Heliyon 10 (2024) e24188

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Leveraging Bayesian deep learning and ensemble methods for uncertainty quantification in image classification: A ranking-based approach

Abdullah A. Abdullah [a,*], Masoud M. Hassan [a], Yaseen T. Mustafa [b]

[a] Computer Science Department, Faculty of Science, University of Zakho, Duhok, Iraq
[b] Environmental Science Department, Faculty of Science, University of Zakho, Duhok, Iraq

ARTICLE INFO

Keywords:
Bayesian deep learning
Uncertainty quantification
Ensemble learning
Ranking-based models
Image classification

ABSTRACT

Bayesian deep learning (BDL) has emerged as a powerful technique for quantifying uncertainty in classification tasks, surpassing the effectiveness of traditional models by aligning with the probabilistic nature of real-world data. This alignment allows for informed decision-making by not only identifying the most likely outcome but also quantifying the surrounding uncertainty. Such capabilities hold great significance in fields like medical diagnoses and autonomous driving, where the consequences of misclassification are substantial. To further improve uncertainty quantification, the research community has introduced Bayesian model ensembles, which combines multiple Bayesian models to enhance predictive accuracy and uncertainty quantification. These ensembles have exhibited superior performance compared to individual Bayesian models and even non-Bayesian counterparts. In this study, we propose a novel approach that leverages the power of Bayesian ensembles for enhanced uncertainty quantification. The proposed method exploits the disparity between predicted positive and negative classes and employs it as a ranking metric for model selection. For each instance or sample, the ensemble's output for each class is determined by selecting the top 'k' models based on this ranking. Experimental results on different medical image classifications demonstrate that the proposed method consistently outperforms or achieves comparable performance to conventional Bayesian ensemble. This investigation highlights the practical application of Bayesian ensemble techniques in refining predictive performance and enhancing uncertainty evaluation in image classification tasks.

1. Introduction

In recent years, machine learning in general and deep learning, in particular, have made remarkable advancements in image processing tasks, such as classification, recognition, and segmentation. Deep learning has proven to be a powerful tool for many applications across various fields. In classification, accurate classification of images is essential for decision-making, which is much easier with the availability of state-of-the-art models and tools. However, another equally critical aspect is the ability to quantify the uncertainty associated with these predictions [1,2]. Uncertainty quantification (UQ) plays a crucial role in situations in which confident and reliable predictions are essential, such as medical diagnosis [2], autonomous systems [3], physics [4–6], and material

[*] Corresponding author.
E-mail address: Abdullah.abdullah@uoz.edu.krd (A.A. Abdullah).

https://doi.org/10.1016/j.heliyon.2024.e24188
Received 31 August 2023; Received in revised form 8 December 2023; Accepted 4 January 2024
Available online 8 January 2024

science [7–10]. Even though traditional deep learning methods can perform well and even achieve state-of-the-art results, they often fail to determine the uncertainties associated with models and data. Without explicit UQ, the model cannot provide sufficient insight into whether the result can be reliable, and their lack in ability to convey prediction confidence will make it difficult for decision-making processes, particularly in decision-sensitive fields such as healthcare [2,11–13]. Several methods, including Bayesian methods and ensemble learning, have been employed and developed to address these shortcomings. Despite the fact that these methods have been around for years; they are underemployed compared to the number of published works in the field.

Bayesian deep learning (BDL) is one of the successful approaches that incorporate Bayesian concepts into neural networks to quantify uncertainty effectively. Traditional deep learning methods often provide point estimates without explicitly accounting for uncertainty, whereas Bayesian deep learning models utilize a posterior probability distribution that relies on prior knowledge distribution and the likelihood of the data being utilized [14]. This Bayesian framework enables uncertainty estimation by representing model weights as random variables. Prior knowledge incorporated into BDL can take different statistical distributions; yet, the most popular deep learning models are Bernoulli/binomial and Gaussian distributions for computer vision tasks [15]. The posterior distribution of BDL is used for prediction and UQ and can be generated by using the prior and likelihood to sample from the posterior distribution. Sampling can be divided into two types: exact and approximate sampling methods. Markov Chain Monte Carlo (MCMC) [16] is one example of an exact sampling method which is the most popular method used to sample from the posterior distribution. However, it has not gained popularity in deep learning models because it is computationally expensive and difficult to scale up for large datasets or a large number of features such as image data [2]. On the other hand, the popular approximate methods for sampling are the Monte Carlo Dropout (MC-Dropout) [14] and Variational Inference (VI) [17]. The MC-Dropout method samples from the Bernoulli/binomial distribution whereas the VI method samples from the Gaussian (normal) distribution as an estimate posterior distribution. These two approximate methods can be scaled up since they have fewer parameters compared to the MCMC [15]. Although the BDL can provide information about the model's uncertainties, it can have a negative impact on the model's performance most of the time particularly in classification tasks [18].

