Computers in Biology and Medicine 135 (2021) 104418

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning

Moloud Abdar [a,*], Maryam Samami [b], Sajjad Dehghani Mahmoodabad [c], Thang Doan [d], Bogdan Mazoure [d], Reza Hashemifesharaki [e], Li Liu [f], Abbas Khosravi [a], U. Rajendra Acharya [g,h,i], Vladimir Makarenkov [j], Saeid Nahavandi [a]

[a] Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Geelong, Australia
[b] Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran
[c] Department of Artificial Intelligence, Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran
[d] Department of Computer Science, McGill University / Mila, Montreal, Canada
[e] Department of Research and Development, Mute Hammer LLC., Santa Monica, USA
[f] Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland
[g] School of Engineering, Ngee Ann Polytechnic, Singapore
[h] Department of Biomedical Engineering, Singapore University of Social Sciences, Singapore
[i] Department of Biomedical Informatics and Medical Engineering, Asia University, Taichung, Taiwan
[j] Department of Computer Science, University of Quebec in Montreal, Montreal, Canada

ARTICLE INFO

Keywords:
Medical image classification
Uncertainty quantification (UQ)
Deep learning
Bayesian deep learning
Monte Carlo dropout
Skin cancer

ABSTRACT

Accurate automated medical image recognition, including classification and segmentation, is one of the most challenging tasks in medical image analysis. Recently, deep learning methods have achieved remarkable success in medical image classification and segmentation, clearly becoming the state-of-the-art methods. However, most of these methods are unable to provide uncertainty quantification (UQ) for their output, often being over-confident, which can lead to disastrous consequences. Bayesian Deep Learning (BDL) methods can be used to quantify uncertainty of traditional deep learning methods, and thus address this issue. We apply three uncertainty quantification methods to deal with uncertainty during skin cancer image classification. They are as follows: Monte Carlo (MC) dropout, Ensemble MC (EMC) dropout and Deep Ensemble (DE). To further resolve the remaining uncertainty after applying the MC, EMC and DE methods, we describe a novel hybrid dynamic BDL model, taking into account uncertainty, based on the Three-Way Decision (TWD) theory. The proposed dynamic model enables us to use different UQ methods and different deep neural networks in distinct classification phases. So, the elements of each phase can be adjusted according to the dataset under consideration. In this study, two best UQ methods (*i.e.*, DE and EMC) are applied in two classification phases (the first and second phases) to analyze two well-known skin cancer datasets, preventing one from making overconfident decisions when it comes to diagnosing the disease. The accuracy and the F1-score of our final solution are, respectively, 88.95% and 89.00% for the first dataset, and 90.96% and 91.00% for the second dataset. Our results suggest that the proposed TWDBDL model can be used effectively at different stages of medical image analysis.

1. Introduction

Accurate and automated medical image classification and segmentation are two extremely important procedures used in clinical research. However, it can be argued that traditional methods performing these tasks are not always efficient when handling large volumes of data. Over the last few years, a large number of traditional machine learning and deep learning methods have been extensively used to perform diagnosis of different diseases, including cancers (*e.g.*, autism spectrum disorder [49], cervical cancer [4], colorectal cancer [50,56,65], coronary artery disease (CAD) [1,72], brain tumour [60], diabetic retinopathy (DR) [16], breast cancer [46], etc).

Despite the state-of-the-art performance of various deep learning methods used in medical image classification and segmentation, they

* Corresponding author.
E-mail address: m.abdar1987@gmail.com (M. Abdar).

https://doi.org/10.1016/j.compbiomed.2021.104418
Received 3 January 2021; Received in revised form 1 April 2021; Accepted 17 April 2021
Available online 28 April 2021

rarely provide uncertainty quantification (estimation) in their outputs, *i.e.*, the model's (called epistemic uncertainty) and data (called aleatoric uncertainty) uncertainties [16,60]. In other words, a blind trust in results obtained by traditional machine learning and deep learning methods can lead to the loss of some sensitive and important features and information [11]. Thus, deep learning-based systems inherently do not calculate any uncertainty related to the model's prediction nor highlight the most important input features for a specific prediction. This is considered as a lack of theoretical understanding of the underlying mechanics of deep learning methods. As a result, they are often referred to as "black boxes" [3,65]. This issue motivated us to design a novel uncertainty quantification (UQ) technique [2].

Generally, it is challenging to train and validate different traditional machine learning and deep learning methods under uncertainty. In this work, we will propose a model for quantifying uncertainty of deep learning predictions carried out for medical image data. Uncertainty is an authoritarian manner to distinguish *what to keep* and *what to modify* as we continuously learn different things in our daily life, and therefore attenuate catastrophic forgetting [14]. Inspired by commonly used UQ techniques and a variety of deep learning architectures, we will propose a hybrid UQ model based on the Three-Way Decisions (TWD) theory [5] and the Bayesian deep learning (BDL) approach. In this regard, we will first apply several well-known UQ methods and then improve the UQ results using the TWD theory. To the best of our knowledge, this is the first study leveraging the TWD theory to enhance the reliability and understanding of the results of deep learning-based methods applied to medical image data.

1.1. Main contributions

Although Deep Neural Networks (DNNs) often provide practitioners and doctors with reasonable predictions in diagnosing a disease, they usually focus on improving accuracy, regardless of estimating uncertainty in the decision making process. As a result, we propose a three-phase uncertainty estimation model to prevent overconfidence in prediction. In summary, the main contributions of our work are as follows:

- Novel hybrid deep learning model for classification of medical images is presented;
- We apply three well-known UQ techniques: Monte Carlo (MC) dropout, Ensemble MC dropout (EMC), and Deep Ensemble (DE), and then use the two best of them (*i.e.*, DE and EMC) in the main model;
- In order to provide an accurate prediction model and further uncertainty estimation, we integrate the UQ methods in it using the TWD theory;
- We define a new certainty threshold for the confidence level of both classification phases;
- Using TWD, we combine two well-known UQ techniques (*i.e.*, EMC and DE);
- We optimize the hyperparameters of deep learning methods using a Bayesian Optimization (BO) technique [53] and network morphisms [63];
- The developed model is tested on two popular skin cancer image datasets; its competitiveness is demonstrated along with the results of uncertainty estimation.

Along with these contributions, we also provide a literature review on uncertainty quantification methods applied to analyze medical data. The rest of the study is organized as follows. In Section 2, we briefly review the application of deep learning and UQ methods in image classification. An in-depth background of the applied methods is presented in this section, whereas some additional details regarding the proposed UQ model are discussed in Section 3. The experimental validation of the proposed model is described in Section 4. The discussion on the obtained results is presented in Section 5. Lastly, the main conclusions are drawn in Section 6.

2. Related work and background

In this section, we briefly review a number of recent studies concerning the use of UQ and deep learning methods in medical data analysis. Then, relevant background of UQ methods is discussed.

2.1. Uncertainty quantification in medical image analysis

Uncertainty Quantification (UQ) is a key factor for an accurate application of machine learning and deep learning methods. A UQ estimation can increase the confidence of results provided by these methods [33]. For example, Bayesian deep learning (BDL) algorithms [16] are well-known methods used to estimate uncertainty. Filos et al. [16] introduced BDL benchmarks which include MC dropout [18], MFVI (Mean-Field Variational Inference) [64], EMC (Ensemble MC) dropout [51], DEs (Deep Ensembles) [30] and α-divergence methods [22,35]. Among others, BDL methods have been tested on Diabetic Retinopathy (DR) data, for which EMC and MC have achieved the best performance in terms of the accuracy and the area under the curve (AUC) metric. Next, the work by Leibig et al. [32] leveraged uncertainty information from deep neural networks (DNNs) used for diabetic detection. A common approach to train Bayesian Convolutional Neural Networks (BCNNs) is the Bernoulli approximate variational inference (BAVI). Leibig et al. [32] computed consequential uncertainty estimations of BCNNs without demanding further labels for obvious uncertain image category. No need for more labels can be considered a good achievement for this approach. Wickstrom et al. [65] employed a CNN model for semantic segmentation of colorectal cancer images. Their results demonstrated that the obtained uncertainty differed significantly for correct and false predictions. In the work of Carneiro et al. [10], an automated-based polyp classification (a five-class classification task) of colonoscopy images using deep learning methods (DenseNet and ResNet) has been introduced. Furthermore, by using confidence calibration (CC), these authors increased the classification Entropy and decreased the standard deviation (STD) of estimated variance. According to the results, the applied uncertainty estimation and confidence calibration algorithms led to a better performance of deep learning methods used for colonoscopy image classification.

2.2. Deep learning-based medical image classification

As mentioned earlier, different deep learning methods have shown outstanding performance not only in medical image classification but also in other domains. In this sub-section, we briefly discuss some of the existing studies which have considered different deep learning methods to perform medical image classification. Ghoneim et al. [21] applied convolutional neural networks (CNNs) and extreme learning machines (ELMs) for classification of cervical cancer. The proposed hybrid CNN-ELM-based model achieved 99.50% and 91.20% accuracies for 2-way and 7-way classification tasks, respectively. Saha et al. [46] introduced a new deep network, named *HscoreNet*, for scoring (scoring layer was a novel concept here) of estrogen and progesterone in breast IHC (Immunohistochemistry) images. Rubin et al. [45] proposed a new deep learning algorithm for classification of a small training sets (label-free cancer cell data), called the transferring of pre-trained generative adversarial networks (TOP-GANs). Coudray et al. [13] applied a CNN model for classification, as well as mutation prediction, of non–small cell lung cancer (histopathology) images. The proposed method automatically classified lung cancer images into three classes: LUAD (Adenocarcinoma), LUSC (squamous cell carcinoma) and normal lung tissues. Based on the obtained results, the authors have emphasized the effectiveness of deep learning in the classification of lung cancer images. Another common challenge in medical image analysis is the collection of high-quality medical image labels which is usually an

expensive, laborious and time consuming task, needing a clinical expertise.

It should be noted that some extensive studies focusing on the ability of UQ method to deal with uncertainties in traditional machine learning and deep learning have been conducted. Indeed, uncertainties in the prediction stages of machine learning methods have been well studies and are often taken into account. In this work, we intend to use the TWD theory to examine uncertainties in the decision stage of deep learning methods in the field of medical image classification.

2.3. Bayesian Deep Learning

In this section, preliminaries of Bayesian neural networks (BNNs) and uncertainty types are reviewed. We discuss an important background features of BNNs as well as those of the approximate variational inference (VI) approach. We have also provide necessary background for UQ and then describe our novel UQ model in the following sub-sections.

Bayesian Neural Networks: The use of the Bayesian model averaging allows one to estimate uncertainty through assigning a distribution over model parameters, followed by the parameters marginalization in order to build a predictive distribution [34]. Since BNNs scale well to high dimensional inputs, such as images, here we focus on BNNs, which are robust to overtting [26]. Predicting the machine learning model outputs using uncertainty estimation on a single observation needs a distribution over possible outcomes. The estimation of uncertainties using Bayesian approaches has been used in many studies for evaluating the validity of different clinical predictions [27,32,61], including large-scale real-world problems [23,32,54]. Given training inputs $X = \{x_1, \ldots, x_N\}$ and their associated outputs $Y = \{y_1, \ldots, y_N\}$, in a Bayesian regression, the parameters ω of the function $y = f^{\omega}(x)$ are inferred. Some prior distributions are used in the space of parameters, $p(\omega)$. Moreover, we need to dene a likelihood distribution $p(y|x,\omega)$. For example, classication models often consider a Softmax likelihood [19]:

$$p(y = d|x, \omega) = \text{Categorical}\left(\frac{\exp\left(f_d^{\omega}(x)\right)}{\sum_{d'} \exp\left(f_{d'}^{\omega}(x)\right)}\right). \tag{1}$$

Let us now consider the posterior distribution through the space of parameters: $p(\omega|X, Y)$, given a dataset X,Y. Using this distribution, the output corresponding to a given input point x^* can be predicted by integrating [19].

$$p(y^*|x^*, X, Y) = \int p(y^*|x^*, \omega) p(\omega|X, Y) d\omega. \tag{2}$$

Placing a prior distribution on the neural network (NN) weights defines a BNN. According to the weight matrices W_i and the bias vectors b_i for the layer i, the standard Gaussian prior distributions are usually assigned to the weight matrices, $p(W_i) = N(0, 1)$ [19].

Uncertainty: Epistemic and aleatoric are two major types of uncertainties that have been extensively studied in the literature. Indeed, we face the *epistemic uncertainty* when a probabilistic model is uncertain and has many possible predictions for the input that can be decreased by observing more data. It is often referred to as model uncertainty [7,16,24]. On the other hand, we have the *aleatoric uncertainty* if the input is not clear and cannot be decreased by gathering more data. In this paper, we provide a novel, simple, but still efficient approach to further address the issue of epistemic uncertainty in deep learning models used for binary classification of medical images.

