Applied Soft Computing 146 (2023) 110666

Contents lists available at ScienceDirect

Applied Soft Computing

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

AUQantO: Actionable Uncertainty Quantification Optimization in deep learning architectures for medical image classification

Zakaria Senousy [a], Mohamed Medhat Gaber [a,b], Mohammed M. Abdelsamea [a,c,*]

[a] School of Computing and Digital Technology, Birmingham City University, Birmingham, UK
[b] Faculty of Computer Science and Engineering, Galala University, Egypt
[c] Department of Computer Science, Faculty of Computers and Information, University of Assiut, Egypt

ARTICLE INFO

Article history:
Received 6 December 2022
Received in revised form 5 June 2023
Accepted 13 July 2023
Available online 25 July 2023

Keywords:
Medical image analysis
Image classification
Deep learning
Convolutional neural networks
Uncertainty quantification
Actionability
XAI

ABSTRACT

Deep learning algorithms have the potential to automate the examination of medical images obtained in clinical practice. Using digitized medical images, convolution neural networks (CNNs) have demonstrated their ability and promise to discriminate among different image classes. As an initial step towards explainability in clinical diagnosis, deep learning models must be exceedingly precise, offering a measure of uncertainty for their predictions. Such uncertainty-aware models can help medical professionals in detecting complicated and corrupted samples for re-annotation or exclusion. This paper proposes a new model and data-agnostic mechanism, called Actionable Uncertainty Quantification Optimization (*AUQantO*) to improve the performance of deep learning architectures for medical image classification. This is achieved by optimizing the hyperparameters of the proposed entropy-based and Monte Carlo (MC) dropout uncertainty quantification techniques escorted by single- and multi-objective optimization methods, abstaining from the classification of images with a high level of uncertainty. This helps in improving the overall accuracy and reliability of deep learning models. To support the above claim, *AUQantO* has been validated with four deep learning architectures on four medical image datasets and using various performance metric measures such as precision, recall, Area Under the Receiver Operating Characteristic (ROC) Curve score (AUC), and accuracy. The study demonstrated notable enhancements in deep learning performance, with average accuracy improvements of 1.76% and 2.02% for breast cancer histology and 5.67% and 4.24% for skin cancer datasets, utilizing two uncertainty quantification techniques, and *AUQantO* further improved accuracy by 1.41% and 1.31% for brain tumor and 4.73% and 1.83% for chest cancer datasets while allowing exclusion of images based on confidence levels.

1. Introduction

Advances in computer-aided diagnosis (CAD) have a substantial impact in reducing the strain on medical professionals performing manual investigations and improving the detection accuracy of various diseases. One of the methods that has been used extensively for automated diagnosis is deep learning. Deep learning approaches have gained massive amounts of headway and achieved exceptional outcomes, driving numerous researchers to provide fair and automated solutions for a few diverse medical image analysis applications. For the image classification task, deep neural networks (DCNN) are considered one of the deep learning approaches that have been commonly used to extract prominent image features for various medical image analysis applications [1].

Medical image classification is the process of categorizing medical images into different classes based on their visual features. It is a complex task due to variability in image characteristics, the complexity of medical conditions, the limited availability of labeled data, and the need for explainable predictions. An example of medical image classification is the use of histopathological images to identify benign and malignant breast tumors. DCNNs can be trained on labeled images to learn the visual characteristics of different types of tumors. The trained models can then be used to classify new images accurately and efficiently.

Despite the ability of DCNNs to demonstrate outstanding performance in image classification tasks [2–4], an initial stage of explainability is required to measure the level of uncertainty in the input samples for medical image analysis applications. Building an uncertainty-aware model can help in identifying any ambiguity that could be therapeutically useful. Uncertainty awareness is also beneficial in terms of actionability to medical samples which could be possibly confusing and challenging to automated

* Corresponding author at: School of Computing and Digital Technology, Birmingham City University, Birmingham, UK.
E-mail address: mohammed.abdelsamea@bcu.ac.uk (M.M. Abdelsamea).

https://doi.org/10.1016/j.asoc.2023.110666

diagnosis systems. It additionally permits clinical experts to rate images that ought to be focused on for manual examination. The motivation for this work is to provide a robust method for image classification models that can help medical practitioners in automated diagnosis. The proposed method includes an uncertainty-aware model to identify ambiguous samples and an automated actionability component to guide medical experts in identifying contaminated samples. By improving the accuracy and interpretability of medical image classification models, this work aims to reduce the strain on medical practitioners and enhance the detection accuracy of various diseases.

In this paper, we propose a model agnostic mechanism, coined Actionable Uncertainty Quantification Optimization (*AUQantO*),[1] to optimize the performance of deep learning architectures for medical image classification. *AUQantO* is guided by uncertainty measurements that help clinical experts refine annotations to develop more reliable DCNN models. *AUQantO* employs either an entropy-based mechanism [5] or a Monte-Carlo (MC) dropout [6] technique to measure uncertainty in images, where a new hyperparameter (i.e., a threshold) is introduced and optimized.

Our approach focuses on determining whether a deep learning architecture is reliable for sample prediction. *AUQantO* maximizes the image confidence score rather than focusing on sample quality uncertainty, and this is done in an automated framework. In other words, our sample exclusion mechanism works on excluding samples based on high uncertainty scores of image predictions generated from deep learning architectures rather on the quality of the image itself during image acquisition (e.g., issues in sample preparation, noises, and artifacts). Based on the aforementioned claim, it is important to highlight that our method does not integrate any relevant information to the quality of images in the training process. The aim of this research work is to maximize the confidence score of deep learning architectures and to generate maximum certainty.

Unlike the work presented in [7] and [8], which are based on training uncertainty-aware models that require specific experimental settings, our method does not require training as it is a model agnostic approach which is utilized as a post-prediction component (i.e., a wrapper method) that can be applied on predictions generated from deep learning models. Moreover, the flexibility of our approach comes from its ability to be applied to any dataset (dataset-agnostic). This makes *AUQantO* a fast and light method compared to stand-alone trainable models. Based on the aforementioned claim, it is worth mentioning that the work proposed in [7] and [8] can be wrapped using *AUQantO* as our method can be utilized to optimize a threshold hyperparameter which is compared against uncertainty measures used by the proposed work in the literature. Therefore, a direct comparison with the proposed work is not possible. To the best of our knowledge, *AUQantO* is the first method to introduce automated actionability based on uncertainty awareness as a model/data set-agnostic approach. The performance of *AUQantO* has been validated using state-of-the-art deep learning architectures (with several optimization methods) on four medical image datasets.

The contributions of the paper can be summarized as follows:

- Introduced a wrapper method accompanied by different uncertainty-aware techniques to measure the uncertainty of predictions generated from deep learning architectures.
- Proposed an optimized automated actionability component for deep learning architectures, which guides medical experts in identifying contaminated samples for re-annotation or exclusion.

[1] https://github.com/zakariaSenousy/AUQantO-Method

The paper is structured as follows. Section 2 presents the relevant background knowledge required for our method. In Section 3, we review related work on uncertainty quantification for medical images. Section 4 discusses, in detail, our proposed *AUQantO* method. Our experimental results and findings are explained in Section 5. Section 6 concludes our work.

2. Background

Deep learning has emerged as a powerful paradigm in the field of artificial intelligence, enabling significant advancements in various domains such as computer vision, natural language processing, and speech recognition. At its core, deep learning involves training neural networks with multiple layers to learn hierarchical representations of data, leading to state-of-the-art performance in complex tasks [9].

Deep learning architectures play a crucial role in the experimental study, encompassing two distinct classes of architectures. The first class comprises single deep learning architectures, including the two-stage CNN [2], the deep spatial fusion CNN (DSF-CNN) [3], and the hybrid LSTM [10]. In this class, the input image is segmented into small patches and processed through a feature extraction network known as the patch-wise network. The extracted feature maps are subsequently fed into an image-wise network for the final classification. The second class encompasses ensemble architectures, such as EMS-Net [4], which involve multiple deep learning models working together to learn image features from different perspectives, thus introducing diversity in the final ensemble of image predictions.

While deep learning models have shown impressive performance, quantifying and managing uncertainty in predictions is crucial for many real-world applications. Uncertainty quantification techniques aim to provide measures of confidence or reliability in deep learning models' predictions. These techniques enable decision-making under uncertain conditions, robustness to noisy or out-of-distribution inputs, and model calibration [11]. Uncertainty quantification techniques, such as Shannon entropy [5] and Bayesian approximation using MC dropout [6], are used in *AUQantO* to measure the confidence and uncertainty levels of deep learning image predictions. Shannon Entropy quantifies ambiguity by analyzing the probability distribution of predicted class labels, while MC Dropout generates multiple probability distributions for each input image to compute the mean prediction and standard deviation. These techniques help identify and exclude uncertain samples, improving the reliability of the overall system.

Optimization methods are essential for training deep learning models by minimizing the loss function and optimizing their parameters. Gradient-based optimization algorithms, like stochastic gradient descent (SGD) and its variants, are widely used in deep learning [12,13]. These methods efficiently update the model parameters by computing gradients through backpropagation and iterative adjustment of the weights.