Another approach used for uncertainty quantification is ensemble methods [19], which can be particularly helpful when the noise ratio in the data is high or the model is complex. Uncertainty quantification in ensemble methods can be performed by combining various models or through fitting one model with diverse hyperparameters and capturing the uncertainty associated with different sources, such as training data, model architectures, or alternative initializations. Several techniques can be used to achieve that such as boosting, bagging, and stacking [1]. Bagging is a technique that involves training multiple models independently on different subsets of the training data. Each model is then applied to the test data to make a prediction, and the final prediction is obtained by averaging the outputs of all models. Boosting involves training models iteratively with the most difficult samples to classify. The idea is to combine several weak models into a single strong model. Stacking involves training multiple models, but instead of just averaging their outputs, it uses a method/model that learns how to integrate the underlying predictions of the base models. Ensemble for uncertainty quantification has been used in many applications, some examples of such applications are computer vision [1] spatiotemporal forecasting [20] mechanical machinery [21]. Published studies have demonstrated the ability and effectiveness of ensemble methods to quantify model uncertainty.

While both Bayesian deep learning models and ensemble methods for uncertainty quantification have their advantages, each has its limitations as well. Bayesian deep learning generates probabilistic predictions, which provide a more comprehensive view of possible outcomes, rather than just a single-point prediction. However, this method is applied on a single model only. By combining predictions from multiple models, ensemble methods can prevent overfitting and improve generalization [22]. However, they cannot provide comprehensive outcomes and uncertainty quantification for each model. To overcome the limitations of both methods, combining the two methodologies can help improve capturing the uncertainty associated with each model; a typical approach to achieve this is through Bayesian model averaging (BMA) [23,24]. The BMA combines the predictions of individual Bayesian models, each representing a different uncertainty, by weighting them according to their posterior probabilities. This process results in an ensemble prediction that represents the uncertainty of combined models and provides a more robust estimate of the posterior distribution. This method can provide a good understanding of the uncertainty of each model also providing more generalized results simultaneously. However, BMA considers all models, even if one model performs poorly in particular instances of the data.

Motivated by the need for reliable uncertainty quantification and improved accuracy of predictive models, this study presents a methodology that leverages Bayesian deep learning and ensemble methods. To achieve this, the proposed method used in this study is based on the ranking of uncertainty/confidence. The method picks only models with low uncertainty (high confidence) samples and averages these selected samples to sample from the posterior distribution. This can alleviate the issue of a model's prediction when it performs poorly only on a particular type of input, which lead to increase the performance.

The rest of this paper is organized as follows: In Section 2, we provide a comprehensive review of related literature in the fields of uncertainty quantification, Bayesian deep learning, and ensemble methods. Section 3 details the proposed methodology, explaining the architecture and training procedures for Bayesian deep learning models, and the ensembling approach. Section 4 presents the experimental setup and evaluation metrics as well as the results and discussion of the findings. Finally, Section 5 concludes the study with a summary of its contributions.