Epistemic Uncertainty: To assess epistemic uncertainty in NN models, a prior distribution is placed over the model weights, which are named BNNs. BNNs apply distributions over these parameters instead of using the deterministic weight parameters of NNs. Afterwards, all possible weights are averaged. This process is referred to as marginalization. If the random output, BNN is illustrated as $f^W(x)$, the model's likelihood can be shown as $p(y|f^W(x))$. Given a dataset $X = \{x_1, \ldots, x_N\}$, $Y = \{y_1, \ldots, y_N\}$, Bayesian inference is applied to calculate the posterior over the weights $p(W|X, Y)$. The posterior takes some plausible parameters of the model. In terms of classification, the output of the model can be computed using a Softmax function. We can, therefore, sample from the probability vector of the result: $p(y|f^W(x)) = Softmax(f^W(x))$. In some approximate inference methods [24], the posterior $p(W|X, Y)$ is fitted with a sample distribution $q_\theta^*(W)$, which is parameterized by θ. To implement the inference, at first, a model should be trained with dropout before any weight layer, and then dropout layer, could be applied at the test stage in order to sample from the approximate posterior. Dropout is considered as a variational Bayesian approximation, where the distribution of approximating is a combination of two Gaussians, which have low variances, and the mean value of one of the Gaussians is zero. The objective function to be minimized (for N data points) can be formulated as follows:

$$L_{dropout}(\theta, p_{drop}) = \frac{-1}{N} \sum_{i=1}^{N} \log p\left(y_i \middle| f^{\widehat{W}_i}(x_i)\right) + \frac{1-p}{2N} \left|\left|\theta\right|\right|^2. \tag{3}$$

It is worth noting that $\widehat{W}_i \sim q_\theta^*(W)$ denotes samples and θ denotes the set of distribution parameters. The epistimic uncertainty induces the uncertainty of the classifier's prediction by marginalizing through the posterior distribution of weights. This uncertainty can be estimated in classification tasks using the MC integration, as follows:

$$p = (y = c|x, X, Y) \approx \frac{1}{T} \sum_{t=1}^{T} softmax\left(f^{\widehat{W}_t}(x)\right). \tag{4}$$

where $q_\theta(W)$ is the Dropout distribution [7,24].

Uncertainty Estimator. The uncertainty of our binary classication is computed by predictive Entropy [17,48], taking into account the average of information which is available in the predictive distribution:

$$H_{pred}(y|x) : = -\sum_{c} p(y = c|x) \log p(y = c|x). \tag{5}$$

The sum is taken over all possible existing classes c (in our case $c \in 0, 1$). The value of H_{pred} is high, if either the value of aleatoric uncertainty or of epistimic uncertainty is high. The probability $p(y = c|x)$ is approximated by T Monte Carlo samples, $\frac{1}{T}\sum_t p_\theta(y = c|x)$. This is achieved by stochastic forward passes through the probabilistic networks. It is worth noting that it is a biased (consistent) estimator of the predictive Entropy in equation (2) [17].

Monte Carlo Dropout (MC). Gal and Ghahramani [18] proposed a Bayesian theory of dropout. They revealed that optimising NNs by dropout (as a standard regularization technique) [55] and L2-regularization can be equivalent to a type of variational inference in a Bayesian machine learning model. Bayesian machine learning model gives as output the probability distribution. The variance of the output probability distribution is predicted to determine the uncertainty of the model for an input sample. Output samples are considered as Monte Carlo samples which are drawn from the posterior distribution of the models by applying standard dropout on DNNs at the test stage. A DNN should be trained using dropout in order to implement MC. Secondly, to compute the inference on each input, the DNN method is carried out T times by applying dropout at the test phase. At any time, the input image is the same but has a different randomly generated dropout mask. The estimators in the case of the mean and the variance of the outputs of Bayesian models are computed as follows:

$$E[y] \approx \frac{1}{T} \sum_{t=1}^{T} \widehat{y}_t(X), \tag{6}$$

$$Var[y] \approx \tau^{-1} I_D + \frac{1}{T} \sum_{t=1}^{T} \widehat{y}_t(X)^T \widehat{y}_t(X) - E[y]^T E[y], \tag{7}$$

Table 1
Illustration of the optimized values of the hyperparameters of different deep learning architectures after using Bayesian optimization for the first dataset.

Hyperparameter	DenseNet201	MobileNetV2	ResNet152V2	InceptionResNetV2
$Units_0$	256	512	320	256
$Activation_0$	Sigmoid	SWish	Swish	Relu
$Units_1$	64	128	128	96
$Activation_1$	Swish	Sigmoid	Sigmoid	Sigmoid
Optimizer	SGD	Adam	Adam	SGD
Learning rate	0.01	0.02	0.0001	0.01
Epoch	31	20	21	22

Table 2
Illustration of the optimized values of the hyperparameters of different deep learning architectures after using Bayesian optimization for the second dataset.

Hyperparameter	DenseNet201	MobileNetV2	ResNet152V2	InceptionResNetV2
$Units_0$	192	192	192	128
$Activation_0$	Relu	Relu	Swish	Relu
$Units_1$	128	64	96	94
$Activation_1$	Relu	Relu	Relu	Relu
Optimizer	Adam	RMSprop	RMSprop	RMSprop
Learning rate	0.01	0.001	0.001	0.001
Epoch	31	20	21	22

Table 3
Comparison of the obtained results of different deep learning methods and three types of UQ methods for the first skin cancer dataset. In this table, Entropy-correct (EC) - indicates the Entropy of correctly classified samples, Entropy-incorrect (EI) - indicates the Entropy of misclassified samples, STD-correct (STD-Co) - indicates the STD of correctly classified samples, STD-incorrect (STD-In) - indicates the STD of misclassified samples.

Method	EC	EI	STD-Co	STD-In	Accuracy (%)	F1-score (%)	AUC
DenseNet + MC	1.974	2.694	0.2488	0.42702	86.15	86.56	0.9420
DenseNet + EMC	1.992	2.687	0.25473	0.4266	85.19	85.16	0.9489
DenseNet + DE	0.4898	0.5162	0.03410	0.1303	89.00	89.08	0.9133
ResNet152V2 + MC	2.0807	3.045	0.2814	0.4772	85.90	85.88	0.9325
ResNet152V2 + EMC	2.792	4.166	0.2553	0.4777	85.45	84.39	0.9412
ResNet152V2 + DE	0.4869	0.5571	0.0497	0.1734	87.42	87.40	0.9077
MobileNetV2 + MC	2.113	3.180	0.2742	0.4512	86.66	86.64	0.9161
MobileNetV2 + EMC	2.746	4.046	0.2396	0.4116	86.20	86.05	0.9418
MobileNetV2 + DE	0.4958	0.6093	0.03971	0.1057	88.48	88.46	0.9014
InceptionResNetV2 + MC	2.076	3.0589	0.2711	0.4497	85.00	84.96	0.9190
InceptionResNetV2 + EMC	2.78493	3.98824	0.24678	0.438516	83.66	81.53	0.9321
InceptionResNetV2 + DE	0.4939	0.5334	0.03806	0.1633	85.45	85.43	0.8930

where $\hat{y}_t(x)$ is the output of the given DNN input sample x, t is a set of dropout masks and τ is a constant selected depending of the model's structure [29].

Deep Ensemble. Lakshminarayanan et al. [30] developed the DE model as a simple and scalable alternative to BNNs. This method computes the uncertainty by collecting various predictions that are achieved by T several individual deterministic models, whose parameters are initialized to random starting points, trained independently [30]. The DE models are not only easy to run, but also parallelizable. They yield more accurate estimations of uncertainty. Despite these advantages, they are usually very computationally expensive [16].

Ensemble Monte Carlo (EMC) Dropout. An Ensemble Monte Carlo (EMC) dropout uses several different MCD models in parallel. By sampling repeatedly from all of the ensemble members (individual models), which require to apply dropout masks for each sample, estimations of each individual models are computed. Afterwards, those estimations are averaged to get the final result of EMC. It is worth noting that EMC dropout is combined by using the two following UQ methods: MC

Table 4
Comparison of the obtained results of different deep learning methods and three types of UQ methods for the second skin cancer dataset. In this table, Entropy-correct (EC) - indicates the Entropy of correctly classified samples, Entropy-incorrect (EI) - indicates the Entropy of misclassified samples, STD-correct (STD-Co) - indicates the STD of correctly classified samples, STD-incorrect (STD-In) - indicates the STD of misclassified samples.

Method	EC	EI	STD-Co	STD-In	Accuracy (%)	F1-score (%)	AUC
DensNet + MC	1.562	1.626	0.5075	0.5615	83.20	83.13	0.6999
DensNet + EMC	2.738	2.793	0.5881	0.6181	81.48	81.48	0.9031
DensNet + DE	0.5435	0.5354	0.0647	0.1170	71.68	71.66	0.7436
$ResNet152V2_2$ + MC	1.623	1.771	0.4907	0.5712	87.33	87.25	0.9435
ResNet152V2 + EMC	2.640	2.765	0.5582	0.6061	79.93	79.91	0.8944
ResNet152V2 + DE	77.98	0.5571	0.0497	0.1734	77.98	77.96	0.8116
MobileNetV2 + MC	1.764	1.614	0.5357	0.5274	44.30	44.00	0.1334
MobileNetV2 + EMC	2.829	2.845	0.6128	0.5984	29.48	28.82	0.2351
MobileNetV2 + DE	0.5639	0.6043	0.0686	0.1175	71.95	71.89	0.7423
InceptionResNetV2 + MC	1.608	1.7044	0.5182	0.5307	47.89	47.71	0.5012
InceptionResNetV2 + EMC	2.753	2.779	0.6056	0.5885	31.41	31.32	0.2622
InceptionResNetV2 + DE	0.5491	0.5507	0.0577	0.1060	74.22	74.21	0.7656

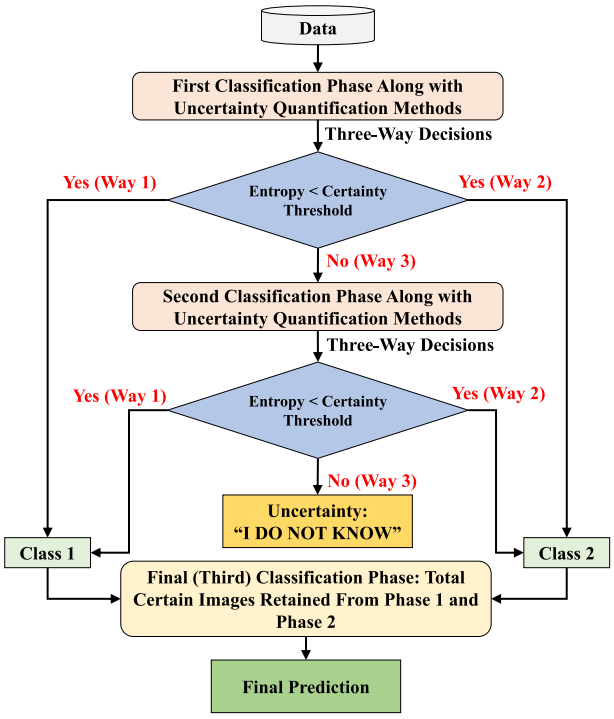


Fig. 1. A general view of the proposed dynamic TWDBDL model.

dropout and Model Ensembling [16].

3. The proposed uncertainty quantification model

In the following, we provide a general explanation of the Three-Way Decision (TWD) theory. Thereafter, we present our new model, called Three-Way Decision-based Bayesian Deep Learning (TWDBDL), for quantifying uncertainties in skin cancer image classification.

3.1. Basic uncertainty quantification methods

Three well-known uncertainty methods, *i.e.*, MC, DE and EMC dropout, will be employed and tested on two different skin cancer datasets. Then, their performances obtained for the first and second datasets will be reported and compared in Table 3 and Table 4, respectively, using various metrics to identify the best UQ method. Thus, for the first dataset, the best two of the three applied UQ methods are used as basic UQ methods in two different classification phases of our TWDBDL model (DE in phase 1 and EMC in phase 2). For for the second dataset, we use EMC only in the both phases because MC and DE provide very mediocre results.

It is important to note that the UQ model we consider here is a dynamic UQ model. This means that the number of classification phases (in this study we have three main phases in which the third phase collects the retained images from the first and second phases), the number of DNNs and UQ methods can be changed based on different case studies. Moreover, the number of methods in each classification phase is totally optional and can always be adjusted. To cover this point, we first discuss a general view of the proposed dynamic TWDBDL model and then the present detailed models used to analyze each of the two considered real datasets.