The optimization methods utilized in this study aim to solve the non-convex objective function associated with the exclusion rate of images and their impact on the accuracy of the deep learning architecture. Due to the stochastic nature of the objective function, finding the optimal threshold hyperparameter involves a random search within a search space that may contain multiple local minima. To tackle this, effective optimization methods for nonconvex problems are employed. These methods include Bayesian Optimization using Gaussian Processes (GP) [14], which approximates objective functions using non-parametric statistical models. Furthermore, Constrained Optimization by Linear Approximation (COBYLA) [15] employs a simplex-based approach, and Dual Annealing [16] applies a generalized simulated annealing algorithm combined with local search. Finally,

the Non-dominated Sorting Genetic Algorithm (NSGA-II) [17] is used for multi-objective optimization, addressing complex interactions and non-linearities through an evolutionary approach. These methods ensure the exploration of the solution space to find the optimal threshold hyperparameter for image exclusion in the context of uncertainty quantification.

3. Related work

Several papers propose various algorithms and frameworks for the processing and analysis of medical data [18–24]. Uncertainty quantification methods are essential to reduce the effect of uncertainties during the decision-making process (actionability). They have been used to solve a wide range of real-world scientific and engineering problems including computer vision, autonomous driving control, risk uncertainties and medical image diagnosis. For instance, various methods have been proposed to introduce uncertainty quantification for computer vision applications such as image/video retrieval [25], semantic segmentation [26], and object detection [27]. Additionally, uncertainty quantification has been used to guarantee safety measures for autonomous driving control where various uncertainty measures can be calibrated for the collision avoidance task [28]. Moreover, risk measures and uncertainty estimates have been studied for deep learning in [29].

In medical image diagnosis, uncertainty quantification is a critical step towards explainable Artificial Intelligence (XAI) [30, 31]. In other words, the uncertainty estimates can provide valuable information about the model's confidence in its predictions, allowing for more transparent and trustworthy explanations of its behavior. This can lead to better understanding and trust in the model's decision-making process, ultimately resulting in improved explainability. A few recently proposed image classification models have been developed in a way to be aware of the uncertainty of the final decision. For instance, an instability map has been used to show regions of ambiguity in a CNN-based model as a measure of uncertainty in the research published in [32]. Fraz et al. [33] has developed a system for micro-vessel segmentation that contained an uncertainty quantification block for histopathological images.

A Bayesian DenseNet-169 has been proposed in [34] for skin lesion images. To create uncertainty measurements, the model triggers dropout layers during the testing phase. Ablation research was carried out to demonstrate how Bayesian deep learning may help diagnostic systems and medical professionals collaborate in the classification of skin lesions. Similarly, a reliable, accurate and active Bayesian network termed (ARA-CNN) for image classification has been presented to categorize colorectal cancer images [35]. For evaluating the uncertainty of the input data, the model is built on residual networks (ResNets) and variational dropout. Another work applied quantitative comparison of MC dropout uncertainty measures for multi-class predictions for medical image analysis [36]. The work introduced in [37] proposed a deep learning model that can handle uncertain inputs. The work uses entropy values and a non-dominant sorting algorithm to identify candidates with the highest entropy value from the dataset.

Although these methods have shown their effectiveness in introducing uncertainty measures in input samples, they lack (1) actionability in identifying the images to be classified/excluded based on the uncertainty of the predictions, and (2) accurate learning strategy to improve model performance (for example, ensemble learning).

Recently, an entropy-based elastic ensemble of DCNNs (3E-Net) has been introduced [38] for breast cancer grading. The 3E-Net model builds an ensemble of image classification networks supported by a patch-wise network (DenseNet-161 [39]) for feature extraction. The model uses entropy to determine the amount of ambiguity in image predictions. Similarly, a model called Multi-level Context and Uncertainty-aware model (MCUa) [40] has been introduced, which employs different levels of spatial feature learning to generate an ensemble of models which support different image scales and architectures for breast cancer histopathology classification. MCUa generates a series of probability distributions using Monte Carlo dropout [6] to determine the amount of ambiguity in the input data. The work presented by Abdelsamea et al. [41] studied actionability in computational pathology applications using uncertainty quantification methods.

Despite the success of recent models in introducing uncertainty-aware components to deep learning models, an automated actionable method that can automatically exclude an optimal number of poor samples is required. This is important in clinical practice to minimize the workload of the medical professional and improve the trust in deep learning models.

4. AUQantO method

In this section, we explain our proposed (*AUQantO*) approach for optimizing uncertainty quantification in deep learning architectures. As illustrated in Fig. 1, an input image is fed into the deep learning architecture for classification. As a pre-stage to our method, *AUQantO* requires deep learning architectures that can generate probability distributions for their input samples. This requirement helps *AUQantO* to generate an uncertainty score for the image probability distribution and decides on the poor medical samples that need to be manually investigated by medical experts.

Consequently, *AUQantO* (as an uncertainty-aware method) has been designed based on Shannon Entropy [5] or MC dropout [6]. Shannon Entropy is based on the image predictions (or the probability distribution of the output, where each value is associated with a class in the training set) generated by the deep learning architecture. Shannon entropy is adopted to generate an uncertainty score to indicate how confident the model is in classifying the input image. On the other hand, MC dropout uses dropout layers in the deep learning architecture network for image classification and activates them during the testing phase, resulting in a list of probability distributions whose mean prediction determines the image's final classification while the standard deviation provides a measure of uncertainty. To automatically exclude poor image samples and keep confident ones for final classification, *AUQantO* introduces a new hyperparameter, which we call threshold (λ). In this work, the optimal threshold value (which aids in excluding the optimal number of poor samples) is explored by single and multi-objective optimization methods.

AUQantO can quantify the uncertainty in medical image samples and automatically tune the threshold hyperparameter against uncertainty values to exclude highly uncertain images. A well optimized threshold in this context would depend on the specific characteristics of the dataset and the deep learning architecture used in the study. It should be set in such a way that it maximizes the trade-off between the accuracy and the robustness of the model. That is, it should be low enough to filter out samples with high uncertainty and reduce the risk of misclassification, but not too low that it excludes informative samples that could contribute to improving model performance.

4.1. Objective function

Here, we explain in detail the single- and multi-objective functions utilized to build our *AUQantO* method. The main purpose of introducing both functions comes from the aim of developing an actionable method that can work on minimizing the number of excluded images from a particular dataset. This is done by

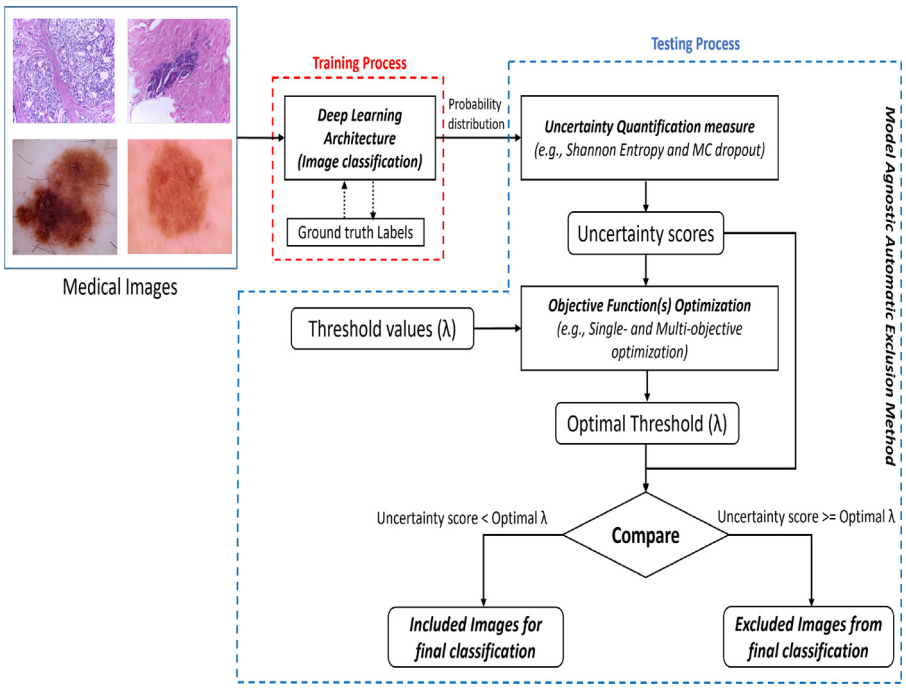

Fig. 1. The *AUQantO* method employs two uncertainty quantification techniques, one using Shannon Entropy on a single probability distribution and the other processing a list of distributions from the dropout layer, to optimize a threshold value for image classification and expert review, with training represented by a red block and testing by a blue block.

considering the levels of certainty of the included and excluded images. In other words, we aim to develop objective functions that work on selecting an optimal hyperparameter threshold for maximizing the rate of highly certain included images to be classified and minimize the rate of excluded samples that are highly uncertain based on probability values generated from deep learning models.