2. Related work

Bayesian deep learning for uncertainty quantification has attracted the attention of many researchers in recent years. BDL is particularly important in healthcare, where it has been applied to various medical imaging classification tasks, including cancer image classification, MRI image classification, histopathology image classification, X-ray image classification, and others. For example, Song et al. [11] proposed a Bayesian deep learning method for reliable oral cancer image classification using Monte Carlo dropout (MC-Dropout). Yadav [12] proposed a Bayesian deep learning-based convolutional neural network for Parkinson's disease classification using functional magnetic resonance images. Their architecture combines a 3D convolutional neural network with a Bayesian network based on the LeNet-5 model. Meanwhile, Thiagarajan et al. [25] used Bayesian neural network classifiers to derive uncertainty estimates for breast histopathology image classification using variational inference to approximate the posterior distribution. In the context of COVID-19 diagnosis, Gour and Jain [26] developed an uncertainty-aware convolutional neural network for COVID-19 X-ray image classification using MC-Dropout to estimate the uncertainty of the model. Subramanian et al. [27] proposed a Bayesian optimization deep learning network to diagnose retinal disease through optical coherence tomography images to optimize the hyperparameters of the network. Additionally, Loey et al. [28] proposed a Bayesian-based optimized deep learning model for COVID-19 patient detection using chest X-ray image data. Variational Inference was used to estimate the posterior distribution. Addressing skin cancer classification, Abdar et al. [13] proposed a three-way decision-based Bayesian deep learning method for uncertainty quantification. They used MC-Dropout to estimate model uncertainty. Although these methods have shown promising results, several challenges remain to be tackled. These include the necessity for diverse and unbiased datasets, potential data biases, and the difficulty surrounding the interpretation of the model's uncertainty estimates.

Over the years, various ensemble-based methods have been proposed for quantifying uncertainty. Althoff et al. [21] utilized dropout neural networks as an ensemble method for hydrological models. Although the proposed approach demonstrated efficiency, the choice of dropout rate was observed to affect the accuracy of the uncertainty estimates. Egele et al. [29] proposed an automated deep ensemble framework named Autodeuq, designed to simplify ensemble training for non-experts. However, they did not validate the reliability of the produced uncertainties. Hoffmann et al. [30] employed ensemble learning to quantify uncertainty in optical measurements; yet, they did not compare their results to Bayesian techniques for evaluating uncertainty. In summary, although ensemble methods exhibit promising potential for uncertainty quantification, certain challenges remain. These challenges encompass the fine-tuning of ensemble hyperparameters, the accurate assessment of uncertainty, and the comparative analysis with Bayesian approaches.

A combination of ensemble models with Bayesian deep learning offers a solution to overcome the limitations inherent in both methods for uncertainty quantification. For this purpose, Pearce et al. [31] proposed using ensembles of neural networks to quantify the predictive uncertainty, revealing that an ensemble composed of Bayesian models effectively captures both aleatoric and epistemic uncertainties. Employing a neural tangent kernel approach, He et al. [32] developed Bayesian deep learning models that enable efficient posterior weights estimation and uncertainty quantification. However, their method focused only on epistemic uncertainty. In the context of intrusion detection, Zhang et al. [33] used an ensemble of Bayesian neural networks, achieving enhanced uncertainty quantification than a single Bayesian neural network. However, their proposed method required tuning of many hyperparameters. Li