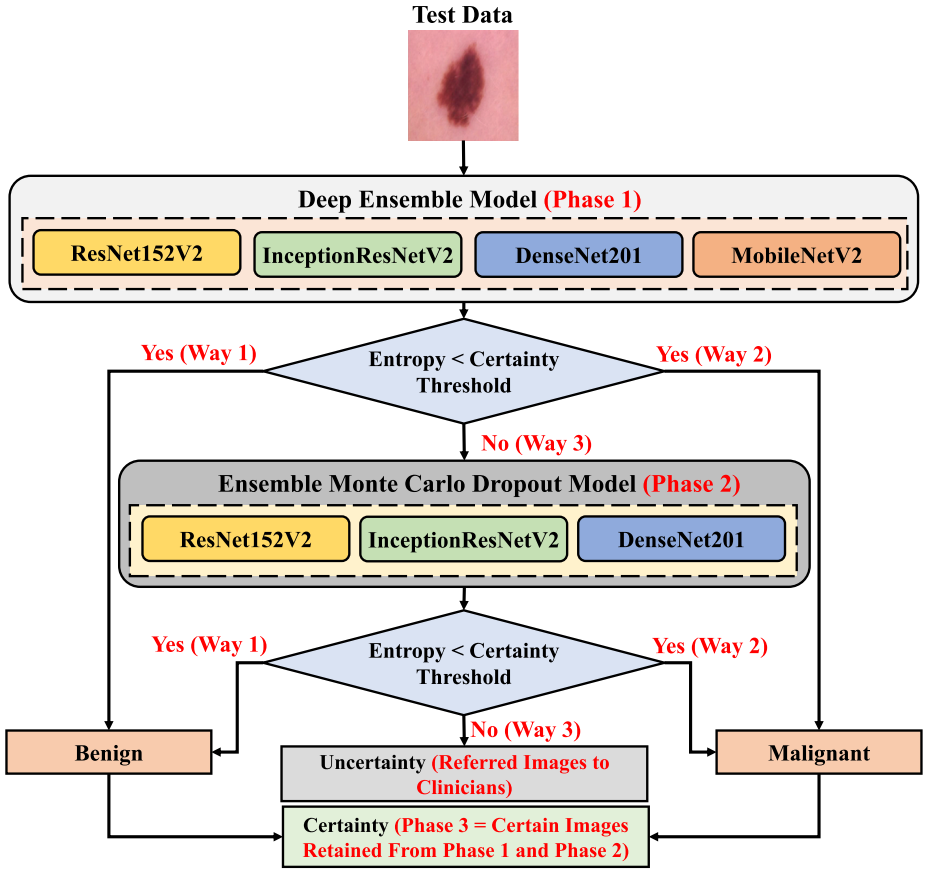


Fig. 2. The block diagram of the applied dynamic TWDBDL model for the first dataset.

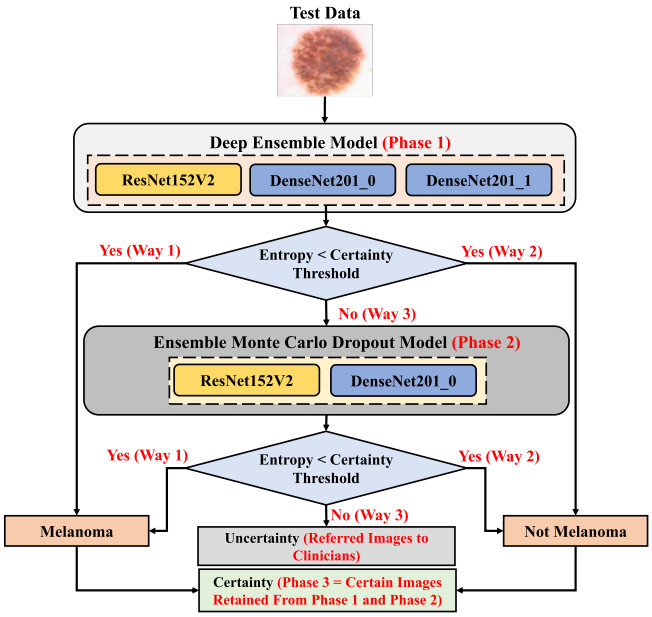


Fig. 3. The block diagram of the applied dynamic TWDBDL model for the second dataset.

3.2. Three-way decision-based uncertainty quantification

We believe that decision-making is the key to achieve the best possible results in any system relying on a"*decision*" process. For this reason, one needs to consider uncertainties in choosing the best options. Moreover, in machine and deep learning, the decision-making process is inevitable. In this area of science, the optimal decision can help achieving the best performance and increase the reliability and trustworthiness of the obtained results. This means that we deal with the model's uncertainties by identifying uncertain points, while conducting different procedures such as classification and segmentation. In 2010, Yao [68] proposed the theory of TWD that is based on the idea of acceptance, rejection and non-commitment. The main notion of the theory can be defined in case of a ternary classification which is related to the evaluation of a set of criteria.

Let U be a finite non-empty set of objects and C be a finite set of conditions. Each member of C may be a criterion. Thresholds on the degrees of satisfiability determine the final decision. The final decisions can be as follows: (a) the object is accepted if it satisfies the set of criteria and if its degree is above a certain level; (b) the object is rejected if it does not satisfy the criteria, meaning that its degree of satisfiability is below a certain level. It is worth noting that an object is considered as a non-commitment if it is neither accepted nor rejected. The last option may also be referred to a deferment decision which needs extra investigation. Whether an object satisfies or does not satisfy the criteria, we cannot determine the subset of objects of the final decisions without considering uncertainty criteria. As a result, we forward uncertain samples to an enhanced model to be further examined in the second phase, as it is illustrated in the block diagram representing our TWDBDL model, Fig. 1.

3.3. The proposed TWD-based UQ dynamic model

As illustrated in Fig. 1, we propose a dynamic multi-phase model in which the TWD-based classification with UQ methods is performed in order to detect uncertain points that will be further examined in the decision-making process of deep learning methods.

In the first phase, presented in Fig. 1, we separately train different DNN models using input images. During training, we use the Bayesian Optimization (BO) to get the best values of hyperparameters (the process

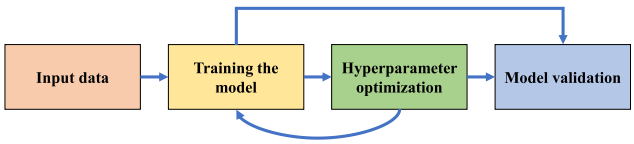


Fig. 4. Bayesian optimization (BO) scheme.

of BO is illustrated in Fig. 4). The BO application helps achieving better results at the test stage. For the first dataset, we used DE as uncertainty quantification method which is applied on the elements of the first phase, *i.e.*, four individual DNN models (ResNet152V2$_1$, DenseNet201$_2$, InceptionResNetV2$_0$ and MobileNetV2$_2$), as shown in Fig. 2. Furthermore, for the second dataset, EMC is applied on the elements of the first phase, ResNet152V2, DenseNet201$_1$ and DenseNet201$_0$, as shown in Fig. 3. Thus, the dynamic ability of the model enables us to change UQ methods and DNNs with respect to the dataset under consideration. Now, we can identify samples for which the DNN models are not certain in their predictions. In other words, the predictions are achieved from M independent trained models. The uncertainty of the probabilistic prediction $\hat{y}$ is computed by estimating its Entropy over its vector of elements $\hat{y}$. For each test sample ($\hat{x}$), the class with the greatest predictive mean is selected as the final output prediction and the Entropy is the measure used for quantifying model uncertainty. Assume that $X = \{x_k | k = 1, 2, \ldots, N\}$ is the considered input dataset, where N denotes the number of samples belonging to two classes (class A and class B), denoted as C_A and C_B, respectively. The input sample, which is passed through each DNN model, is denoted as X_k, where the output of each DNN model is denoted as y_k. In general, the state of every sample can be defined as follows:

$$\begin{cases} \text{DNN is uncertain}, & \text{for } y_k > H_{pred}; \\ X_k \in C_A \text{ or } C_B, & \text{for } y_k < H_{pred}, \end{cases} \tag{8}$$

where H_{pred} illustrates the Etropy that is the threshold used to define the model's uncertainty. It is worth mentioning that the classified samples can be categorized into one of the following categories: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The DNN is certain for some of these samples that can be called retrained samples. However, the other samples for which the DNN is uncertain are considered as referred ones. Referred samples are then forwarded to the second phase.

At the second phase, the proposed model is trained with the same training data that are used at the first phase. Instead of referring uncertain samples to an expert, these samples are fed to the second phase in which a proper UQ method is applied on DNNs to determine whether a sample belongs to a certain class or not, as illustrated in Fig. 1. In other words, dropout is implemented several times at the test time to sample from the approximate posterior and average the stochastic feed forward MC sampling. This gives us an Entropy to estimate the uncertainty. The test samples are then given to the DNN models at the test time. The empirical average of the model's predictions over MC iterations is considered as the estimation of the output for the unseen data, which is given as [17]:

$$\mu_{\text{pred}} \approx \frac{1}{T} \sum_{t=1}^{T} p(\hat{y} | \hat{x}, x, y), \tag{9}$$

where the predictive mean is denoted as $\hat{y}$. Finally, the Entropy estimates provided by each individual DNN model are averaged in case of unknown samples to detect uncertain samples.

It should be noted that the ensemble model (*i.e.*, EMC) applied to process the first dataset consists, at the second phase, of three individual models (ResNet152V2$_2$, DenseNet201$_0$ and InceptionResNetV2$_0$) presented in Fig. 2. However the ensemble model (*i.e.*, EMC) used to process the second dataset includes, at the second phase, two elements:

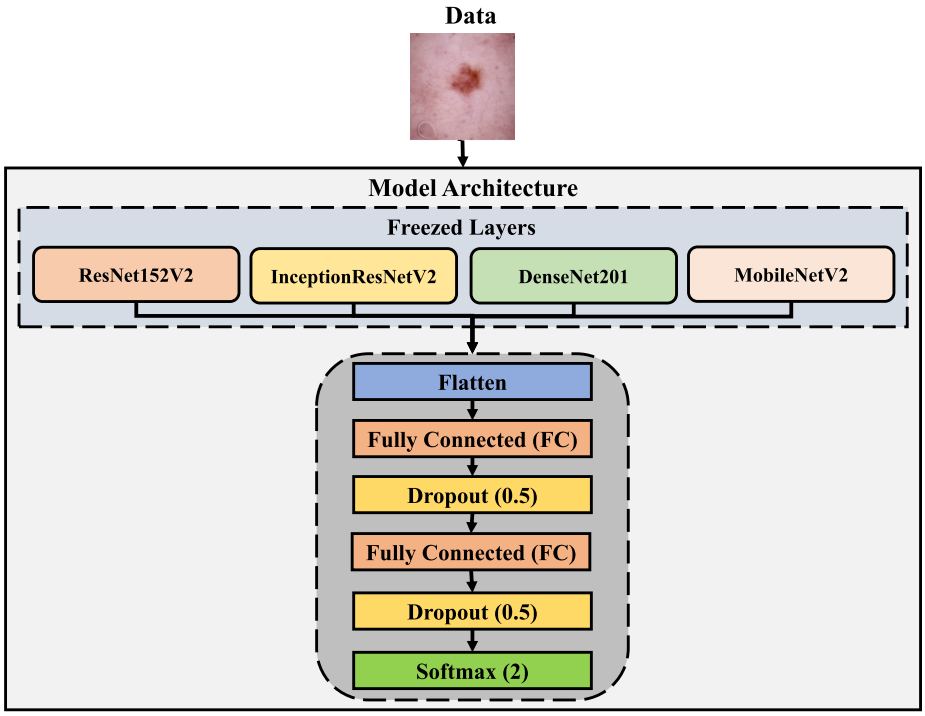


Fig. 5. DNN's block diagram (the backbone of DNN methods).

$ResNet152V2_2$ and $DenseNet201_0$, as illustrated in Fig. 3. As mentioned earlier, at each phase of the model, the UQ methods and DNNs can be changed to make the model more suitable to the data at hand. Then, the referral samples are given to the ensemble model as unseen data at the test time. Finally, as depicted in Fig. 1, the output result of a given model determines whether an unseen sample belongs to class A, class B, or to none of them. If the EMCD model is uncertain about selecting a particular class for a given sample, the model classifies it as uncertain (as indicated: "Uncertainty"). In other words, the model expresses it is as: "*I don't know*" or "*I am not certain*".

3.4. Bayesian Optimization (BO)

Bayesian Optimization (BO) is a probabilistic technique used for the hyperparameter tuning. It is based on the Bayesian theorem. BO has two major components: an acquisition function and a surrogate model. It creates a probabilistic surrogate model, generally by considering a Gaussian process [52,66] or a tree-based model [62]. It sets a Gaussian process prior over the optimization functions to present assumptions about the optimized function. Then, it collects the information from the previous sample to reform the posterior. Then, an acquisition function is chosen to create a utility function from the posterior of the model [52]. Indeed, BO maps several different hyperparameters configurations of their performance with various uncertainty measures [62]. Hyperparameters play an important role in traditional machine learning and deep learning models as they control the training process of the applied models and thus significantly affect their efficiency [67].