4.1.1. Single-objective function

Our objective function has introduced a new hyperparameter (λ) has been introduced by our objective function, to be tuned based on the generated uncertainty scores. This is by checking if the input image has an uncertainty score greater than the λ then the image will be excluded from the final classification process, otherwise, the image will be classified by the model. More precisely, to calculate the optimal threshold value (λ), we introduce a single-objective function to be minimized. Our objective function has two terms that have been designed to encode the confidence of probability distributions for both included and excluded images. We used cross-entropy for the probability distributions against the ground-truth labels. For example, we customized the cross-entropy equation by multiplying the probability distribution of a given image by the one hot-encoding labeling for the same image. Consequently, we formulate H_{exc} and H_{inc} to present a summation of cross-entropy values for excluded and included images, respectively. H_{exc} and H_{inc} can be represented as:

$$H_{exc} = \sum_{i=1}^{n}\sum_{j=1}^{c} p_{ij} \times q_{ij} \tag{1}$$

$$H_{inc} = \sum_{i=1}^{m}\sum_{j=1}^{c} p_{ij} \times q_{ij} \tag{2}$$

where p_{ij} represents the probability value j over c class probability values, while q_{ij} represents the hot encoding value j over c class categories of image i over either n excluded images or m included images.

The average cross-entropy for both excluded and included images is then calculated by dividing H_{exc} and H_{inc} by n excluded images and m included images, respectively. Using single-objective optimization methods, the main target is to reach an optimal λ which minimizes the summation of the two terms of the objective function.

For example, a possible scenario to minimize the cross-entropy of excluded images H_{exc} is to have cases where images are misclassified with high confidence. This means that the evaluation of the cross-entropy equation (assuming we have a classification problem of two classes) will have a very small probability value p (tends to zero) for the correct class multiplied by $q = 1$ to represent the one hot encoding of the correct class. While, for an incorrect class, a very high probability value p is multiplied by $q = 0$. A similar scenario could happen for the maximization of the cross-entropy of included images H_{inc} by having images that are correctly classified with high confidence and by subtracting this term from a value of one, we convert it into a term to be minimized instead. Both scenarios for included and excluded images lead to a very small value for the output of the objective function, and hence we can reach high level of trustworthiness for included images that are classified by a deep learning architecture and exclude images that are truly uncertain with high confidence for further annotation and investigation by medical experts. The single-objective function can be defined as:

$$F(\lambda) = argmin_{(SE|\sigma<\lambda \| SE|\sigma\geq\lambda)}\left(\left(\frac{H_{exc}}{n}\right) + \left(1 - \frac{H_{inc}}{m}\right)\right) \tag{3}$$

where $F(\lambda)$ is the output of the single-objective function and λ is the optimal threshold value. λ is verified by Shannon Entropy SE or MC dropout's standard deviation σ to differentiate between included and excluded image groups and measure the average cross-entropy.

4.1.2. Multi-objective function

As can be noticed from the above-mentioned single-objective function, that we have two terms to work on both included and

excluded images. The two terms can be presented in two separate objective functions that can be optimized simultaneously to reach the optimal threshold which achieves the selection of (1) highly certain images to be included in the final classification and (2) highly uncertain images to be excluded from classification and to be returned to medical experts for manual exploration. In that sense, we introduce a multi-objective function with the target of maximizing the rate of included images and minimizing the rate of excluded images based on their uncertainty and confidence of deep learning architecture's predictions.

Our multi-objective function can be defined as:

$$\min \{F_{exc}(\lambda), F_{inc}(\lambda)\} \tag{4}$$

where:

$$F_{exc}(\lambda) = H_{exc}/n$$
$$F_{inc}(\lambda) = 1 - (H_{inc}/m)$$

subject to:

$$n \geq 1$$
$$\lambda \leq SE|\sigma \quad for\ F_{exc}(\lambda)$$
$$m \geq 1$$
$$\lambda > SE|\sigma \quad for\ F_{inc}(\lambda)$$
$$\lambda_{\min} \leq \lambda \leq \lambda_{\max},$$

where $F_{exc}(\lambda)$ and $F_{inc}(\lambda)$ represent the objective functions which are based on the average cross-entropy of excluded and included images, respectively. The number of excluded images n and included images m in the multi-objective function is subject to the number of images not less than the value of one. Moreover, λ value is subject to a pre-defined range ($\lambda_{\min}$ to $\lambda_{\max}$) while using a multi-objective optimization method.

Algorithm 1 presents a description of the workflow of our *AUQantO* method. As described in the algorithm, our method works on the probability distributions of input images generated from the deep learning architectures used in the study. The probability distribution(s) is/are used by either Shannon Entropy or MC dropout to generate uncertainty measures. To prepare for the optimization stage, we use either single- or multi-mode for applying single- or multi-objective functions. Once we reach the optimal threshold using various optimization methods, we compare the uncertainty scores generated for all dataset images with the optimal threshold. The highly uncertain images are automatically excluded from the final classification.

4.2. Computational complexity

The computational complexity of our method depends on some important stages: the uncertainty measure, and the optimization methods associated with the objective functions. For the uncertainty measure, we used the Shannon entropy which has the complexity of O(n) independent of the number of classes associated with the classification problem. In MC dropout, as described in the background section, we apply a number of test passes over the CNN network to generate a number of probability distributions that are used later for measuring the uncertainty. This indicates that the complexity depends on the time spent t processing the number of test passes used and the settings associated with the CNN used. Hence, we present the complexity for the MC-dropout method as O($t*hwcfl$) where h and w indicate the height and width of the input sample, c represents channels, f represents the number of filters, and l represents the number of layers. In our study, we used four optimization methods for two objective functions (single- and multi-objective functions). The four optimization methods have complexity order of: O(n^3) for Gaussian processes [42], O($n * k^2$) per iteration for COBYLA [15], O(n) for dual annealing [43] (where n is the number of variables and k is the number of constraints), and O(MN^2) for NSGA-II (where M is the number of objectives and N is the population size) [44].

Algorithm 1: AUQantO Method

Input: Images from Dataset D
Output: Decision of automated classification/exclusion based on uncertainty
$D = x_1, x_2, ..., x_n$ // n images from Dataset D
/* Deep Learning Architecture stage */
Predictions = [ ] // Empty list to store all probability distributions generated for all dataset images
for $x \in D$ **do**
  // DeepLearningModel is a model function which takes image x as an input and generates probability distribution p
  $p_i = DeepLearningModel(x_i)$
  Predictions.append(p_i)
end
/* Uncertainty measure stage */
// UncertaintyMeasure is a function that uses either Shannon Entropy or Monte-Carlo dropout for generating uncertainty scores
$Scores = UncertaintyMeasure(Predictions)$
$ThresholdRange = \lambda_1, \lambda_2, ..., \lambda_n$ // A list of threshold values (λ) range that is initialized to be used by the optimization methods for finding for the optimal threshold based on objective functions
/* Objective functions */
if *Mode* == "*Single*" **then**
  /* Use the Single-objective function */
  $F(\lambda) = argmin_{(SE|\sigma<\lambda||SE|\sigma\geq\lambda)}((H_{exc}/n) + (1 - H_{inc}/m))$
end
else
  /* Use the multi-objective function */
  $F(\lambda) = \min\{F_{exc}(\lambda), F_{inc}(\lambda)\}$ // check equation 10
end
/* Optimization method */
// OptimizationMethod function uses either single or multi-objective function, the threshold range list, and the optimization method algorithm (e.g., NSGA-II)
$OptimalThreshold = OptimizationMethod(F, ThresholdRange, algorithm)$
/* Comparison of optimal threshold against uncertainty scores of dataset images */
for $score \in Scores$ **do**
  if $score_i > OptimalThreshold$ **then**
    automatically Exclude image x_i from final classification due to high uncertainty
    Return to a medical professional for further investigation
  end
  else
    Generate final image classification for included image x_i
  end
end
Evaluate the performance of Included images
Evaluate the performance of Excluded images

5. Experimental study

We validated *AUQantO* with 16 different case studies, where the case studies are associated with four different deep learning architectures on two medical datasets using both Shannon-entropy and MC dropout uncertainty quantification methods. Also, we used the best performing *AUQantO* version to be applied on two other datasets. A 5 × 4 nested cross-validation has been used to evaluate the performance of the methods in all the case studies. A 5 × 4 Nested cross-validation indicates that we divide a particular dataset into 5 folds. Four out of the five folds are treated as a smaller dataset with 4-fold cross-validation where 3 folds are used for training and 1-fold used for validation. This process is repeated among the 4 folds until we have each fold as validation set. Then, the optimal model hyperparameter which achieves the highest validation accuracy is selected to be applied on the testing fold. Each fold of the 5 folds is selected as the testing fold and the remaining 4 folds as training-validation folds until we evaluate the average testing accuracy of the 5 folds. The 5 × 4 nested cross-validation makes the splitting process of the utilized datasets as follows: 20% for testing and 80% for training-validation (60% training and 20% validation).

5.1. Datasets

In this work, we applied the following main medical image datasets to the 16 case scenarios of *AUQantO*:

Table 1
Description of the utilized datasets.

Dataset	Cancer type	# of samples	# of classes	Classes distribution
Histopathology microscopic	Breast	400	4	Normal: 100 Benign: 100 InSitu: 100 Invasive: 100
Dermoscopic	Skin	3297	2	Benign: 1800 Malignant: 1497

5.1.1. Breast cancer dataset

BreAst Cancer Histology images (BACH) dataset [45] of hematoxylin & eosin stained breast cancer histology images divided into two parts (A and B). Images of part B were provided for pixel-wise classification tasks. Consequently, in this work, we used images of part A of the dataset which is composed of 400 microscopy images of size 2048 × 1536 pixels and 20× magnification level. The 400 images are divided into four groups (normal, benign, in situ, and invasive).