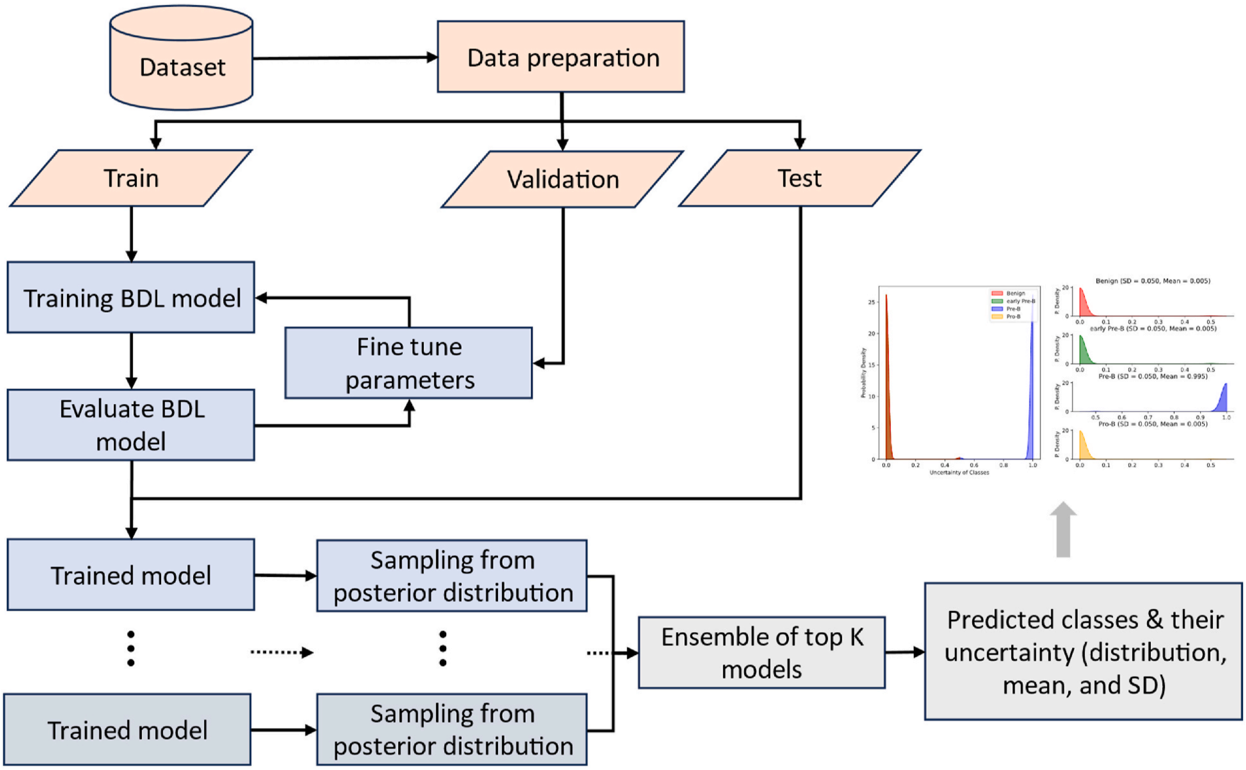
Dataset
Data preparation
Train
Validation
Test
Training BDL model
Fine tune parameters
Evaluate BDL model
Trained model
Sampling from posterior distribution
Trained model
Sampling from posterior distribution
Ensemble of top K models
Predicted classes & their uncertainty (distribution, mean, and SD)
Benign
early Pre-B
Pre-B
Pro-B
Probability Density
Uncertainty of Classes
Benign (SD = 0.050, Mean = 0.005)
early Pre-B (SD = 0.050, Mean = 0.005)
Pre-B (SD = 0.050, Mean = 0.995)
Pro-B (SD = 0.050, Mean = 0.005)
P. Density

Fig. 1. Flowchart of the utilized method for uncertainty quantification.

et al. [34] developed a variational Bayesian deep-learning approach for hydrological modeling that efficiently estimates model uncertainties. However, their method's assumption of Gaussian approximation for posterior distributions may not always be applicable. Delving into the domain of distribution shift, Seligmann et al. [35] performed a large-scale evaluation of Bayesian deep learning models. They found that Bayesian ensembles performed best for uncertainty quantification under a distribution shift. In most cases, when ensemble methods and Bayesian deep learning are utilized, the BMA is used, or take the predictions of all models are predicted.

In most instances where ensemble methods and Bayesian deep learning are utilized, they often involve Bayesian Model Averaging (BMA) or the aggregation of predictions from all models. The objective of this work is to integrate both ensemble and Bayesian models, retaining only those models with low uncertainty (high confidence) prediction. Subsequently, these selected predictions are averaged to yield samples from the posterior distribution. This can alleviate the issues related to poor model performance on particular types of inputs, potentially leading to enhanced overall performance.