3.4.1. Block diagram of the model

In this study, four different well-known deep learning architectures (*i.e.*, ResNet152V2, MobileNetV2, DenseNet201 and InceptionResNetV2) are used as pre-trained deep learning models on ImageNet. The models' weights are frozen during the training stage. However, as it is presented in Fig. 5, two new fully connected layers, followed by a dropout layer with dropout probability of 0.5, are added to each model. As reported in Table 1 and Table 2, the number of neurons in two fully connected layers is separately determined using BO for each individual pre-trained model. $Unites_0$ represents the number of neurons used in the first layer, while $Units_1$ indicates the numbers of neurons used in the second layer. Finally, a two-dimensional output layer is set as the top layer of the model to classify images.

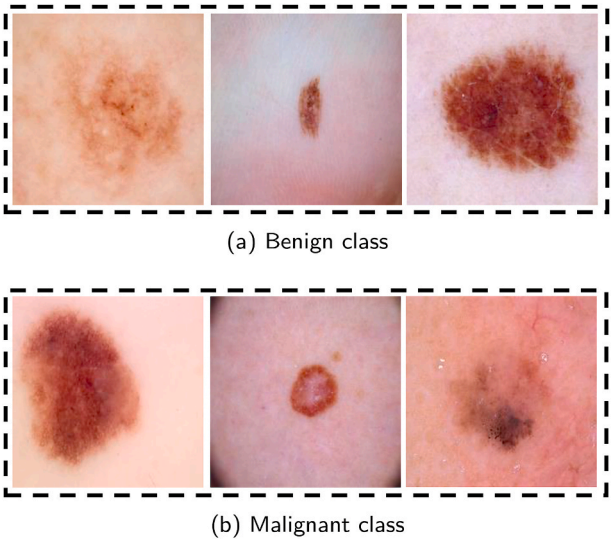


Fig. 6. Some image samples from the first considered skin cancer dataset: (a) Benign and (b) Malignant classes.

4. Experimental validation

In this section, more details of both considered skin cancer datasets, experimental setup, and evaluation metrics, are given. Then, the obtained experimental results are reported.

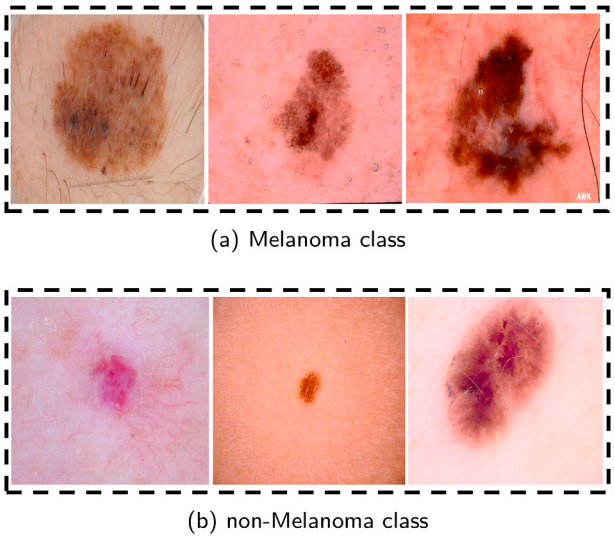


Fig. 7. Some image samples from the second considered skin cancer dataset: (a) Melanoma and (b) non-Melanoma classes.

4.1. Dataset and experimental setup

In the first experiment of this study, we considered a Kaggle Skin Cancer dataset[1] with two classes of images: Benign (class 0) and Malignant (class 1). It contains 2637 training images (1440 Benign and 1197 Malignant) and 660 test images (360 Benign and 300 Malignant, some examples are shown in Fig. 6). Each image is of size 224× 244 pixels.

In the second experiment, we considered the ISIC 2019 dataset is chosen as the second dataset[2] with two classes of images: Melanoma cases (class 0) and non-Melanoma cases (class 1). It contains 7234 training images (3618 Melanoma and 3616 non-Melanoma) and 1808 test images (904 Melanoma and 904 non-Melanoma) with the size of 224 × 244 pixels (some examples are shown in Fig. 7). The images included in our second experimental dataset can be downloaded from this repository: https://drive.google.com/file/d/1XNDq0J86kl9MlRfp8V684YwRyERGrtlB/view?usp=sharing. It should be noted that this dataset is a part of the ISIC 2019 challenge dataset. The non-Melanoma (class 1) class includes images from all classes of the ISIC 2019 challenge dataset, except for those from the Melanoma class.

Here, we provide further details regarding the experimental settings of the proposed model. As mentioned earlier, the MC dropout is applied to quantify uncertainties in which $T = 50$ stochastic forward passes through the DNN model are averaged. In order to achieve faithful results, we repeated each experiment three times and then we reported the mean accuracy and F1-score values. The hyperparameters of each individual model were tuned by applying Bayesian optimization technique. Tables 1 and 2 report the values of the tuned hyperparameters for each applied DNN model, obtained for the first and second datasets, respectively. The default values of the learning rate for the Adam and SGD optimizers were 0.001 and 0.01, respectively. Finally, early stopping technique was also used to prevent the overfitting [42]. To investigate the performance of all applied UQ methods, we calculated the values of the F1-score, Accuracy, Entropy, Standard Deviation (STD) and the receiver operator characteristic (ROC) with its area under the curve (AUC), metrics as it has been done in the recent studies in the field [1,16,47,72].

[1] https://www.kaggle.com/fanconic/skin-cancer-malignant-vs-benign.
[2] https://www.kaggle.com/andrewmvd/isic-2019.

Table 5
Percentage of uncertain samples of the first dataset returned to specialists. In this table, Accuracy of 20% indicates the achieved accuracy when 20% of samples with the highest Entropy were removed and Accuracy of 50% indicates the achieved accuracy when 50% of samples with the highest Entropy were removed from the first dataset.

Method	20%		50%	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
DensNet + MC	89.09	89.08	91.86	95.76
DensNet + EMC	88.18	88.17	92.72	96.22
DensNet + DE	89.09	89.08	90.67	95.11
ResNet152V2 + MC	85.90	85.88	89.69	94.56
ResNet152V2 + EMC	84.23	84.16	90.60	95.07
ResNet152V2 + DE	87.42	87.40	89.91	94.68
MobileNetV2 + MC	87.12	87.10	90.90	93.28
MobileNetV2 + EMC	86.13	86.01	92.12	95.89
MobileNetV2 + DE	86.13	86.01	90.80	95.18
InceptionResNetV2 + MC	85.00	84.96	89.45	94.43
InceptionResNetV2 + EMC	82.29	82.07	89.42	94.41
InceptionResNetV2 + DE	85.45	85.43	87.93	93.57

4.2. Experimental results

In this section, an extensive experimental evaluation of the proposed TWDBDL model is provided separately for the first, second and final classification phases. First, to select the best models for the first and second phases, we took into consideration the performance of each individual DNN model and each individual uncertainty method (*i.e.*, MC, EMC and DE). They are presented in Tables 3 and 4 for the first and second datasets, respectively.

As can be observed from the results obtained for the first dataset (they are listed in Table 3), DE clearly outperforms MC and EMC. Thus, DE was used to combine the best individual DNN models to develop an ensemble model, which yielded reasonable measures for the criteria in the first classification phase of our model as illustrated in Fig. 2. Although DenseNet with the DE technique provided the best accuracy and F1-score, it did not recognize uncertain samples. According to Table 3, since EMC achieved the best AUC for each DNN model, we used it at the second classification phase of the proposed TWDBDL model. Moreover, the loss-epoch and accuracy-epoch curves (DenseNet201, ResNet152V2, MobileNetV2 and InceptionResNetV2 methods) for the first dataset are presented in Fig. 18 in A.1. To assess the performances of the four models (*i.e.*, DenseNet201, ResNet152V2, MobileNetV2 and InceptionResNetV2) using the three uncertainty techniques in terms of the number of referred and retained data of the first dataset, we considered 50% and 20% of all test samples with the highest Entropy as uncertain samples. We then considered the rest of the dataset (50% and 80% of data) as retained data. In Table 5, the applied models for each uncertainty technique, for both 50% and 20% of data, are evaluated separately on the retained data based on the diagnostic accuracy and F1-score, as a function of the uncertainty rate. As reported in Table 5, the performance of all considered techniques has improved with the increase of the referral rate (the rate of the uncertain samples). When we removed 50% of all test samples with the highest Entropy from the first skin cancer dataset, the performance of all models has improved significantly compared to the case where 20% of all test samples with the highest Entropy were removed.

As indicated in Table 4, when the second dataset was considered, EMC outperformed the other UQ methods (*i.e.*, MC and DE). Thus, we combined the best individual DNN models to design an ensemble model based on EMC, giving us reasonable measures for the criteria in both the first and second classification phases of our proposed model, as illustrated in Fig. 3. The loss-epoch and accuracy-epoch curves (for the

Table 6
Percentage of uncertain samples of the second dataset returned to specialists. In this table, Accuracy of 20% indicates the achieved accuracy when 20% of samples with the highest Entropy were removed and Accuracy of 50% indicates the achieved accuracy when 50% of samples with the highest Entropy were removed from the second dataset.

Method	20%		50%	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
DensNet + MC	81.72	81.13	83.78	91.17
DensNet + EMC	78.38	77.63	80.37	83.65
DensNet + DE	71.68	71.66	69.55	73.65
ResNet152V2 + MC	85.80	85.73	86.52	85.88
ResNet152V2 + EMC	77.45	77.44	79.43	84.36
ResNet152V2 + DE	77.98	78.01	76.19	86.48
MobileNetV2 + MC	41.31	41.29	41.40	45.87
MobileNetV2 + EMC	31.69	31.38	29.91	33.49
MobileNetV2 + DE	71.95	72.02	72.35	76.39
InceptionResNetV2 + MC	85.00	84.96	89.45	94.43
InceptionResNetV2 + EMC	34.10	34.09	32.47	41.26
InceptionResNetV2 + DE	74.22	74.24	72.49	76.56

DenseNet201 and ResNet152V2 methods) for the second dataset are presented in Fig. 18 in A.1. Similarly, to the experiments with the first dataset, we considered 50% and 20% of all test samples with the highest Entropy values as uncertain samples to investigate the performances of the four deep learning models (*i.e.*, DenseNet201, ResNet152V2, MobileNetV2 and InceptionResNetV2). We then considered the rest of the dataset (50% and 80% of the data) as retained data. In Table 6, the applied models in terms of each uncertainty technique for both 50% and 20% are evaluated separately on the retained data based on the diagnostic accuracy and F1-score, as a function of the uncertainty rate. As reported in Table 6, the performance of all considered techniques have improved in terms of F1-score with the increase in the referral rate of the second dataset (the rate of the uncertain samples). When we removed 50% of all test samples with the highest Entropy, the performance of all models improved significantly in terms of F1-score compared to the case where 20% of all test samples with the highest Entropy were removed. Although the accuracy rate of 7 UQ methods has increased when 50% of all test samples with the highest Entropy were removed from the data, compared to the case when 20% of them were removed, it was not the case of the remaining 5 UQ methods, including - DensNet using DE, ResNet152V2 using DE, MobileNetV2 using EMC, InceptionResNetV2 using EMC, and InceptionResNetV2 using DE.

In addition, we have visualized the regions of the input samples that were the most important for predictions of the applied deep learning methods. The Gradient-weighted Class Activation Mapping (Grad-CAM) was used here for the first (see Fig. 8) and second (see Fig. 9) datasets.

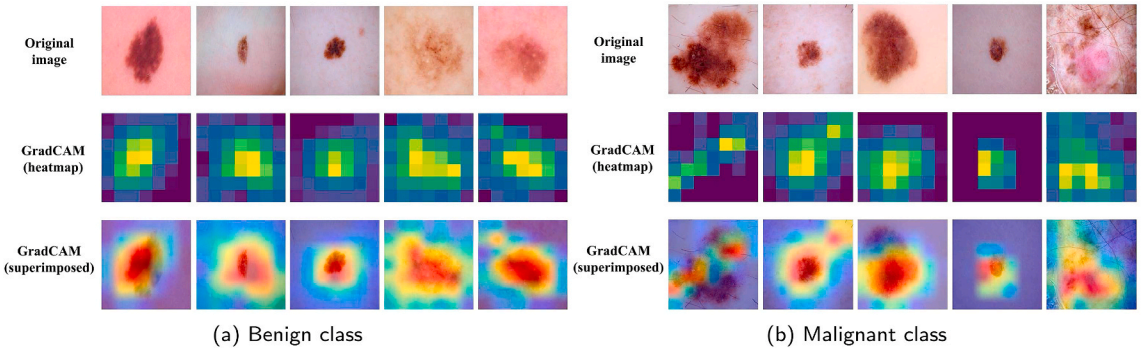


Fig. 8. Grad-CAM visualizations shown for ten image samples (five images for each class from the first dataset).