5.1.2. Skin cancer dataset

Skin cancer dataset [46] is introduced by the International Skin Imaging Collaboration (ISIC). More than 2000 people contributed 33,126 dermoscopic images of benign and malignant skin lesions. For computational and memory efficiency, we utilized a smaller version of the dataset[2] which comprises of 3297 image samples (with 224 × 224 pixels) distributed between the two classes of skin lesions as 1800 images for benign and 1497 for malignant. Table 1 presents a description of the utilized datasets.

5.2. Experimental setup

To evaluate the single architectures (Two-stage CNN, DSF-CNN, and Hybrid LSTM) on a skin dataset, we used resized images of 224 ×224 pixels. During the training of a patch-wise network of single architectures, we extracted overlapped image patches of size 112 × 112 pixels from input images using a patch stride of 56. For the image-wise network of single architectures, we extracted non-overlapped image patches using a patch-stride of 112. We used data augmentation to rotate the training patches 90 degrees while flipping them horizontally and vertically. Adam optimizer [13] has been used to reduce the loss function of the networks. Patch-wise and image-wise networks are trained using a learning rate of 0.0001 and a batch size of 32. We used 7 and 4 epochs for training patch-wise and image-wise networks of Two-stage CNN, respectively. While for the other single architectures (DSF-CNN and Hybrid LSTM), we used 2 and 4 epochs for training patch-wise and image-wise networks, respectively.

In the BACH dataset, we used the original image size (2048 × 1536 pixels) as input to the single architectures (two-stage CNN, DSF-CNN, and hybrid LSTM). We extracted overlapping image patches of size (512 × 512 pixels) using a patch stride of 256 to train the patch-wise network. Non-overlapped image patches are used for the image-wise network of the single architectures (using patch stride 512). For single architectures except for Two-stage CNN, we utilized 2 epochs for the training of patch-wise networks and 4 epochs for the training of image-wise networks. While for Two-stage CNN, we utilized 8 training epochs for patch-wise networks of Two-stage CNN and 6 epochs for training image-wise networks.

Lastly, we employed an ensemble architecture (EMS-Net) for the two datasets (BACH and skin). We utilized the exact hyperparameter settings for the BACH dataset as described in [4]. This is by utilizing two image scale levels (scale 1: 448 × 336, scale 2: 296 × 224) for the three pre-trained DCNN models. We extracted patches of size 224 × 224 pixels and fine-tuned the pre-trained DCNN models based on the BACH dataset. We changed the number of neurons in the last fully connected layer of the pre-trained models to 4 (where BACH has 4 classes). Moreover, during the training, we used patch strides of 28 and 9 for scales 1 and 2, respectively, while during testing, we used patch strides of 56 and 18 for scales 1 and 2, respectively. Finally, we followed the same augmentation settings similar to the single architectures and we used Adam optimizer with learning rate of 0.0001 and batch size of 32. We utilized 2 epochs for training the pre-trained models of the ensemble architecture.

We applied a similar strategy to the one used for the EMS-Net on the BACH dataset to the skin dataset. We utilized two image scales (scale 1: 224 × 224, scale 2: 112 × 112) for the three pre-trained DCNN models. we extracted patches of size 112 × 112 and 56 × 56 for scales 1 and 2, respectively. We modified the number of neurons in the last fully connected layer of the pre-trained models to 2 (where the skin dataset has 2 classes). We used patch strides of 56 for scale 1 and 28 for scale 2 during the training and testing phases. We used 2 epochs for training pre-trained models applied on scale 1 and 6 epochs for training the pre-trained model applied on scale 2. Finally, the remaining settings in terms of data augmentation and Adam optimizer are the same as EMS-Net on BACH dataset. As can be seen from the settings, we employed to evaluate the *AUQantO* method using different dataset image settings (e.g., image scales) and different deep learning architectures including pre-trained DCNN models.

We employed the four optimization methods explained earlier (Bayesian optimization using GP, COBYLA, dual annealing, and NSGA-II) to all case studies. In the single-objective optimization methods (Bayesian optimization using GP, COBYLA, and dual annealing), we set the λ range from 1×10^{-9} to 2 while the evaluation step is set to 50. In COBYLA, the initial search point is set to 0.01. Finally, in the multi-objective optimization method (NSGA-II), we set the number of variables to 1 as we optimize only one hyperparameter (e.g., λ), number of objectives to 2, number of generations to 50, population size to 1, and we utilized the same λ range as in the single-objective optimization.

To perform the uncertainty measure using the Bayesian approximation with MC dropout to the deep learning architectures, we employed 50 test runs (which has been proved to be adequate to establish a valid distribution) for each image.

5.3. Results and analysis

In this work, we introduce three different metrics to measure the effectiveness and robustness of *AUQantO*. First, we introduce Weighted Average Accuracy (*WAA*), which measures average classification accuracy weighted by the included images in each test fold. Second, Accuracy Difference (*AD*) measures the difference between the accuracy of included images and the accuracy of excluded images. Third, The Abstain Percentage (*AP*) calculates the proportion of excluded images in each dataset compared to

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

Table 2
Average test performance (without image exclusion — AUQantO method) for case studies conducted on the BACH dataset.

Architecture	Precision	Recall	AUC	Accuracy
Two-stage CNN	88.50%	88.25%	93.39%	88.25%
DSF-CNN	91.50%	91.25%	98.78%	91.25%
Hybrid-LSTM	90.60%	90.25%	98.97%	90.25%
EMS-Net	93.42%	93.25%	97.07%	93.25%

Table 3
Average test performance (without image exclusion — AUQantO method) for case studies conducted on the Skin dataset.

Architecture	Precision	Recall	AUC	Accuracy
Two-stage CNN	84.01%	83.90%	83.89%	83.90%
DSF-CNN	90.26%	90.14%	90.21%	90.14%
Hybrid-LSTM	90.14%	90.02%	90.10%	90.02%
EMS-Net	91.41%	91.30%	91.36%	91.30%

the total number of images. The three metrics are formulated as follows.

$$WAA = \frac{1}{\sum_{i=1}^{r} L_i} \sum_{i=1}^{r} Acc_i \times L_i \tag{5}$$

$$AD = WAA_{inc} - WAA_{exc} \tag{6}$$

$$AP = \left(\frac{\sum_{i=1}^{r} V_i}{D} \right) \times 100 \tag{7}$$

where L_i is the number of images (whether they are included or excluded images) in fold i. Acc_i is the classification accuracy in fold i on a total number of r folds. Acc equals to $(TP + TN)/(TP + TN + FP + FN)$, where TP and TN represent the correct predictions by our model, while FP and FN represent the incorrect predictions. WAA_{inc} and WAA_{exc} are the weighted average accuracy for included and excluded images, respectively. V_i is the number of images excluded in fold i, while D is the total number of images in each dataset. Furthermore, we use other metrics such as precision, recall, and Area Under the Receiver Operating Characteristic (ROC) Curve score (AUC). Similar to *WAA*, these metrics are also weighted by the number of included images in each test fold.

5.3.1. Performance of deep learning architectures without AUQantO

Tables 2 and 3 show the average test performance for all case studies before applying the *AUQantO* method to exclude images. After evaluating the deep learning models on each dataset, it can be noticed that EMS-Net has higher test accuracy on BACH (93.25%) and skin (91.30%) datasets among all deep learning architectures. This is due to the usage of an ensemble architecture that applies diversity in learning the variety of image features. While, in terms of single deep learning architectures, we can see that DSF-CNN and Hybrid-LSTM have high average test accuracy comparable to EMS-Net. DSF-CNN achieved an average test accuracy of 91.25% and 90.14% on BACH and skin datasets, respectively, while Hybrid-LSTM achieved an average test accuracy of 90.25% and 90.02% on BACH and skin datasets, respectively.

5.3.2. Performance of AUQantO method on BACH dataset

Table 4 demonstrates the performance of *AUQantO* (in terms of the weighted average test precision, recall, AUC, and accuracy of included images) using Shannon Entropy and MC dropout with four optimization methods (Bayesian optimization using GP, COBYLA, Dual Annealing, and NSGA-II). *AUQantO* shows a significant improvement in all case studies applied in the BACH data set using the four optimization methods. Also, this improvement showed the capability of *AUQantO* in automatically excluding poor samples that are misclassified or even the uncertain images that are correctly classified. Moreover, NSGA-II demonstrated the highest average test accuracy among other optimization methods for three case studies: Two-stage (90.29%) and DSF-CNN (97.61%) using Shannon Entropy and Two-stage using MC dropout (89.72%). Dual annealing and COBYLA showed the highest test accuracy performance on EMS-Net (Entropy: 94.78% and MC: 94.68%) and Hybrid-LSTM (Entropy: 92.65% and MC: 92.23%), respectively. GP showed the highest accuracy on DSF-CNN using MC dropout (97.92%). In terms of the performance of single architectures, DSF-CNN showed higher performance for all optimization methods on BACH dataset compared to Two-stage CNN and Hybrid-LSTM.

The results presented in Table 4 showed the performance of the included images using the *AUQantO* method. To show how effective our method is in excluding poor samples, we present the test results of the excluded images by *AUQantO* using the four optimization methods in Table 5. This aids in linking the high performance of our method resulting from including certain images and the low performance resulting from the excluded uncertain images.

