated deep learning pipeline for cervical cancer classification, Expert Syst. Appl. 141 (2020) 112951.
[13] O.-J. Skrede, S. De Raedt, A. Kleppe, T.S. Hveem, K. Liestøl, J. Maddison, H. A. Askautrud, M. Pradhan, J.A. Nesheim, F. Albregtsen, et al., Deep learning for prediction of colorectal cancer outcome: a discovery and validation study, Lancet 395 (10221) (2020) 350–360.
[14] M. Saha, I. Arun, R. Ahmed, S. Chatterjee, C. Chakraborty, HscoreNet: a Deep network for estrogen and progesterone scoring using breast IHC images, Pattern Recogn. 102 (2020) 107200.
[15] R. Stoean, Analysis on the potential of an EA–surrogate modelling tandem for deep learning parametrization: an example for cancer classification from medical images, Neural Comput. Appl. 32 (2) (2020) 313–322.
[16] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.
[17] A.G.C. Pacheco, A.-R. Ali, T. Trappenberg, Skin Cancer Detection Based on Deep Learning and Entropy to Detect Outlier Samples, 2019 arXiv preprint arXiv:1909.04525.
[18] A. Dascalu, E.O. David, Skin cancer detection by deep learning and sound analysis algorithms: a prospective clinical study of an elementary dermoscope, EBioMedicine 43 (2019) 107–113.
[19] T.Y. Tan, L. Zhang, C.P. Lim, Intelligent skin cancer diagnosis using improved particle swarm optimization and deep learning models, Appl. Soft Comput. 84 (2019) 105725.
[20] M. Abdar, M. Samami, S.D. Mahmoodabad, T. Doan, B. Mazoure, R. Hashemifesharaki, L. Liu, A. Khosravi, U.R. Acharya, V. Makarenkov, et al., Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning, Comput. Biol. Med. (2021) 104418.
[21] J. Postels, F. Ferroni, H. Coskun, N. Navab, F. Tombari, Sampling-free epistemic uncertainty estimation using approximated variance propagation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2931–2940.
[22] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.
[23] D.J.C. MacKay, A practical Bayesian framework for backpropagation networks, Neural Comput. 4 (3) (1992) 448–472.
[24] T. Pearce, F. Leibfried, A. Brintrup, Uncertainty in neural networks: approximately bayesian ensembling, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 234–244.
[25] J.Z. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax-Weiss, B. Lakshminarayanan, Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness, 2020 arXiv preprint arXiv:2006.10108.
[26] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving Neural Networks by Preventing Co-adaptation of Feature Detectors, 2012 arXiv preprint arXiv:1207.0580.

[27] A. Labach, H. Salehinejad, S. Valaee, Survey of Dropout Methods for Deep Neural Networks, 2019 arXiv preprint arXiv:1904.13310.

[28] N.C.F. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. 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), IEEE, 2018, pp. 168–172.

[29] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014 arXiv preprint arXiv:1409.1556.