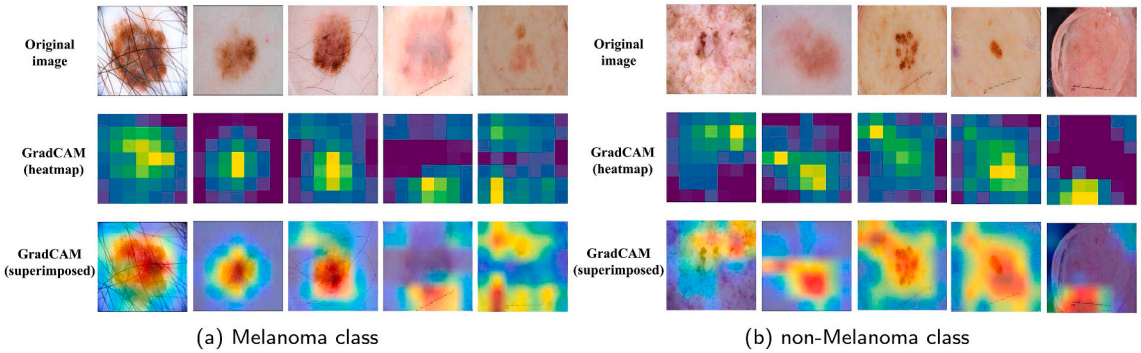


Fig. 9. Grad-CAM visualization are indicated for ten images (five images for each class of the second dataset).

Table 7
The results obtained in the first classification phase of the proposed TWDBDL model applied to two considered skin cancer datasets. In this table, f1-0 is the F1-score of class 0, f1-1 is the F1-score of class 1 and F1-score is the overall F1-score.

Datasets	Method	EC	EI	STD-Co	STD-In	f1-0	f1-1	Accuracy (%)	F1-score (%)	AUC
First	TWDBDL (DE)	0.6207	0.6335	0.1351	0.2377	91.00	89.00	87.55	90.19	0.9377
Second	TWDBDL (EMC)	2.624	2.835	0.4707	0.4934	91.00	92.00	89.39	92.00	0.9700

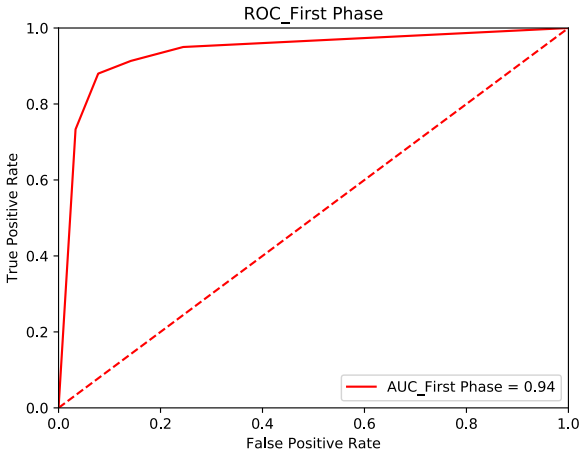


Fig. 10. The ROC curve of the proposed TWDBDL model constructed for its first phase when applied to the first considered skin cancer dataset.

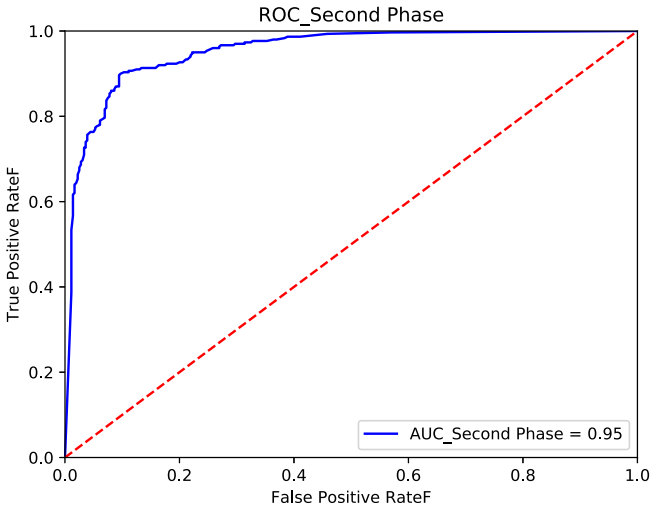


Fig. 12. The ROC curve of the proposed TWDBDL model constructed for its second phase when applied to the first considered skin cancer dataset.

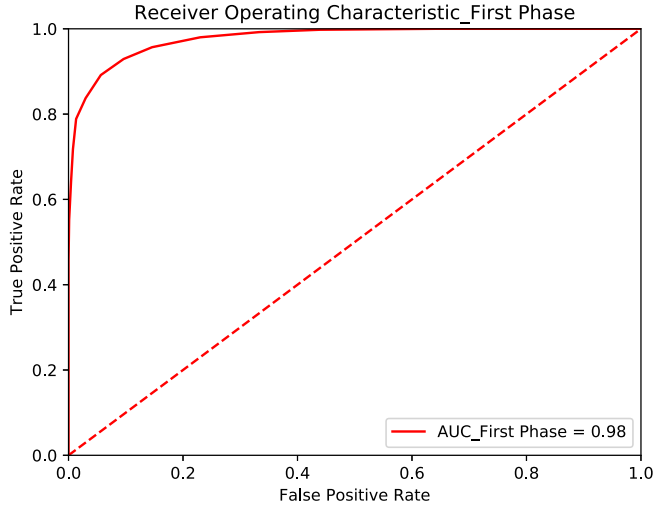


Fig. 11. The ROC curve of the proposed TWDBDL model constructed for its first phase when applied to the second considered skin cancer dataset.

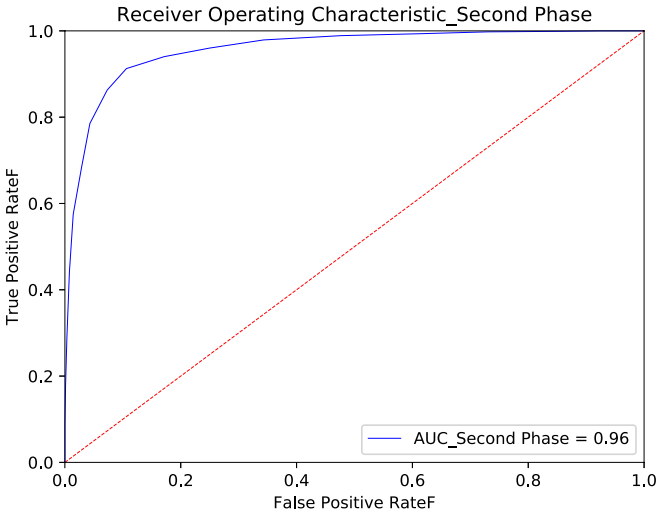


Fig. 13. The ROC curve of the proposed TWDBDL model constructed for its second phase when applied to the second considered skin cancer dataset.

Thus, we have visualized the heatmaps and superimposed five images for each class. In this study, we have detected uncertain images in the first classification phase and then referred them to the second classification phase for further evaluation instead of removing them. In order to be more confident in the predicted results, we have presented a two-phase model for assessing uncertain samples.

4.2.1. *The first phase results*

The results of the first classification phase provided by our TWDBDL model applied to the first considered skin cancer dataset are presented in the first row of Table 7, including the assessment of the performance of the DE method comprising the ResNet152V2, DenseNet201, Inception-ResNetV2 and MobileNetV2 architectures. The 232 uncertain samples were referred to the second phase. It should be noted that only the retained samples were considered to estimate the accuracy, F1-score and STD of the first phase. The ROC curve for the first classification phase is shown in Fig. 10.

The results of the first phase of the second dataset are presented in Table 7, including the assessment of performance of EMC method containing ResNet152V2, DenseNet201 architectures. The 375 uncertain samples are referred to the second phase. It should be noted that only retained samples are considered to estimate the accuracy, F1-score and STD of the first phase. The ROC of the first phase is shown in Fig. 11.

4.2.2. *The second phase results*

In this section, we present the performance of the applied ensemble model integrating the DenseNet201,ResNet152V2 and InceptionResNetV2 architectures using EMCD in Table 8, on quantifying the uncertainty of the referred samples of the first dataset. It is important to note

Table 8
The results obtained in the second classification phase of the proposed TWDBDL model applied to two considered skin cancer datasets.

Datasets	Method	EC	EI	STD-Co	STD-In	f1-0	f1-1	Accuracy (%)	F1-score (%)	AUC
First	TWDBDL (EMC)	2.860	3.995	0.2677	0.4724	91.00	89.00	93.04	90.00	0.95
Second	TWDBDL (EMC)	2.78191	2.708	0.43290	0.46577	90.00	90.00	99.61	90.00	0.96

Table 9
The results obtained in the final phase by the proposed TWDBDL model for two considered skin cancer datasets.

Datasets	Method	f1-0	f1-1	Accuracy (%)	F1-score (%)	AUC
First	TWDBDL	92.00	83.00	88.95	89.00	0.92
Second	TWDBDL	91.00	91.00	90.96	91.00	0.97

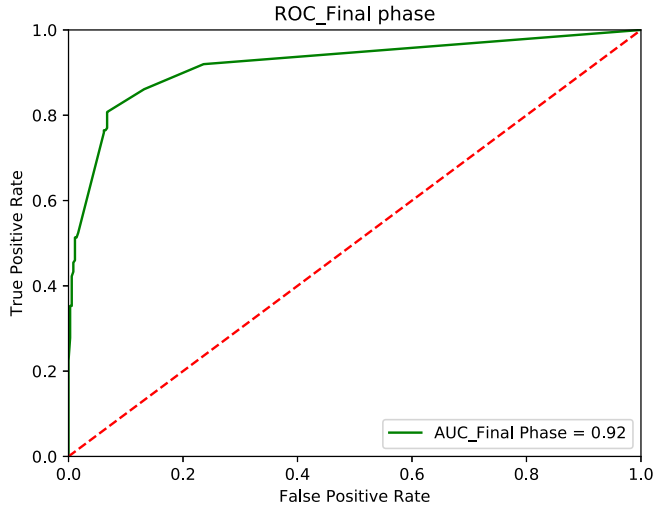


Fig. 14. The ROC curve of the proposed TWDBDL model constructed for its when final phase when applied to the first considered skin cancer dataset.

that only the samples for which the given model was certain have been used in the second phase. As shown in Tables 7 and 8, the accuracy and the AUC values in the second phase were better than in the first phase for the first skin cancer dataset. The ROC curve of the second phase for the first considered skin cancer dataset is illustrated in Fig. 12. Overall, the model was uncertain for 117 test samples out of 232 considered in the second phase. It is worth noting that most of these 117 samples belonged to the class 1 (Malignant cases), while most of the retained samples (115 images) belonged to class 0 (Benign cases). The obtained results suggest that the proposed TWDBDL model is robust and reliable for the uncertain samples as it does not tend to make overconfident predictions.

The performance of the model including the ensemble of the ResNet152V2$_2$ and DenseNet2010 architectures using EMC, applied to the second considered skin cancer dataset, is provided in the second row of Table 8. As shown in Table 8 the accuracy and AUC values in the second phase are higher compared to those of the first phase. The ROC curve of the second phase for the second considered dataset is illustrated in Fig. 13. Here, the model was uncertain for 115 test samples out of 375 considered in the second phase. These uncertain samples in the final phase should be transferred to clinicians. Based on these outcomes, we achieved our goal of constructing a more trustable model that says *I don't know* when it is not confident (certain) regarding its predictions.

4.2.3. The final phase results

In this classification phase, we sum up certain samples of the first and second classification phases to evaluate the overall accuracy, F1-score and AUC of the proposed model. As reported in Table 9, for the first considered skin cancer dataset, the final accuracy is higher than the first phase accuracy, while the values of F1-score and of AUC of the final phase are lower than those of the first phase. The ROC curve of the final phase for the first considered dataset is shown in Fig. 14. As indicated in Tables 7 and 8, F1-score of class 0 in the final phase is higher than those obtained in the first and second phases, whereas F1-score of class 1 in the final phase is lower than those in the first and second phases. As mentioned in section 4.2.2, most of the referred samples removed from

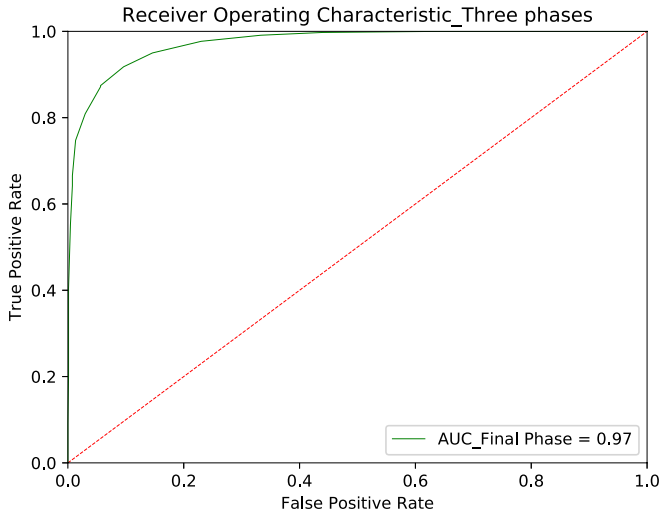


Fig. 15. The ROC curve of the proposed TWDBDL model constructed for its when final phase when applied to the second considered skin cancer dataset.

the dataset belong to class 1. They are not considered when it comes to estimating of the accuracy, F1-score and AUC. Some of these uncertain samples, referred to radiologists for further investigation, are recognized correctly by the ensemble model in both first and second phases. As a result, in the final phase, we have a slight decrease in F1-score of class 1 (Malignant class), total F1-score and AUC. This point will be discussed in more detail in Section 5 to show the importance and advantages of the proposed model.