In Table 5, as described above, we present the performance of *AUQantO* on the excluded poor samples along with the abstain percentage (which presents the number of excluded images to the total number of images). As shown in Table 5, the excluded image accuracy (in all case studies) varies between 20% to 74% which indicates how effective our method is in excluding poor samples. Moreover, Hybrid-LSTM excluded the least number of images with the lowest abstain percentage for all optimization methods among all architectures using the two uncertainty measures (entropy and MC dropout). Furthermore, the evaluation of Hybrid-LSTM on BACH dataset showed very low excluded images accuracy for both uncertainty measures: 25%, 42.11%, 20%, and 25% were reported for GP, COBYLA, dual annealing, and NSGA-II, respectively, using Shannon Entropy. While for MC dropout, accuracy of 27%, 43.22%, 22.00%, and 26.00% were reported for GP, COBYLA, dual annealing, and NSGA-II, respectively. This shows that *AUQantO* is effective and successful in minimizing the exclusion rate by excluding the most challenging and poor samples that require manual investigation by medical experts.

To further demonstrate the effectiveness of our method in excluding poor image samples, Fig. 2 shows (1) Accuracy Improvement, which presents the level of improvement (in terms of test accuracy) reported after using *AUQantO* by excluding the uncertain samples and (2) Accuracy Difference (AD), which represents the difference between average test accuracy of included and excluded images.

Figs. 2(a) and (b) show the accuracy improvement reported in all deep learning models in the BACH data set using the four optimization methods by Shannon Entropy and MC dropout, respectively, which confirms the effectiveness of *AUQantO*. As seen in Figs. 2(a) and (b), our method succeeded in improving the accuracy of all deep learning models using both uncertainty measures. DSF-CNN reported the highest level of accuracy improvement for both uncertainty measures, where accuracy values of around 2% to almost 7% were achieved using optimization methods. The other deep learning architectures (Two-stage, Hybrid-LSTM, and EMS-Net) reported an improvement of less than 2.5% for all optimization methods in both uncertainty measures. Figs. 2(c) and (d) show the difference in accuracy between included and excluded images in all deep learning models on the BACH dataset using the four optimization methods by Shannon Entropy and MC dropout, respectively. Hybrid-LSTM showed the highest accuracy difference with all optimization methods (Figs. 2(c) and (d)) achieving an accuracy difference of around 65% to 70% for GP, dual annealing, and NSGA-II and 50% for COBYLA. The other deep learning architectures reported an accuracy difference of almost 20% to 30% for the optimization methods used.

Table 4
Average test performance of included images using AUQantO method for case studies conducted on BACH dataset using Shannon Entropy and MC dropout as uncertainty measures.

Architecture (uncertainty measure)	Optimization method															
	GP				COBYLA				Dual annealing				NSGA-II			
	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC	Accuracy
Two-Stage (Entropy)	89.72%	89.34%	94.57%	89.34%	89.23%	88.99%	94.22%	88.99%	89.53%	89.16%	94.42%	89.16%	90.64%	90.29%	95.10%	90.29%
DSF-CNN (Entropy)	94.79%	94.57%	99.22%	94.57%	93.70%	93.35%	99.17%	93.35%	93.88%	93.58%	99.18%	93.58%	97.84%	97.61%	99.60%	97.61%
Hybrid-LSTM (Entropy)	92.00%	91.58%	99.15%	91.58%	92.91%	92.65%	99.26%	92.65%	91.41%	91.14%	99.04%	91.14%	92.60%	92.27%	99.20%	92.27%
EMS-Net (Entropy)	94.95%	94.75%	97.04%	94.75%	94.69%	94.44%	96.93%	94.44%	94.92%	94.78%	97.66%	94.78%	94.90%	94.68%	97.09%	94.68%
Two-Stage (MC)	89.34%	89.10%	94.21%	89.10%	89.02%	88.81%	94.05%	88.81%	89.29%	89.12%	94.08%	89.12%	89.94%	89.72%	94.32%	89.72%
DSF-CNN (MC)	98.12%	97.92%	99.61%	97.92%	95.15%	94.97%	99.26%	94.97%	96.24%	96.00%	99.28%	96.00%	97.03%	96.76%	99.40%	96.76%
Hybrid-LSTM (MC)	91.56%	91.32%	99.24%	91.32%	92.47%	92.23%	99.47%	92.23%	91.34%	91.16%	99.06%	91.16%	92.29%	92.03%	99.36%	92.03%
EMS-Net (MC)	94.81%	94.63%	97.32%	94.63%	94.47%	94.24%	97.07%	94.24%	94.88%	94.68%	97.47%	94.68%	94.76%	94.59%	97.21%	94.59%

Table 5
Average test accuracy of excluded images and (abstain percentage of dataset images) using AUQantO method for case studies conducted on BACH dataset.

Architecture (uncertainty measure)	Optimization method			
	GP	COBYLA	Dual annealing	NSGA-II
Two-Stage (Entropy)	71.85% (13.25%)	72.49% (10.50%)	72.46% (11.25%)	74.24% (27.50%)
DSF-CNN (Entropy)	68.00% (12.50%)	71.79% (9.75%)	71.43% (10.50%)	73.83% (26.75%)
Hybrid-LSTM (Entropy)	25.00% (2.00%)	42.11% (4.75%)	20.00% (1.25%)	25.00% (3.00%)
EMS-Net (Entropy)	57.90% (4.75%)	63.64% (5.50%)	58.82% (4.25%)	66.67% (6.00%)
Two-Stage (MC)	69.86% (11.25%)	71.45% (9.25%)	71.57% (10.50%)	72.67% (22.50%)
DSF-CNN (MC)	74.10% (28.00%)	70.97% (15.50%)	70.67% (18.75%)	72.52% (22.75%)
Hybrid-LSTM (MC)	27.00% (2.50%)	43.22% (5.00%)	22.00% (1.75%)	26.00% (3.25%)
EMS-Net (MC)	62.90% (4.50%)	67.50% (5.00%)	61.24% (4.50%)	65.67% (5.20%)

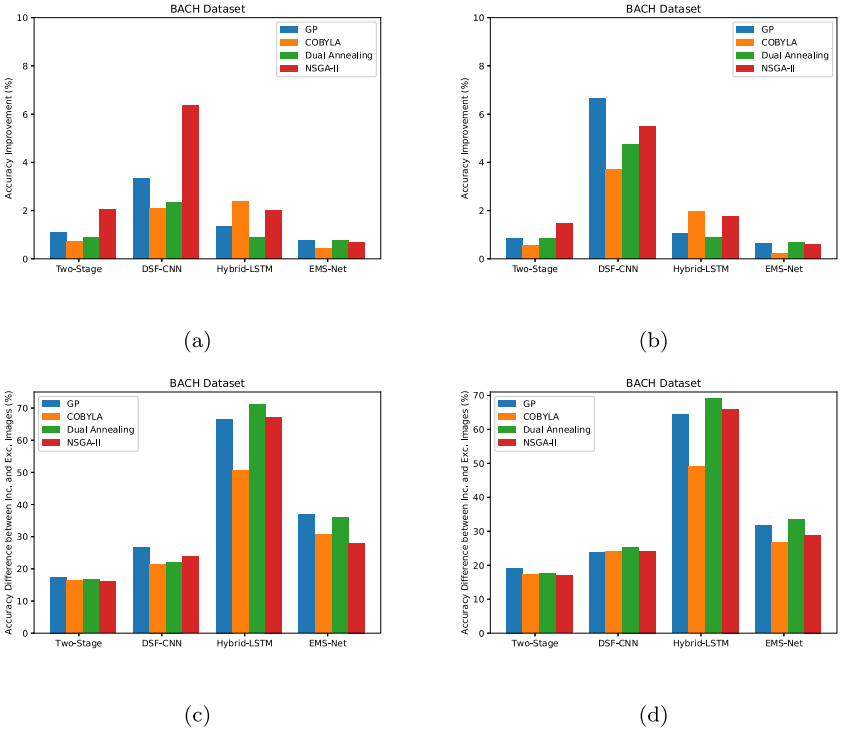

Fig. 2. Accuracy Improvement using AUQantO method and Accuracy Difference (AD) between included and excluded images for all deep learning architectures on BACH dataset using Shannon Entropy (a & c) and MC dropout (b & d) as uncertainty measures.

5.3.3. Performance of AUQantO method on skin dataset

Here, we describe the experimental study of the case studies conducted on the Skin dataset using Shannon Entropy and MC dropout as uncertainty quantification measures. Table 6 presents the average test performance (precision, recall, AUC, and accuracy) of the images included in the final classification. NSGA-II showed the highest average test accuracy among other optimization methods for the following 5 case studies: Two-stage (Entropy: 93.46% and MC: 89.54%), DSF-CNN (Entropy: 99.08% and MC: 96.39%) and Hybrid-LSTM using Entropy of accuracy equals to 96.67%. GP showed the highest level of accuracy for Hybrid LSTM using MC (95.44%). EMS-Net on both uncertainty measures have comparable records (varies between approximately 94% and almost 97%) among all optimization methods.