3. Methodology

Bayesian ensemble deep learning is utilized in this study to quantify the uncertainty of the model outputs. The Bayesian method for deep learning is an excellent tool for uncertainty quantification; however, Bayesian models for deep learning cannot or marginally improve the model performance and can reduce the performance in many cases for image classification tasks. One way to improve performance in such cases is to utilize ensembles for Bayesian deep learning models. The goal of this work is to leverage Bayesian deep learning and ensemble methods to quantify uncertainty using Bayesian model averaging. Unlike the methods that consider all models for averaging, the proposed method only considers those models with the lowest uncertainty associated with them. This can be performed by ranking and selecting the top K model with the lowest uncertainty for the BMA. Details of the proposed method are presented in the following sections. Fig. 1 shows the flowchart of the proposed method for uncertainty quantification.

3.1. Datasets

The performance of the method proposed in this study was evaluated using two small-image classification datasets. The first dataset is called acute lymphoblastic leukemia (ALL) which has 3 K images with a low noise level and four different classes. The second dataset is a breast cancer dataset with grayscale images of breast cancer and has three classes.

3.1.1. Acute lymphoblastic leukemia (ALL)

The first dataset utilized in this study was the ALL image dataset collected by Mehrad et al. [36]. ALL refers to a malignant neoplasm cancer that affects the bone marrow and the blood. The bone marrow, which resides within bones, is responsible for producing blood cells. In ALL, immature white blood cells called lymphoblasts are excessively produced and fail to function properly, outnumbering normal blood cells. Consequently, this condition can lead to anemia, infection, and other severe health complications [37]. Leukemia, the broader category of ALL, can be classified into two main types: acute and chronic. Chronic leukemia is not very aggressive compared to the acute type, develops more slowly, and may not require immediate intervention [38]. In contrast, acute leukemia is more aggressive and necessitates immediate treatment, which is marked by the rapid production of underdeveloped blood cells. ALL, classified as acute leukemia, affects lymphoid cells, which are a particular variety of white blood cells responsible for enhancing the immune system in the human body. Children are more susceptible to ALL than adults and are characterized by rapid overproduction of immature lymphoblasts or lymphocytes. Immediate medical intervention is essential to prevent life-threatening consequences [39].

In this study, a dataset comprising Peripheral Blood Smear (PBS) images was used to diagnose ALL. Obtained from the bone marrow laboratory of Taleqani Hospital in Tehran, Iran, the dataset utilized ALL images from 89 individuals suspected of having ALL and collected 3256 PBS images [39]. Within the dataset, 25 individuals had benign conditions, whereas 64 individuals tested positive for a particular malignant lymphoblast type. Mehrad et al. [36] used this dataset to classify images into four categories: benign, early pre-B, pre-B, and pro-B. Some samples of these four classes are depicted in Fig. 2 ((A) Benign, (B) early Pre-B, (C) Pre-B, and (D) Pro-B). These images, which were captured using a Zeiss camera at 100× magnification, were improved and saved as JPG files. Despite its relatively small size, this dataset allowed a high level of accuracy to be achieved, which motivated its use in the present study.

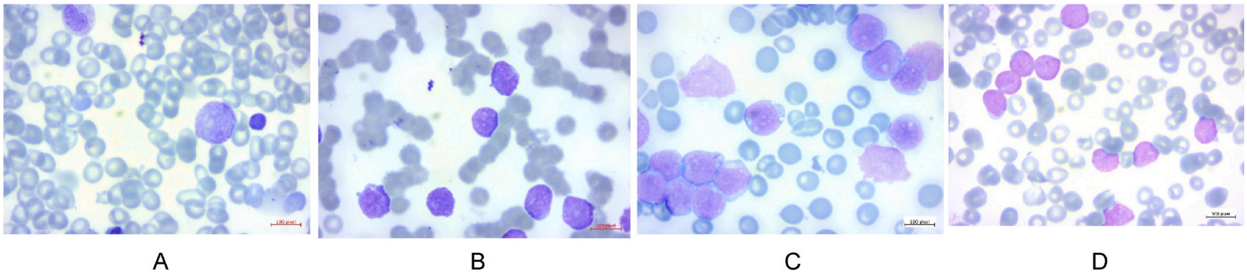

Fig. 2. Samples from the ALL dataset: (A) Benign, (B) early Pre-B, (C) Pre-B, and (D) Pro-B.