As indicated in Table 9, for the second considered skin cancer dataset, the final accuracy is higher than the first phase accuracy and lower than the second phase. The value of F1-score in the final phase is lower than in the first phase, but greater than in the second phase. The ROC curve of the final phase for the second considered skin cancer dataset is illustrated in Fig. 15. As reported in Tables 7 and 8, the F1-score of classes 0 and 1 in the final phase are higher than in the second phase. F1-score of class 0 in the final phase remains the same as in the first phase, though it decreases compared to that in the second phase.

5. Discussion

5.1. Evaluation of the proposed model

In this section, we highlight the advantages of the proposed UQ model discussing its performances on the two considered skin cancer datasets. In Table 10,the recall and precision values per class for the two datasets are reported.

5.1.1. First dataset

As indicated in Table 10, the precision of class 0 (Benign cases) in the final classification phase is greater than in the first and second classification phases. However, the precision of the final phase of class 1 (Malignant cases) is lower than in the first and second phases. The recall value of class 0 for the first phase is the same as for the final phase, while it grows in the second phase compared to the first phase. However, there is a decreasing trend in the recall value of class 1 in the second and the final phases compared to the first phase. As a result, the F1-score of class 0 is greater in the final phase compared to the first two phases; however, the F1 value of class 1 in the final phase dropped compared to the first and second phases. Although the overall F1-score and AUC decreased in the final phase compared to the first and second phases, the value of the F1-score of class 0 and the accuracy increased in the final phase. It should be mentioned that the main goal of this study was not to present a state-of-the-art model, but assess the performance of TWD combined with uncertainty methods to improve the diagnosis performance of

Table 10
Precision, recall, F1-score, Accuracy and AUC results obtained for the three phases of the proposed TWDBDL model for the two skin cancer datasets considered in this study.

Dataset	Phases	P-0	P-1	R-0	R-1	F1-0	F1-1	F1-T	Accuracy	AUC
Dataset 1	1	92.00	88.00	90.00	90.00	91.00	89.00	90.00	87.85	0.9377
	2	90.00	90.00	92.00	89.00	91.00	89.00	90.00	93.04	0.9500
	3	93.00	81.00	90.00	86.00	92.00	83.00	89.00	88.95	0.9200
Dataset 2	1	89.00	94.00	94.00	89.00	91.00	92.00	92.00	89.39	0.9700
	2	89.00	91.00	91.00	90.00	90.00	90.00	90.00	99.61	0.9600
	3	89.00	93.00	94.00	88.00	91.00	91.00	91.00	90.96	0.9700

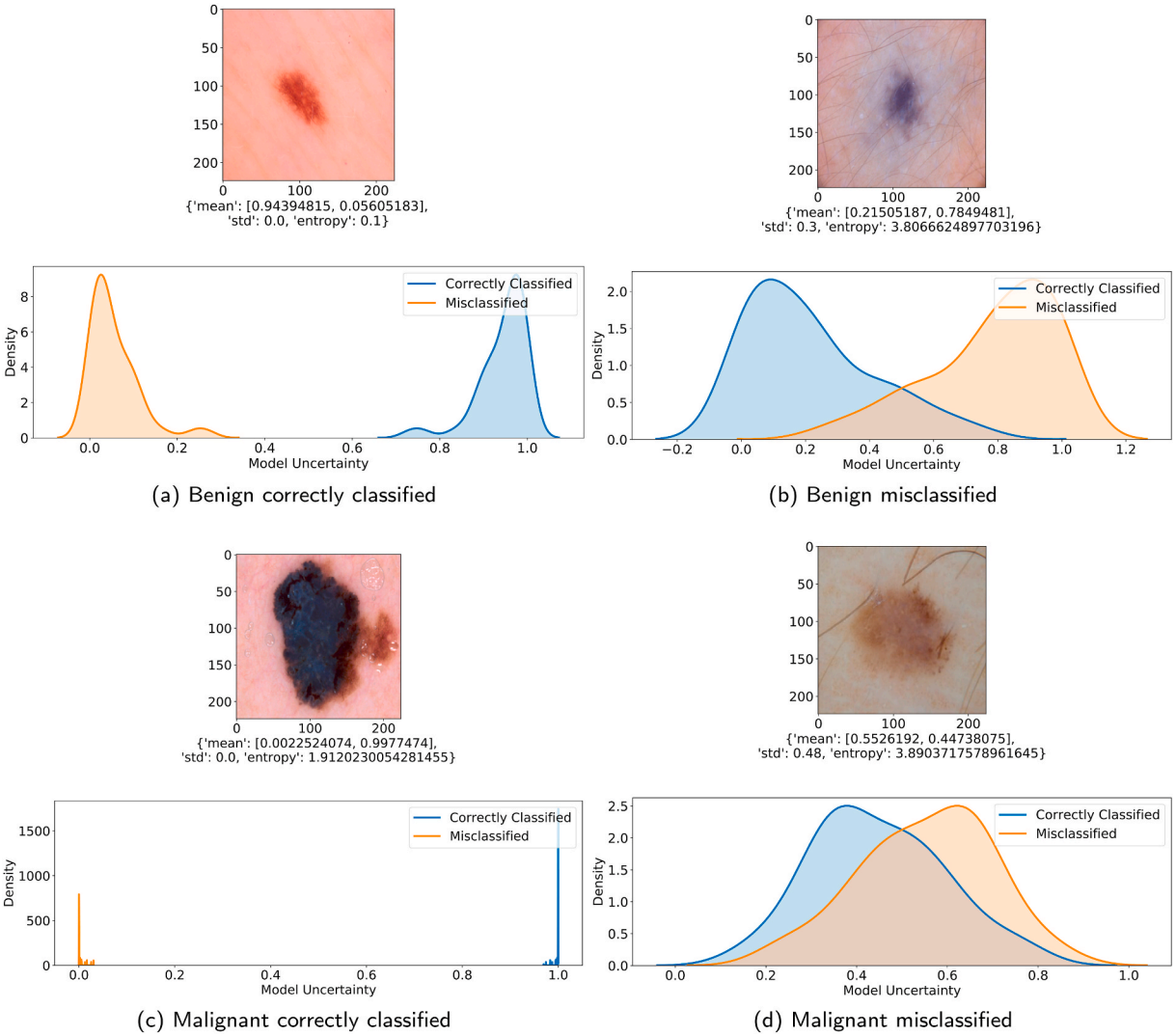


Fig. 16. Four test input images with the corresponding STD and Entropy values and the corresponding predictive distributions generated by BDLs (taken from the first considered skin cancer dataset).

DNNs and BDLs. It is clear that the proposed model cannot improve the diagnostic in case of all criteria for both classes, though we managed to improve the diagnostic performance of our model for class 0 (Benign) as well as the overall accuracy. It is critically important to differentiate Benign patients from Malignant ones as the cost of a mistake could be fatal.

5.1.2. Second dataset

As reported in Table 10, the precision of class 0 (Melanoma cases) in the final phase remains the same during the first, second and final classification phases. Although in the second phase the precision of class 1 (non-Melanoma cases) decreased compared to that of the first phase, the value of the final phase is greater than that in the second phase. When it comes to the value of the recall associated with the class 0, it remains stable in the final phase compared to the first phase, being higher than the recall value of the second phase. However, the recall value of class 1 in the final phase is lower than those of the first and second phases. Thereby, the F1-score value of class 0 in the final phase remains stable compared to the first phase, while it decreases in the second phase. Regarding the F1-score value of class 1, it decreases in the final phase compared to the first phase; however, the value of the final phase is higher than that of the second phase. The overall F1-score decreases in the final phase compared to the first one. Thus, we can see that the overall accuracy increases compared to that of the first phase, but it

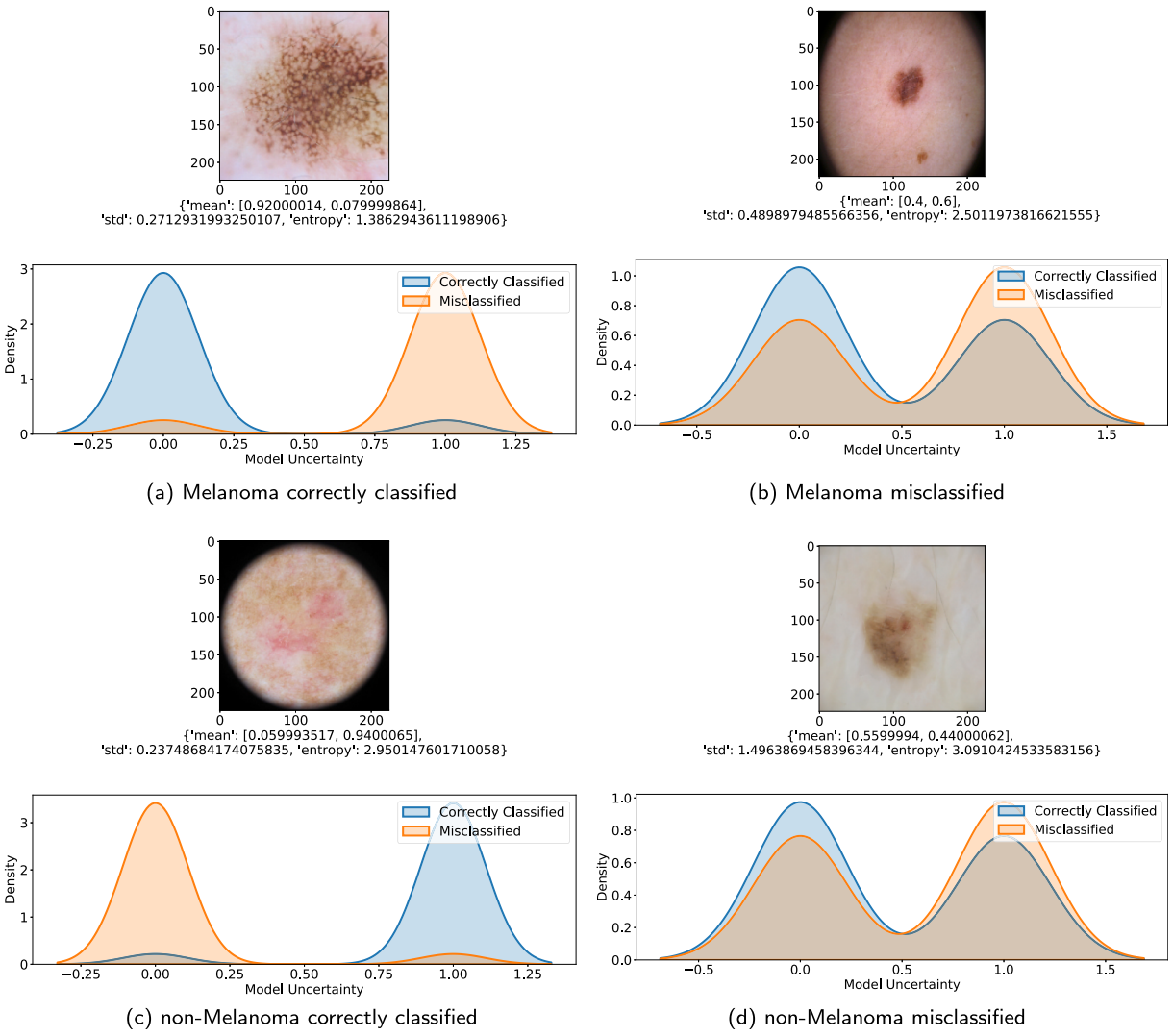


Fig. 17. Four test input images with the corresponding STD and Entropy value and the corresponding predictive distributions generated by BDLs (taken from the second considered skin cancer dataset).

declines compared to the second phase. Finally, the AUC value of the final phase is the same as in the first phase and it drops in the second phase.

As mentioned earlier, our objective was not only to improve the overall prediction performance, we also aim at defining a model having more confidence in its predictions. Clearly, the proposed model is able to deal with uncertainties in its predictions. This is vital as it helps to distinguish the patients having skin cancer from the healthy ones.