In terms of single architectures, DSF-CNN showed high accuracy with all optimization methods using the entropy method (95.58%, 93.88%, 94.12%, and 99.08% for GP, COBYLA, dual annealing, and NSGA-II, respectively). Also, DSF-CNN achieved higher accuracy for all optimization methods except for GP using MC dropout, where accuracy measures of 93.23%, 93.93%, and 96.39%

Table 6
Average test accuracy of included images using AUQantO method for case studies conducted on Skin dataset using Shannon Entropy and MC dropout as uncertainty measures.

Architecture (uncertainty measure)	Optimization method															
	GP				COBYLA				Dual annealing				NSGA-II			
	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC	Accuracy
Two-Stage (Entropy)	91.18%	91.01%	91.17%	91.01%	90.95%	90.79%	90.94%	90.79%	91.88%	91.75%	91.87%	91.75%	93.56%	93.46%	93.50%	93.46%
DSF-CNN (Entropy)	95.62%	95.58%	95.60%	95.58%	93.96%	93.88%	93.90%	93.88%	94.16%	94.12%	94.12%	94.12%	99.08%	99.08%	99.08%	99.08%
Hybrid-LSTM (Entropy)	93.83%	93.75%	93.81%	93.75%	93.60%	93.51%	93.60%	93.51%	93.44%	93.36%	93.42%	93.36%	96.70%	96.67%	96.66%	96.67%
EMS-Net (Entropy)	96.28%	96.25%	96.28%	96.25%	96.71%	96.69%	96.72%	96.69%	95.63%	95.60%	95.63%	95.60%	96.71%	96.69%	96.72%	96.69%
Two-Stage (MC)	88.19%	88.07%	88.11%	88.07%	88.03%	87.92%	87.89%	87.92%	89.48%	89.34%	89.44%	89.34%	89.66%	89.54%	89.54%	89.54%
DSF-CNN (MC)	94.79%	94.73%	94.71%	94.73%	93.29%	93.23%	93.20%	93.23%	93.99%	93.93%	93.90%	93.93%	96.39%	96.39%	96.23%	96.39%
Hybrid-LSTM (MC)	95.49%	95.44%	95.48%	95.44%	93.23%	93.14%	93.19%	93.14%	93.68%	93.63%	93.65%	93.63%	94.65%	94.62%	94.60%	94.62%
EMS-Net (MC)	94.85%	94.82%	94.84%	94.82%	94.85%	94.82%	94.84%	94.82%	94.85%	94.82%	94.84%	94.82%	94.85%	94.82%	94.84%	94.82%

Table 7
Average test accuracy of excluded images and (abstain percentage of dataset images) using AUQantO method for case studies conducted on SKIN dataset.

Architecture (uncertainty measure)	Optimization method			
	GP	COBYLA	Dual annealing	NSGA-II
Two-Stage (Entropy)	56.10% (20.38%)	56.30% (19.99%)	57.37% (22.84%)	63.46% (31.88%)
DSF-CNN (Entropy)	68.90% (20.38%)	60.54% (11.22%)	57.42% (10.83%)	77.16% (40.79%)
Hybrid-LSTM (Entropy)	57.27% (10.22%)	57.76% (9.77%)	56.95% (9.16%)	68.26% (23.42%)
EMS-Net (Entropy)	65.27% (15.89%)	66.12% (17.65%)	64.77% (13.86%)	66.14% (17.65%)
Two-Stage (MC)	61.40% (13.83%)	62.36% (13.86%)	61.97% (18.11%)	61.23% (18.23%)
DSF-CNN (MC)	63.80% (14.83%)	59.46% (9.13%)	60.95% (11.50%)	73.03% (26.66%)
Hybrid-LSTM (MC)	64.30% (17.50%)	56.22% (8.52%)	57.40% (10.04%)	61.09% (13.80%)
EMS-Net (MC)	66.14% (11.65%)	66.14% (11.65%)	66.14% (11.65%)	66.14% (11.65%)

were reported by COBYLA, dual annealing, and NSGA-II, respectively. For GP using MC dropout, Hybrid-LSTM showed an accuracy of 95.44% which is slightly higher than the one reported by DSF-CNN. From Table 6, we can conclude that NSGA-II of multi-objective optimization showed the highest performance levels for almost all case studies. Similarly to the evaluation conducted on the BACH dataset, we introduce in Table 7 the performance of the excluded images by *AUQantO* from the Skin dataset to link between the included and excluded images.

Table 7 demonstrates the performance of our method on excluded images and the associated abstain percentage of datasets for all case studies on the skin dataset. For entropy as an uncertainty measure, the two-stage and hybrid-LSTM interchangeably showed the lowest excluded image accuracy for all optimization methods with an excluded image accuracy of around 56% to 68%. For MC dropout as an uncertainty measure, Hybrid-LSTM showed the lowest excluded image accuracy for all optimization methods except GP, where excluded image accuracy measures of 56.22%, 57.40%, and 61.09% have been achieved by COBYLA, dual annealing, and NSGA-II, respectively. For GP on the skin data set, DSF-CNN showed the lowest excluded image accuracy of 63.80%. Moreover, generally, the least abstain percentage was obtained by Hybrid-LSTM on the skin dataset.

Fig. 3 confirms the effectiveness of our method on Skin dataset, where the accuracy improved in all case studies (Figs. 3(a) and (b)). The accuracy improvement for the skin dataset using the two uncertainty measures varies between almost 4% to 10% for all deep learning architectures and all optimization methods. Figs. 3(c) and (d) show the accuracy difference (AD) between included and excluded images for Shannon Entropy and MC dropout uncertainty measures, respectively. The obtained accuracy differences among all deep learning architectures and all optimization methods have comparable records with accuracy difference varies between 22% and 38% (Figs. 3(c) and (d)).

5.3.4. Statistical significance measurement of AUQantO

To confirm the effectiveness of *AUQantO*, we utilized Paired t-test and Wilcoxon Signed-Rank as statistical significance measurements to test whether there is a significant difference in test results before and after using *AUQantO*. Paired t-test measures the significant difference between two populations when the distribution of the differences between the two samples accounts for non-normality. Wilcoxon is the non-parametric version of paired t-test. Table 8 presents the *p*-value on the four metrics used in our study (precision, recall, AUC, and accuracy). It can be seen from the table that all reported p-values are less than 0.05 which means that we reject the null hypothesis. This reflects that the true mean of test results is not equivalent between the two populations (Results before and after *AUQantO*). In other words, *AUQantO* improves the effectiveness of the underlying model with statistical significance.

5.3.5. AUQantO against literature methods

In this section, we compare our *AUQantO* method against literature methods using the two main datasets applied in this study (BACH and skin datasets). For BACH dataset, we compared our method against the following: (1) DCNN + SVM model [47] which uses a pre-trained DCNN for contextual feature extraction and SVM for classification and (2) InceptionV3 + Ensemble model [48] which uses InceptionV3 as a feature extractor where these extracted features are then passed to a second layer of ensemble prediction fusion using gradient boosting machine, logistic regression, and support vector machine (SVM) to refine predictions. For the skin dataset, we compared our method against the following: (1) DIMLP-ensemble [49] which uses the CNN-based two-step rule extraction technique for two CNN-based subnets to improve CNN interpretability, and (2) TWDBDL [50] which introduces a Bayesian deep learning strategy based on three-way decision using uncertainty quantification for the classification of skin cancer images. As noticed from Table 9, our method surpasses the other literature methods confirming the effectiveness of our approach.

5.3.6. AUQantO on other datasets

To confirm the experimental study conducted above on BACH and skin datasets, we selected one of the single architectures (DSF-CNN [3]) to be the backbone deep learning architecture for *AUQantO* on two more datasets. DSF-CNN showed high performance on different scenarios done for *AUQantO*.

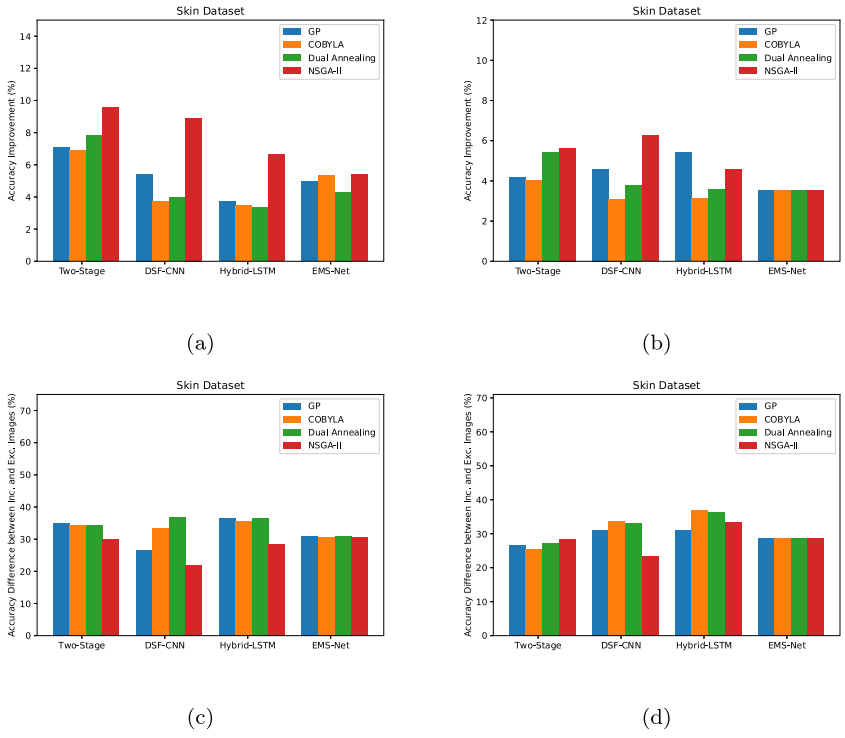

Fig. 3. Accuracy Improvement using AUQantO method and accuracy difference (AD) between included and excluded images for all deep learning architectures on Skin dataset using Shannon Entropy (a & c) and MC dropout (b & d) as uncertainty measures.