3.1.2. *Breast cancer dataset*

Breast cancer is a type of malignancy that originates in the breast tissue and affects both women and men, although it is more commonly found in women. The disease is characterized by uncontrolled growth of abnormal cells within the breast tissue, leading to tumor formation. While some tumors are benign and lack the ability to spread, others are malignant and have the potential to metastasize to other parts of the body [40]. One of the most popular methods to diagnose breast cancer is by imaging techniques such as mammography or ultrasound [41]. The key to improved breast cancer survival rates and recovery is early detection and prompt medical attention [42].

Breast cancer detection and diagnosis have undergone significant advancements with the aid of medical image datasets. Al-Dhabyani et al. [41] contributed to this progress by introducing a Breast Ultrasound Dataset in 2018. This collection is specifically designed for breast cancer analysis using ultrasound scans and includes a diverse range of cases divided into three classes: normal, benign, and malignant. The dataset consisted of 780 grayscale breast ultrasound images gathered from 600 female subjects aged between 25 and 75 years old. A few samples from the breast cancer dataset are depicted in Fig. 3 ((A) Normal, (B) Benign, and (C) Malignant). The images had dimensions of approximately 500 × 500 pixels on average. The dataset was chosen for this study because of its relatively compact size and grayscale format, which facilitate classification tasks. The authors demonstrated the effectiveness of the dataset by employing machine-learning algorithms for breast cancer classification, detection, and segmentation.

3.2. *Data preparation*

In this study, we optimized the performance of our deep learning models through a series of steps during the data preparation. First, the images were resized to 128 × 128 pixels. This adjustment not only reduced computational demands but also maintained adequate resolution, allowing for accurate image analysis. Furthermore, addressing the challenge of class imbalance is crucial for achieving good classification performance of minority classes. This was accomplished by utilizing class weighting, in which each class was assigned a different weight. The objective was to test the performance of the ensemble model of Bayesian deep learning models and the uncertainty associated with them rather than performing state-of-the-art results. Therefore, this study did not utilize any type of data augmentation, resampling, or other techniques to improve accuracy. Although these methods might have enhanced the accuracy, this work intentionally kept things simple to stay focused on the intended goals to be achieved. In the utilized model setup, the data were split into three subsets: 20 % allocated to the test set, 10 % assigned to the validation set, and the remaining 70 % dedicated to the training set. Finally, all data were normalized to ensure that the feature scales were similar for all the utilized models and data. This step is important for the generalizability and convergence of the model.

3.3. *Bayesian method for uncertainty*

Bayesian methods in deep learning have gained significant attention owing to their ability to quantify uncertainty in predictive models. Bayesian methods estimate uncertainty by treating model parameters (weights) as probability distributions rather than as fixed values. BDL offers advantages over conventional deep-learning models for image classification tasks by quantifying uncertainty, improving robustness, enabling transfer learning, improving model calibration, and supporting generalization. These advantages make BDL an appealing approach for different applications that require accurate predictions and reliable uncertainty estimates.

BDL relies on Bayes' theorem, which is a fundamental principle in probability theory. The Bayes' theorem equation [15] can be written as follows:

$$P(H|E) = \frac{P(E|H)\ \ P(H)}{P(E)} \tag{1}$$

The term $P(H|E)$ represents the posterior probability of the hypothesis when the evidence is confirmed to be true. The term $P(E|H)$ which is called likelihood, which indicates the probability of evidence occurring under the assumption that the hypothesis is valid. The