5.2. Relationship between entropy, STD and misclassification

5.2.1. First dataset

Four test input images from the first dataset with the STD and Entropy values and the corresponding predictive distributions generated by BDLs are presented in Fig. 16. According to Fig. 16, the predictive uncertainty is greater for misclassified samples compared to the correctly classified samples.

Fig. 16 shows the distribution of predictive uncertainty values for four selected test images that are classified correctly (in blue) and wrongly (in red). The class with the highest value of the Softmax output for the predictive distribution mean is considered as the final output and the predictive Entropy of the estimated output distribution is considered as the estimated epistemic uncertainty. A model can be uncertain with high or low Entropy, which is associated to the predictive posterior distribution. Wider output posterior distributions reflect lower confidence of the model. Fig. 16 (a) presents an image that is correctly classified in the Benign class. It is clear that in this case the model is highly certain about its prediction (STD = 0 and Entropy = 0.1). Fig. 16 (b) illustrates a Benign sample that is misclassified as Malignant. The high values of STD and Entropy (STD = 0.3 and Entropy = 3.8066) and a wider posterior distributions prove that the model is uncertain. Fig. 16 (c) presents an image that is classified correctly as a Malignant sample. The model is certain about this case (STD = 0.0 and Entropy = 1.9120). Fig. 16 (d) presents a misclassified Malignant sample that was wrongly diagnosed as Benign. The uncertainty of the model is expressed through its high Entropy (STD = 0.48 and Entropy = 3.8903) and wide posterior distributions.

5.2.2. Second dataset

Furthermore, the STD and Entropy values with the corresponding predictive distributions generated by BDLs for four test samples from the second dataset are illustrated in Fig. 17. Once again, here, the estimated uncertainty is higher for misclassified samples than for correctly classified ones. Fig. 17 shows the distribution of predictive uncertainty for four selected random test images from the second dataset. The misclassified samples are shown in red and correctly classified samples in

Table 11
Comparison of the performance of the proposed TWDBDL model with some existing ML models used to classify skin cancer data.

Study	Year	# of Classes	Model	Accuracy	F1-score	AUC	Uncertainty
Esteva et al. [15]	2017	3	CNN	72.10	N/A	N/A	No
Esteva et al. [15]	2017	9	CNN	55.40	N/A	N/A	No
Mobiny et al. [36]	2019	7	Bayesian DenseNet-169	83.59	N/A	N/A	Yes
Bologna and Fossati [9]	2020	2	CNN + VDIMLP[a]	84.90	N/A	N/A	No
Combalia et al. [12]	2020	9	TA + MCD[b]	N/A	N/A	N/A	Yes
Lee and Renee [31]	2020	2	CNN	82.90	N/A	N/A	No
Pacheco and Krohling [40]	2021	9	ResNet-50	91.30	N/A	0.865	No
Samrat and Ganguly [38]	2021	2	CNN	60.00	69.04	N/A	No
Bhardwaj and Rege [8]	2021	2	SVM[c]	86.00	68.57	N/A	No
Khan et al. [25]	2021	7	CNN	86.50	86.28	N/A	No
Wang et al. [59]	2021	7	STCN[d]	80.60	N/A	0.790	No
Our study (Dataset 1)	2021	2	TWDBDL	88.95	89.00	0.920	Yes
Our study (Dataset 2)	2021	2	TWDBDL	90.96	91.00	0.970	Yes

[a] VDIMLP = Virtual Discretized Interpretable a Multi-Layer Perceptron.
[b] TA + MCD = Test Augmentation þ MC dropout.
[c] SVM = Support Vector Machine.
[d] STCN = Self-supervised Topology Clustering Network.

blue. As mentioned earlier, the predictive Entropy of the estimated distributions is considered as estimated epistemic uncertainty, whereas a wider output posterior distribution reflects a low confidence of the model. Fig. 17 (a) presents an image that is correctly classified in the Melanoma class. It shows that the model is certain about its prediction (STD = 0.2712 and Entropy = 1.3862). Fig. 17 (b) illustrates a Melanoma sample that is misclassified as a non-Melanoma image. The high value of STD and Entropy (STD = 0.4898 and Entropy = 2.5011) with wide posterior distributions indicate that the model is uncertain in its prediction. Fig. 17 (c) presents a test image that is classified correctly as a non-Melanoma sample. Here, the model is certain in its prediction with low Entropy (STD = 0.2374 and Entropy = 2.9501). Fig. 17 (d) presents a misclassified image of non-Melanoma, wrongly diagnosed as Melanoma. The high value of Entropy (STD = 1.4963 and Entropy = 3.0910) and wide posterior distributions suggest that the model is not certain about its outcome here.

5.3. Comparison with existing prediction models

In order to highlight the results provided by the proposed TWDBDL model, we compare its performance with those of the existing efficient models used to classify different skin cancer datasets (see Table 11). It the addition to the models' performances, we are also interested in the uncertainty of classification results.

Table 11 reports the results provided by the proposed TWDBDL model in its final phase for both skin cancer datasets. According to these results, the TWDBDL model is highly competitive compared to its counterparts, being able to estimate the uncertainty accurately (for both considered datasets). Moreover, even though Rasul et al. [43] obtained slightly better results for the second dataset, they did not use the second dataset alone to design and train their model. Rasul et al. first used the ISIC 2018 dataset for segmentation and the second dataset for classification.

By observing the results presented in Table 11, we can notice that most of the existing ML models dealing with skin cancer classification (*e.g.*, Yu et al. [69], Roslin [44], Zhang et al. [70], Pathan et al. [41], Tang et al. [58], Zhuang et al. [71] and Tan et al. [57]) do not consider the uncertainty of the model's output. A few studies considering the model's uncertainty include those of Mobiny et al. [36] and Combalia et al. [12]. Mobiny et al. [36] tested their model on the HAM10000 dataset (seven classes) and achieved 90.00% accuracy by referring 35% of the samples (approximately 1 out of 3 images) to physicians. The TWDBDL model introduced here refers to physicians and clinicians much fewer samples. Another important advantage of our model is that in each classification phase, we use multiple models instead of a single one. In other words, our model has a set of decision processes instead of a single decision process. Combalia et al. [12] developed their model for the ISIC 2018 and 2019 challenges, achieving balanced accuracies of 76.00% and 64.00%, respectively.

Even though our model has several important advantages, it still has a great potential for improvement which need to be addressed in the future. For instance, the performance of the proposed TWDBDL model can be improved using an attention-based mechanism [6]. Moreover, some other recent UQ methods, such as MC-DropConnect [37], can be integrated in it to get a better quantification output. The weight of each individual classifier used at each step of the presented model can be considered as well [28]. Finally, both the robustness of the TWDBDL model against noise (noise classifications [20,47]) and the impact of various fusion models [5,39] can be studied.

6. Conclusion

In this study, we introduced a new, simple, but yet very efficient uncertainty quantification model based on the Three-Way Decision (TWD) theory and applied it to analyze two well-known skin cancer image datasets. The main goal of our work was not to introduce a new state-of-art deep learning model, but assess the performance of uncertainty quantification models using both Bayesian CNNs and TWD in order to improve the performance of computer-aided diagnostic systems. Our novel hybrid dynamic TWDBDL model allowed us to apply different UQ methods and different DNNs in distinct classification phases. Thus, we were able to select the most appropriate elements of the model in each phase, adjusting them to the data at hand. In our study, two UQ methods were used in two classification phases, preventing one from making overconfident skin cancer classification decisions. DE and EMC were applied in the first and second classification phases for the first considered skin cancer dataset. However, only EMC was applied in both the first and second classification phases for the second considered skin cancer dataset. The accuracy, F1-score and AUC of the final phase of the model were, respectively, 88.95%, 89.00% and 0.92 for the first considered dataset, and, respectively, 90.96%, 91.00% and 0.97 for the second considered dataset. These results are very encouraging. One of the advantages of our model is an automated differentiation between two classes of benign and malignant melanoma cases as well as non-melanoma cases (since the cost of misdiagnosis can be fatal). To further deal with uncertainty, a Bayesian optimization method was employed to tune the hyperparameters of all deep learning architectures used in our work. The proposed TWDBDL model can be effectively incorporated into various computer-aided diagnostic systems which certainly need to integrate new tools for estimating uncertainty of the models' predictions.

In the future, we plan to develop a weighted ensemble model based

on the TWD theory to enhance the uncertainty awareness of different deep learning models. Furthermore, non-probabilistic decision theory, called Info-Gap decision theory, can also be considered to optimize the robustness of failure. Moreover, the proposed model should be tested for noise detection. Finally, we plan to provide a confidence score for the TWD theory procedure outputs to make the model's decisions more effective.

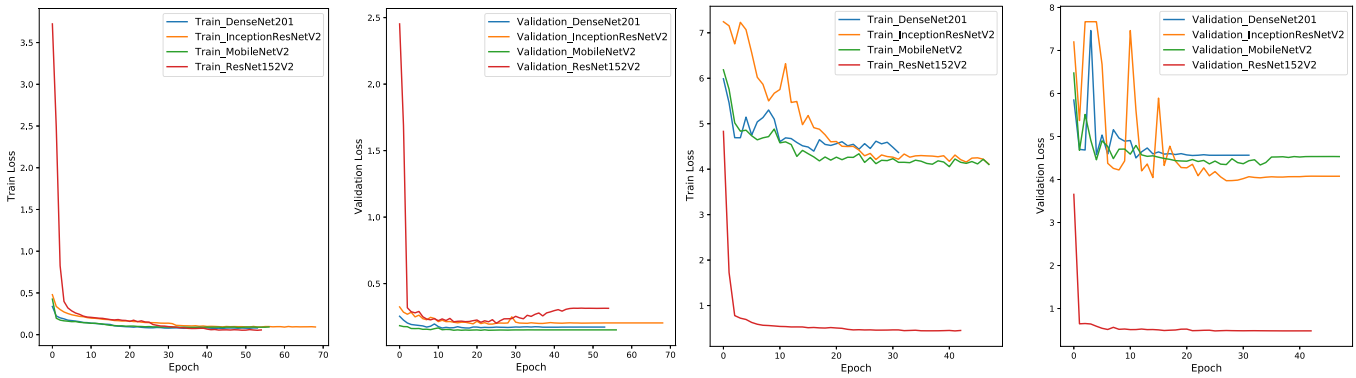
Acknowledgment

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

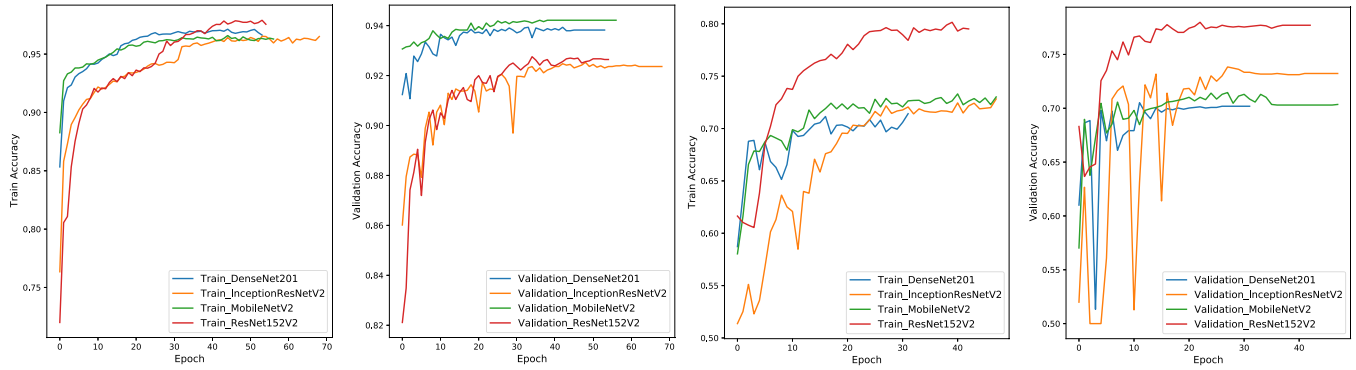
Appendix

A.1. Accuracy-Epoch vs Loss-Epoch curves

The loss-epoch and accuracy-epoch curves (DenseNet201, ResNet152V2, MobileNetV2, and InceptionResNetV2 methods) for the first and second datasets are presented in Fig. 18.



(a) Loss-Epoch for the first dataset (b) Loss-Epoch for the second dataset



(c) Accuracy-Epoch for the first dataset (d) Accuracy-Epoch for the second dataset

Fig. 18. Accuracy-Epoch vs Loss-Epoch curves for the first and second datasets.

References

[1] M. Abdar, U.R. Acharya, N. Sarrafzadegan, V. Makarenkov, Ne-nu-svc: a new nested ensemble clinical decision support system for effective diagnosis of coronary artery disease, IEEE Access 7 (2019) 167605–167620.