Table 8
Paired t-test and Wilcoxon Signed-Rank as statistical significance measurement on results before and after AUQantO method using BACH and Skin datasets.

P-value on metric	BACH		Skin	
	Paired t-test	Wilcoxon Signed-Rank	Paired t-test	Wilcoxon Signed-Rank
P-value on precision	3.75×10^{-8}	7.95×10^{-7}	9.70×10^{-17}	7.90×10^{-7}
P-value on recall	3.74×10^{-8}	7.94×10^{-7}	6.23×10^{-17}	7.90×10^{-7}
P-value on AUC	9.60×10^{-8}	2.32×10^{-6}	1.14×10^{-16}	7.896×10^{-7}
P-value on accuracy	3.74×10^{-8}	7.94×10^{-7}	6.23×10^{-17}	7.90×10^{-7}

Table 9
Comparison against literature methods.

Dataset	Method	Accuracy
BACH	DCNN + SVM [47]	90.00%
	InceptionV3 + Ensemble [48]	87.50%
	***AUQantO* DSF-CNN using Entropy (NSGA-II)**	**97.61%**
	***AUQantO* DSF-CNN using MC Dropout (GP)**	**97.92%**
Skin	DIMLP-ensemble [49]	84.90%
	TWDBDL [50]	88.95%
	***AUQantO* DSF-CNN using Entropy (NSGA-II)**	**99.08%**
	***AUQantO* DSF-CNN using MC Dropout (NSGA-II)**	**96.39%**

Table 10
Hyperparameter settings used for AUQantO (DSF-CNN).

Hyperparameter	Value
Patch Stride (Patch-wise Network)	56
Patch Stride (Image-wise Network)	112
Optimizer	Adam [13]
Learning Rate	0.0001
Batch Size	32
Training Epochs (Patch-wise Network)	4
Training Epochs (Image-wise Network)	6
λ Range	1×10^{-9} to 2
Evaluation Step (Single-Objective)	50
Initial Search Point (COBYLA)	0.01
Number of Variables (NSGA-II)	1
Number of Objectives (NSGA-II)	2
Number of Generations (NSGA-II)	50
Population Size (NSGA-II)	1
MC Dropout Test Runs	50

First, we used Brain tumor Magnetic Resonance (MR) image dataset[3] which comprises of 3264 images containing three different types of brain tumor images (glioma tumor, meningioma tumor, and pituitary tumor) and normal images. The dataset images are distributed as 500 images for normal, 926 for glioma, 937 for meningioma, and 901 for pituitary. Second, we used Chest Computed Tomography (CT) scan cancer dataset[4] which is divided into 4 classes: normal (215 images), adenocarcinoma (338 images), large cell carcinoma (187 images), and squamous cell carcinoma (260 images) giving us a total number of 1000 chest scan images.

For evaluating DSF-CNN on brain and chest datasets, we first resized the images from the chest and brain dataset to the size of 224 × 224 due to the variable image sizes of the original samples. Then we applied data augmentation by rotating the training patches by 90 degrees with horizontal and vertical flipping. Table 10 shows our hyperparameter settings that have been used by *AUQantO*.

3 https://www.kaggle.com/sartajbhuvaji/brain-tumor-classification-mri
4 https://www.kaggle.com/mohamedhanyyy/chest-ctscan-images

Table 11
AUQantO Performance with backbone DSF-CNN on Brain and Chest datasets..

Dataset	Optimization methods	Entropy						MC Dropout					
		Before AUQantO	After AUQantO	Acc. improvement	Excluded images acc.	AD	AP	Before AUQantO	After AUQantO	Acc. improvement	Excluded images acc.	AD	AP
Brain	GP	95.77%	96.72%	0.95%	50.00%	46.72%	2.02%	95.77%	97.02%	1.25%	60.71%	36.31%	4.29%
	COBYLA	95.77%	96.58%	0.81%	44.00%	52.58%	1.53%	95.77%	96.76%	0.99%	60.34%	36.42%	3.55%
	Dual annealing	95.77%	96.57%	0.80%	50.00%	46.57%	1.72%	95.77%	96.92%	1.15%	63.37%	33.55%	4.35%
	NSGA-II	95.77%	98.84%	3.07%	67.50%	31.34%	9.80%	95.77%	97.74%	1.97%	72.99%	24.75%	9.19%
Chest	GP	84.23%	86.43%	2.20%	22.86%	63.57%	3.50%	84.23%	85.10%	0.87%	44.44%	40.66%	2.70%
	COBYLA	84.23%	87.34%	3.11%	15.91%	71.43%	4.40%	84.23%	84.39%	0.16%	28.57%	55.82%	0.70%
	Dual annealing	84.23%	86.28%	2.05%	31.58%	54.70%	3.80%	84.23%	84.90%	0.67%	74.41%	10.49%	8.60%
	NSGA-II	84.23%	95.78%	11.55%	36.41%	59.37%	19.50%	84.23%	89.84%	5.61%	54.26%	35.58%	16.40%

Table 11 presents the results of applying our *AUQantO* method using Shannon Entropy and MC Dropout as uncertainty measures and the DSF-CNN deep learning architecture as the backbone network on Brain and Chest datasets, respectively. The table shows the accuracy performance before and after using *AUQantO*, the improvement in accuracy, the accuracy of the excluded images, the difference in accuracy between the accuracy of the included images and the accuracy of the excluded images, and the percentage of abstaining from the dataset. It can be seen from the table that our method is highly efficient in improving performance.

6. Conclusion and future work

In this paper, we introduce a model and a data-agnostic method, which we call Actionable Uncertainty Quantification Optimization (*AUQantO*) for optimizing uncertainty quantification in deep learning architectures. *AUQantO* can measure the uncertainty level in medical images and exclude poor samples based on a hyperparameter (e.g., threshold) that is optimized using single and multi-objective optimization methods. We validated and evaluated the performance of our method in several different case studies using two commonly used uncertainty measures. Moreover, we validated the performance of the best performing version of *AUQantO* on two other datasets. Experimental results showed a favorable performance in the exclusion of highly uncertain images, confirming its automated actionability with different deep learning architectures.

Future research directions include the trial of other optimization methods and the aid of action taking through visual explanation of the classified image. Furthermore, multiple uncertainty quantification methods are planned to be combined and incorporated to enrich the framework and introduce better sample exclusion. Also, the adoption of machine teaching is planned by the authors.

CRediT authorship contribution statement

Zakaria Senousy: Methodology, Software, Validation, Visualisation, Writing – original draft. **Mohamed Medhat Gaber:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Mohammed M. Abdelsamea:** Conceptualization, Methodology, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgment

The authors would like to thank the anonymous reviewers for their invaluable feedback, which significantly contributed to enhancing the overall quality of the paper.

References

[1] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (1) (2017) 221–248, http://dx.doi.org/10.1146/annurev-bioeng-071516-044442, PMID: 28301734.
[2] K. Nazeri, A. Aminpour, M. Ebrahimi, Two-stage convolutional neural network for breast cancer histology image classification, in: Image Analysis and Recognition, Springer International Publishing, 2018, pp. 717–726.
[3] Y. Huang, A.C.-S. Chung, Improving high resolution histology image classification with deep spatial fusion network, in: Computational Pathology and Ophthalmic Medical Image Analysis, Springer International Publishing, 2018, pp. 19–26.
[4] Z. Yang, L. Ran, S. Zhang, Y. Xia, Y. Zhang, EMS-net: Ensemble of multiscale convolutional neural networks for classification of breast cancer histology images, Neurocomputing 366 (2019) http://dx.doi.org/10.1016/j.neucom.2019.07.080.
[5] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423, http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x.
[6] 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.
[7] F.C. Ghesu, B. Georgescu, E. Gibson, S. Guendel, M.K. Kalra, R. Singh, S.R. Digumarthy, S. Grbic, D. Comaniciu, Quantifying and leveraging classification uncertainty for chest radiograph assessment, in: D. Shen, T. Liu, T.M. Peters, L.H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan (Eds.), Medical Image Computing and Computer Assisted Intervention, MICCAI 2019, Springer International Publishing, Cham, 2019, pp. 676–684.
[8] F.C. Ghesu, B. Georgescu, A. Mansoor, Y. Yoo, E. Gibson, R. Vishwanath, A. Balachandran, J.M. Balter, Y. Cao, R. Singh, S.R. Digumarthy, M.K. Kalra, S. Grbic, D. Comaniciu, Quantifying and leveraging predictive uncertainty for medical image assessment, Med. Image Anal. 68 (2021) 101855.
[9] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.
[10] R. Yan, F. Ren, W. Zihao, L. Wang, T. Zhang, Y. Liu, X. Rao, C. Zheng, F. Zhang, Breast cancer histopathological image classification using a hybrid deep neural network, Methods 173 (2020) http://dx.doi.org/10.1016/j.ymeth.2019.06.014.
[11] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J.V. Dillon, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Curran Associates Inc. Red Hook, NY, USA, 2019.
[12] L. Bottou, F.E. Curtis, J. Nocedal, Large-scale machine learning with stochastic gradient descent, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2010, pp. 641–648.
[13] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2014.
[14] C.E. Rasmussen, Gaussian processes in machine learning, in: O. Bousquet, U. von Luxburg, G. Ratsch (Eds.), Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, TÜBingen, Germany, August 4 - 16, 2003, Revised Lectures, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 63–71.
[15] M.J.D. Powell, A direct search optimization method that models the objective and constraint functions by linear interpolation, in: S. Gomez, J.-P. Hennart (Eds.), Advances in Optimization and Numerical Analysis, Springer Netherlands, Dordrecht, 1994, pp. 51–67.