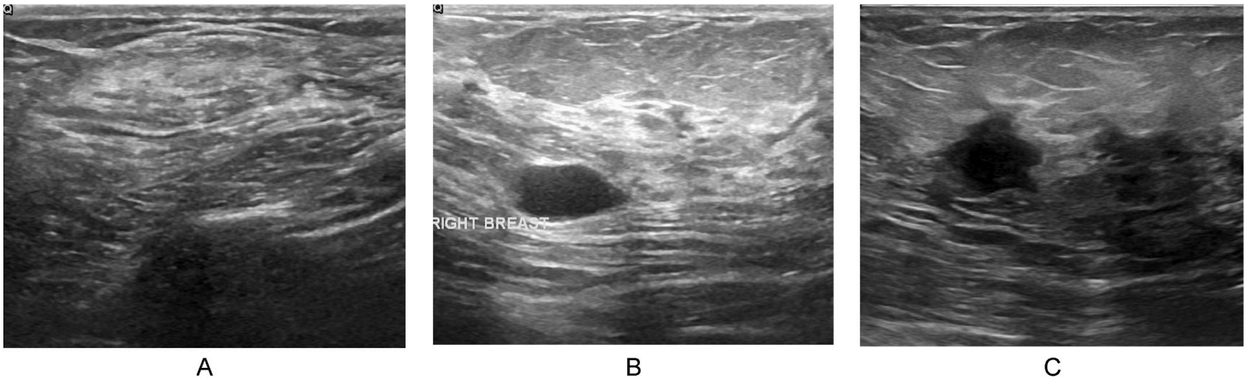

Fig. 3. Samples from breast cancer dataset: (A) Normal, (B) Benign, and (C) Malignant.

term $P(H)$ refers to the prior assumption or probability of the hypothesis. Finally, the term $P(E)$ refers to the probability that the evidence is true.

In the context of BDL, equation (1) can be rewritten as:

$$p(\omega|D) = \frac{p(D|\omega)\ \ p(\omega)}{p(D)} \tag{2}$$

Where $p(\omega|D)$ refers to the posterior probability of weights given data, $p(D|\omega)$ indicates the likelihood of weights given data, $p(\omega)$ is the prior probability of weights, and $p(D)$ is the marginal likelihood of data.

In classification problems [18], the predictive distribution based on equation (2) can be calculated as:

$$p(y^{'}|x^{'}, D) = \int p(y^{'}|x^{'}, \omega)p(\omega|D)d\omega \approx \int p(y^{'}|x^{'}, \omega)q(\omega)d\omega \tag{3}$$

The true posterior distribution in equation (3) which is $p(\omega|D)$ (or expressed as $p(\omega|x, y)$ for deep learning models), representing the model parameters ω, is inherently untraceable. However, it can be approximated with a simpler distribution $q(\omega)$, which must be traceable. In this study, the MC-Dropout method was employed to draw samples from the posterior distribution. MC-Dropout is a simple yet highly effective method for quantifying uncertainty in deep-learning models. It capitalizes on dropout layers that are used during training to prevent overfitting. During testing, MC-Dropout performs multiple forward passes through the network, retaining active dropout layers instead of deactivating them as in the conventional dropout method. MC-Dropout uses the Bernoulli distribution for each run, which leads to the generation of a binomial distribution over multiple-run averaging. MC-Dropout can have different rates for different models and can be managed through adjusting hyperparameters. Thus, the output of the network is transformed into a sample originating from the approximate posterior distribution of the model's parameters. The predictive distribution for MC-Dropout [18]can be calculated as

$$p(y^{'}|x^{'}, D) = \int p(y^{'}|x^{'}, \omega)p(\omega|D)d\omega \approx \frac{1}{T}\sum_{t=1}^{T} p(y^{'}|x^{'}, \omega) \tag{4}$$

The term $p(y^{'}|x^{'},\omega)$, in equation (4), is the predicted probability for each class in the t-th dropout sample, and T denotes the number of samples used to predict the binomial distribution. The summation of the predicted values for each class can be used to draw the predicted destruction, subsequently quantifying the associated uncertainty in the predictions.

3.4. Ensemble approach incorporating BDL models based on ranking

Ensemble methods, which refer to the use of multiple models to produce a common output, have been employed in deep-learning models for a relatively long time. The main purpose of employing ensemble models is to provide more robust predictive results compared to those produced by individual models. Bayesian model averaging is a well-established method that integrates Bayesian models and ensembles by averaging the predictions of given outputs. For classification, Bayesian model averaging calculates the mean for each class by averaging the predictions of the classes from each model.