[2] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, et al., A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges, 2020 arXiv preprint arXiv:2011.06225.

[3] G. Alain, Y. Bengio, Understanding intermediate layers using linear classifier probes, in: International Conference on Learning Representations Workshop, 2017, pp. 1–4.

[4] Z. Alyafeai, L. Ghouti, A fully-automated deep learning pipeline for cervical cancer classification, Expert Syst. Appl. 141 (2020), 112951.

[5] M.E. Basiri, M. Abdar, M.A. Cifci, S. Nemati, U.R. Acharya, A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques, Knowl. Base Syst. 198 (2020), 105949.

[6] M.E. Basiri, S. Nemati, M. Abdar, E. Cambria, U.R. Acharya, Abcdm: an attention-based bidirectional cnn-rnn deep model for sentiment analysis, Future Generat. Comput. Syst. 115 (2021) 279–294.

[7] L. Bertoni, S. Kreiss, A. Alahi, Monoloco: Monocular 3d pedestrian localization and uncertainty estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6861–6871.

[8] A. Bhardwaj, P.P. Rege, Skin lesion classification using deep learning, in: Advances in Signal and Data Processing, Springer, 2021, pp. 575–589.

[9] G. Bologna, S. Fossati, A two-step rule-extraction technique for a cnn, Electronics 9 (2020) 990.

[10] G. Carneiro, L.Z.C.T. Pu, R. Singh, A. Burt, Deep Learning Uncertainty and Confidence Calibration for the Five-Class Polyp Classification from Colonoscopy, Medical Image Analysis, 2020, p. 101653.

[11] J.P. Cohen, M. Luck, S. Honari, Distribution matching losses can hallucinate features in medical image translation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 529–536.

[12] M. Combalia, F. Hueto, S. Puig, J. Malvehy, V. Vilaplana, Uncertainty estimation in deep neural networks for dermoscopic image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 744–745.

[13] N. Coudray, P.S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, A. Tsirigos, Classification and mutation prediction from

non–small cell lung cancer histopathology images using deep learning, Nat. Med. 24 (2018) 1559–1567.
[14] S. Ebrahimi, M. Elhoseiny, T. Darrell, M. Rohrbach, Uncertainty-guided continual learning with bayesian neural networks. International Conference on Learning Representations, 2019, pp. 1–16.
[15] 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 (2017) 115–118.
[16] A. Filos, S. Farquhar, A.N. Gomez, T.G. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. de Kroon, Y. Gal, A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks, in: NIPS Bayesian Deep Learning Workshop, 2019, pp. 1–12.
[17] Y. Gal, Uncertainty in Deep Learning, vol. 1, University of Cambridge, 2016.
[18] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: representing model uncertainty in deep learning, in: International Conference on Machine Learning, 2016, pp. 1050–1059.
[19] Y. Gal, Z. Ghahramani, A theoretically grounded application of dropout in recurrent neural networks, in: Advances in Neural Information Processing Systems, 2016, pp. 1019–1027.
[20] B. Gao, H. Gouk, T.M. Hospedales, Searching for Robustness: Loss Learning for Noisy Classification Tasks, 2021 arXiv preprint arXiv:2103.00243.
[21] A. Ghoneim, G. Muhammad, M.S. Hossain, Cervical cancer classification using convolutional neural networks and extreme learning machines, Future Generat. Comput. Syst. 102 (2020) 643–649.
[22] P. Henderson, T. Doan, R. Islam, D. Meger, Bayesian policy gradients via alpha divergence dropout inference, in: NIPS Bayesian Deep Learning Workshop, 2017, pp. 1–12.
[23] G. Kahn, A. Villaflor, V. Pong, P. Abbeel, S. Levine, Uncertainty-aware Reinforcement Learning for Collision Avoidance, 2017 arXiv preprint arXiv: 1702.01182.
[24] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in Neural Information Processing Systems, 2017, pp. 5574–5584.
[25] M.A. Khan, Y.D. Zhang, M. Sharif, T. Akram, Pixels to classes: intelligent learning framework for multiclass skin lesion localization and classification, Comput. Electr. Eng. 90 (2021), 106956.
[26] A. Kirsch, J. van Amersfoort, Y. Gal, Batchbald: efficient and diverse batch acquisition for deep bayesian active learning, in: Advances in Neural Information Processing Systems, 2019, pp. 7026–7037.
[27] I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, Artif. Intell. Med. 23 (2001) 89–109.
[28] A. Koohestani, M. Abdar, A. Khosravi, S. Nahavandi, M. Koohestani, Integration of ensemble and evolutionary machine learning algorithms for monitoring diver behavior using physiological signals, IEEE Access 7 (2019) 98971–98992.
[29] A. Labach, H. Salehinejad, S. Valaee, Survey of Dropout Methods for Deep Neural Networks, 2019 arXiv preprint arXiv:1904.13310.
[30] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Information Processing Systems, 2017, pp. 6402–6413.
[31] K.W. Lee, R.K.Y. Chin, The effectiveness of data augmentation for melanoma skin cancer prediction using convolutional neural networks, in: 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), IEEE, 2020, pp. 1–6.
[32] C. Leibig, V. Allken, M.S. Ayhan, P. Berens, S. Wahl, Leveraging uncertainty information from deep neural networks for disease detection, Sci. Rep. 7 (2017) 1–14.
[33] G. Luo, S. Dong, W. Wang, K. Wang, S. Cao, C. Tam, H. Zhang, J. Howey, P. Ohorodnyk, S. Li, Commensal correlation network between segmentation and direct area estimation for bi-ventricle quantification, Med. Image Anal. 59 (2020) 101591.
[34] W.J. Maddox, P. Izmailov, T. Garipov, D.P. Vetrov, A.G. Wilson, A simple baseline for bayesian uncertainty in deep learning, in: Advances in Neural Information Processing Systems, 2019, pp. 13153–13164.
[35] B. Mazoure, R. Islam, Alpha-divergences in Variational Dropout, 2017 arXiv preprint arXiv:1711.04345.
[36] A. Mobiny, A. Singh, H. Van Nguyen, Risk-aware machine learning classifier for skin lesion diagnosis, J. Clin. Med. 8 (2019) 1241.
[37] A. Mobiny, P. Yuan, S.K. Moulik, N. Garg, C.C. Wu, H. Van Nguyen, Dropconnect is effective in modeling uncertainty of bayesian deep networks, Sci. Rep. 11 (2021) 1–14.
[38] S. Mukherjee, D. Ganguly, Transfer learning in skin lesion classification, in: Proceedings of International Conference on Frontiers in Computing and Systems, Springer, 2021, pp. 343–349.
[39] S. Nemati, R. Rohani, M.E. Basiri, M. Abdar, N.Y. Yen, V. Makarenkov, A hybrid latent space data fusion method for multimodal emotion recognition, IEEE Access 7 (2019) 172948–172964.
[40] A.G.C. Pacheco, R. Krohling, An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification, IEEE J. Biomed. Health Informat. (2021).
[41] S. Pathan, K.G. Prabhu, P. Siddalingaswamy, Automated detection of melanocytes related pigmented skin lesions: a clinical framework, Biomed. Signal Process Contr. 51 (2019) 59–72.
[42] L. Prechelt, Early stopping-but when?, in: Neural Networks: Tricks of the Trade Springer, 1998, pp. 55–69.
[43] M.F. Rasul, N.K. Dey, M. Hashem, A comparative study of neural network architectures for lesion segmentation and melanoma detection, in: 2020 IEEE Region 10 Symposium (TENSYMP), IEEE, 2020, pp. 1572–1575.
[44] S.E. Roslin, et al., Classification of melanoma from dermoscopic data using machine learning techniques, Multimed. Tool. Appl. (2018) 1–16.
[45] M. Rubin, O. Stein, N.A. Turko, Y. Nygate, D. Roitshtain, L. Karako, I. Barnea, R. Giryes, N.T. Shaked, Top-gan: stain-free cancer cell classification using deep learning with a small training set, Med. Image Anal. 57 (2019) 176–185.
[46] M. Saha, I. Arun, R. Ahmed, S. Chatterjee, C. Chakraborty, Hscorenet: A Deep Network for Estrogen and Progesterone Scoring Using Breast Ihc Images, Pattern Recognition, 2020, 107200.
[47] M. Samami, E. Akbari, M. Abdar, P. Plawiak, H. Nematzadeh, M.E. Basiri, V. Makarenkov, A mixed solution-based high agreement filtering method for class noise detection in binary classification, Phys. Stat. Mech. Appl. (2020), 124219.
[48] C.E. Shannon, A mathematical theory of communication, Bell Syst. Techn. J. 27 (1948) 379–423.
[49] Z. Sherkatghanad, M. Akhondzadeh, S. Salari, M. Zomorodi-Moghadam, M. Abdar, U.R. Acharya, R. Khosrowabadi, V. Salari, Automated detection of autism spectrum disorder using a convolutional neural network, Front. Neurosci. 13 (2019).
[50] 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 (2020) 350–360.
[51] L. Smith, Y. Gal, Understanding Measures of Uncertainty for Adversarial Example Detection, 2018 arXiv preprint arXiv:1803.08533.
[52] J. Snoek, H. Larochelle, R.P. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems, 2012, pp. 2951–2959.
[53] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, R. Adams, Scalable bayesian optimization using deep neural networks. International Conference on Machine Learning, 2015, pp. 2171–2180.
[54] F. Soboczenski, M.D. Himes, M.D. O'Beirne, S. Zorzan, A.G. Baydin, A.D. Cobb, Y. Gal, D. Angerhausen, M. Mascaro, G.N. Arney, et al., Bayesian deep learning for exoplanet atmospheric retrieval, in: Bayesian Deep Learning, NeurIPS 2018, 2018, pp. 1–6.
[55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[56] 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 (2020) 313–322.
[57] T.Y. Tan, L. Zhang, C.P. Lim, Adaptive melanoma diagnosis using evolving clustering, ensemble and deep neural networks, Knowl. Base Syst. 187 (2020), 104807.
[58] P. Tang, Q. Liang, X. Yan, S. Xiang, D. Zhang, Gp-cnn-dtel: global-part cnn model with data-transformed ensemble learning for skin lesion classification, IEEE J. Biomed. Health Informat. 24 (10) (Oct. 2020) 2870–2882, https://doi.org/10.1109/JBHI.2020.2977013.
[59] D. Wang, N. Pang, Y. Wang, H. Zhao, Unlabeled skin lesion classification by self-supervised topology clustering network, Biomed. Signal Process Contr. 66 (2021), 102428.
[60] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, Neurocomputing 338 (2019) 34–45.
[61] S. Wang, R.M. Summers, Machine learning and radiology, Med. Image Anal. 16 (2012) 933–951.
[62] J. Waring, C. Lindvall, R. Umeton, Automated Machine Learning: Review of the State-Of-The-Art and Opportunities for Healthcare, Artificial Intelligence in Medicine, 2020, 101822.
[63] T. Wei, C. Wang, Y. Rui, C.W. Chen, Network morphism, in: International Conference on Machine Learning, 2016, pp. 564–572.
[64] Y. Wen, P. Vicol, J. Ba, D. Tran, R. Grosse, Flipout: efficient pseudo-independent weight perturbations on mini-batches, in: International Conference on Learning Representations, 2018, pp. 1–16.
[65] K. Wickstrøm, M. Kampffmeyer, R. Jenssen, Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps, Med. Image Anal. 60 (2020), 101619.
[66] M. Wistuba, N. Schilling, L. Schmidt-Thieme, Scalable Gaussian process-based transfer surrogates for hyperparameter optimization, Mach. Learn. 107 (2018) 43–78.
[67] J. Wu, X.Y. Chen, H. Zhang, L.D. Xiong, H. Lei, S.H. Deng, Hyperparameter optimization for machine learning models based on bayesian optimization, J. Electr. Sci. Technol. 17 (2019) 26–40.
[68] Y. Yao, Three-way decisions with probabilistic rough sets, Inf. Sci. 180 (2010) 341–353.
[69] L. Yu, H. Chen, Q. Dou, J. Qin, P.A. Heng, Automated melanoma recognition in dermoscopy images via very deep residual networks, IEEE Trans. Med. Imag. 36 (2016) 994–1004.

[70] J. Zhang, Y. Xie, Q. Wu, Y. Xia, Medical image classification using synergic deep learning, Med. Image Anal. 54 (2019) 10–19.
[71] D. Zhuang, K. Chen, J.M. Chang, Cs-af: A Cost-Sensitive Multi-Classifier Active Fusion Framework for Skin Lesion Classification, 2020 arXiv preprint arXiv: 2004.12064.
[72] M. Zomorodi-moghadam, M. Abdar, Z. Davarzani, X. Zhou, P. Pławiak, U. R. Acharya, Hybrid particle swarm optimization for rule discovery in the diagnosis of coronary artery disease, Expet Syst. (2019), e12485.