[16] Y. Xiang, D. Sun, W. Fan, X. Gong, Generalized simulated annealing algorithm and its application to the thomson model, Phys. Lett. A 233 (3) (1997) 216–220, http://dx.doi.org/10.1016/S0375-9601(97)00474-X.
[17] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Trans. Evol. Comput. 6 (2) (2002) 182–197, http://dx.doi.org/10.1109/4235.996017.
[18] Q. Nie, Y.-b. Zou, J.C.-W. Lin, Feature extraction for medical CT images of sports tear injury, Mob. Netw. Appl. 26 (1) (2021) 404–414.
[19] A. Belhadi, Y. Djenouri, V.G. Diaz, E.H. Houssein, J.C.-W. Lin, Hybrid intelligent framework for automated medical learning, Expert Syst. 39 (6) (2022) e12737.
[20] J.M.-T. Wu, G. Srivastava, J.C.-W. Lin, Q. Teng, A multi-threshold ant colony system-based sanitization model in shared medical environments, ACM Trans. Internet Technol. 21 (2) (2021) http://dx.doi.org/10.1145/3408296.
[21] Y. Djenouri, A. Belhadi, G. Srivastava, J.C.-W. Lin, Secure collaborative augmented reality framework for biomedical informatics, IEEE J. Biomed. Health Inf. 26 (6) (2022) 2417–2424, http://dx.doi.org/10.1109/JBHI.2021.3139575.
[22] B. Han, R.H. Jhaveri, H. Wang, D. Qiao, J. Du, Application of robust zero-watermarking scheme based on federated learning for securing the healthcare data, IEEE J. Biomed. Health Inf. 27 (2) (2023) 804–813, http://dx.doi.org/10.1109/JBHI.2021.3123936.
[23] R.K. Dhanaraj, R.H. Jhaveri, L. Krishnasamy, G. Srivastava, P.K.R. Maddikunta, Black-hole attack mitigation in medical sensor networks using the enhanced gravitational search algorithm, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 29 (Supp02) (2021) 297–315, http://dx.doi.org/10.1142/S021848852140016X.
[24] A. Khamparia, D. Gupta, V.H.C. de Albuquerque, A.K. Sangaiah, R.H. Jhaveri, Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning, J. Supercomput. 76 (11) (2020) 8590–8608.
[25] G. Dorta, S. Vicente, L. de Agapito, N.D.F. Campbell, I.J.A. Simpson, Structured uncertainty prediction networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5477–5485.
[26] J. Mukhoti, Y. Gal, Evaluating Bayesian deep learning methods for semantic segmentation, 2018, arXiv, arXiv:1811.12709.
[27] A. Harakeh, M. Smart, S. Waslander, BayesOD: A Bayesian approach for uncertainty estimation in deep object detectors, 2020, pp. 87–93, http://dx.doi.org/10.1109/ICRA40945.2020.9196544.
[28] R. Michelmore, M. Wicker, L. Laurenti, L. Cardelli, Y. Gal, M. Kwiatkowska, Uncertainty quantification with statistical guarantees in end-to-end autonomous driving control, in: 2020 IEEE International Conference on Robotics and Automation, ICRA, 2020, pp. 7344–7350, http://dx.doi.org/10.1109/ICRA40945.2020.9196844.
[29] I. Osband, Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout, in: Workshop on Bayesian Deep Learning, NIPS, 2016.
[30] F. Doshi-Velez, B. Kim, Towards A rigorous science of interpretable machine learning, 2017, arXiv: Machine Learning.
[31] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.
[32] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. Snead, Y.W. Tsang, N. Rajpoot, MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images, Med. Image Anal. 52 (2019) 199–211, http://dx.doi.org/10.1016/j.media.2018.12.001.
[33] M. Fraz, S.A. Khurram, S. Graham, M. Shaban, A. Loya, N. Rajpoot, FABnet: Feature attention-based network for simultaneous segmentation of microvessels and nerves in routine histology images of oral cancer, Neural Comput. Appl. 32 (2020) http://dx.doi.org/10.1007/s00521-019-04516-y.
[34] A. Mobiny, A. Singh, Risk-aware machine learning classifier for skin lesion diagnosis, J. Clin. Med. 8 (2019) 1241, http://dx.doi.org/10.3390/jcm8081241.
[35] L. Raczkowski, M. Mozejko, J. Zambonelli, E. Szczurek, ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning, Sci. Rep. 9 (2019) http://dx.doi.org/10.1038/s41598-019-50587-1.
[36] R. Camarasa, D. Bos, J. Hendrikse, P. Nederkoorn, E. Kooi, A. van der Lugt, M. de Bruijne, Quantitative comparison of Monte-Carlo dropout uncertainty measures for multi-class segmentation, in: C.H. Sudre, H. Fehri, T. Arbel, C.F. Baumgartner, A. Dalca, R. Tanno, K. Van Leemput, W.M. Wells, A. Sotiras, B. Papiez, E. Ferrante, S. Parisot (Eds.), Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis, Springer International Publishing, Cham, 2020, pp. 32–41.
[37] U. Ahmed, J.C.-W. Lin, Robust adversarial uncertainty quantification for deep learning fine-tuning, J. Supercomput. (2023).
[38] Z. Senousy, M.M. Abdelsamea, M.M. Mohamed, M.M. Gaber, 3E-net: Entropy-based elastic ensemble of deep convolutional neural networks for grading of invasive breast carcinoma histopathological microscopic images, Entropy 23 (5) (2021) http://dx.doi.org/10.3390/e23050620.
[39] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 2261–2269.
[40] Z. Senousy, M. Abdelsamea, M.M. Gaber, M. Abdar, R.U. Acharya, A. Khosravi, S. Nahavandi, MCUa: Multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification, IEEE Trans. Biomed. Eng. (2021) 1, http://dx.doi.org/10.1109/TBME.2021.3107446.
[41] M.M. Abdelsamea, U. Zidan, Z. Senousy, M.M. Gaber, E. Rakha, M. Ilyas, A survey on artificial intelligence in histopathology image analysis, WIREs Data Min. Knowl. Discov. e1474.
[42] H. Liu, Y. Ong, X. Shen, J. Cai, When Gaussian process meets big data: A review of scalable GPs, IEEE Trans. Neural Netw. Learn. Syst. 31 (2018) 4405–4423.
[43] L. Ingber, Simulated annealing: Practice versus theory, Math. Comput. Modelling 18 (11) (1993) 29–57.
[44] M. Abdel-Basset, L. Abdel-Fatah, A.K. Sangaiah, Chapter 10 - metaheuristic algorithms: A comprehensive review, in: A.K. Sangaiah, M. Sheng, Z. Zhang (Eds.), Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications, in: Intelligent Data-Centric Systems, Academic Press, 2018, pp. 185–231.
[45] G. Aresta, T. Araújo, S. Kwok, S.S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q.D. Vu, M.N.N. To, E. Kim, J.T. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun, K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia, P. Aguiar, BACH: Grand challenge on breast cancer histology images, Med. Image Anal. 56 (2019) 122–139, http://dx.doi.org/10.1016/j.media.2019.05.010.
[46] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, H.P. Soyer, A patient-centric dataset of images and metadata for identifying melanomas using clinical context, Sci. Data 8 (1) (2021) 34, http://dx.doi.org/10.1038/s41597-021-00815-z.
[47] R. Awan, N.A. Koohbanani, M. Shaban, A. Lisowska, N. Rajpoot, Context-aware learning using transferable features for classification of breast cancer histology images, in: Image Analysis and Recognition, Springer International Publishing, 2018, pp. 788–795.
[48] Y.S. Vang, Z. Chen, X. Xie, Deep learning framework for multi-class breast cancer histology image classification, in: A. Campilho, F. Karray, B. ter Haar Romeny (Eds.), Image Analysis and Recognition, Springer International Publishing, Cham, 2018, pp. 914–922.
[49] G. Bologna, S. Fossati, A two-step rule-extraction technique for a CNN, Electronics 9 (6) (2020) http://dx.doi.org/10.3390/electronics9060990.
[50] M. Abdar, M. Samami, S. Dehghani Mahmoodabad, T. Doan, B. Mazoure, R. Hashemifesharaki, L. Liu, A. Khosravi, U.R. Acharya, V. Makarenkov, S. Nahavandi, Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning, Comput. Biol. Med. 135 (2021) 104418.