The proposed method leverages Bayesian model averaging through the use of a ranking mechanism. While the conventional BMA uses averages for all models, the proposed method just averages the top-K model with the lowest uncertainty. The metric used to determine the top K predictions involves calculating the sum of the difference between the predicted probability of the "positive" class and that of the remaining "negative" classes. The predicted class is then determined by utilizing Softmax, a technique commonly used for multiclass classification, which assigns the class with the highest probability as the predicted class. It is important to note that the probability provided by Softmax sums to one, indicating the relative likelihood of each class. The difference utilized for ranking purposes can have a value between zero and one. A value of zero indicates that all classes have the same probability and there is high uncertainty in the model prediction. Conversely, a value of one denotes the situation where the positive class is assigned a probability of one while the remaining classes have probabilities of zero, which means that the model is very confident. However, it is worth noting that these two values, zero and one, are two extreme cases that rarely occur. Most of the time, the difference approaches one when the model is confident while deviating from one otherwise. This difference can be calculated using the following equation.

$$Diffrence = \frac{\sum_{i=1}^{n} Class_P - Class_i}{n - 1} \tag{5}$$

The term $Class_P$ refers to the highest probability associated with the predicted class "positive class". The term $Class_i$ refers to all probabilities of classes. The division by $n - 1$ is used to find the average of all differences between the positive and negative classes. The discrepancies (*Diffrence*) for all models are calculated and sorted. Subsequently, the highest-ranking K models are chosen from the sorted models based on the magnitude of the discrepancy, signifying lower uncertainty. These chosen models then undergo the BMA procedure. For example, if K is set to 3, then the top three models are chosen with the highest discrepancy between positive and negative classes, which means the lowest uncertainty; then, the BMA is applied to find the prediction of each class.

Algorithm 1summarizes the proposed ranking-based Bayesian deep learning ensemble method for uncertainty quantification for image classification.

Algorithm 1. Pseudocode of the proposed Ranking-based Bayesian deep learning ensemble method

```
Input: Training, validation, and test data
Output: Predicted class, posterior predictive distribution (mean and standard deviation)
  Method:
  Begin
    1) Load training, validation, and test data.
    2) Resize images to 128 × 128 pixels.
    3) Normalize data to have a similar scale.
    4) Find class weights of testing data to address imbalanced data issues.
    5) Load pre-trained CNN models on ImageNet.
    6) Tailor dense layers into a pre-trained model for classification.
    7) Train the model with MC-Dropout.
    8) Evaluate the model on testing data using MC-Dropout.
    9) Repeat steps 1 to 8 for different models.
    10) For each instance or sample predict the output classes.
    11) Find the difference between the predominant "positive" class and other "negative classes" using Eq. (5).
    12) Select Top K models for the ensemble and calculate BMA.
  End
```

4. Experiments and results

To check the effectiveness of the proposed method, we utilized five deep-learning models based on CNN architectures. These models were trained and tested using two separate image datasets, ALL and Breast Cancer, to evaluate the performance of the proposed method for tasks related to image classification. The training phase involved varying the number of epochs using the early stopping criteria. Typically, most of the models were trained for 30–60 epochs.

In this study, an assessment was conducted to determine the effectiveness of the proposed method using five pre-trained deep learning models, namely, DenseNet, MobileNet, MobileNetv2, VGG16, and VGG19 from the TensorFlow library. These models, distinguished by their unique architectural attributes, were chosen as prime candidates for evaluating the performance of the proposed method. The use of pre-trained versions of these models, which were initially trained on the ImageNet dataset, was extended to the two datasets used in this study for the classification task. To facilitate efficient processing and feature extraction, a global average pooling layer was incorporated following the output of each model. Subsequently, the resulting feature vectors were channeled through a series of three dense layers, 1024, 64, and 4/3 neurons per layer, respectively. These layer configurations were specifically tailored to accommodate the characteristics of each dataset. The Softmax activation function was applied to the final layer, facilitating the generation of class probabilities for each instance or sample.

Targeted fine-tuning was implemented to optimize model performance. For this purpose, the final ten layers of each model were fine-tuned using the abovementioned datasets, while the remaining layers were frozen to retain their original pre-trained features on ImageNet. Throughout the training phase, the Adam optimizer and Rectified Linear Unit (ReLU) activation functions were maintained across all model variants. It is important to highlight that the main goal of this research was not solely to attain state-of-the-art results within the utilized datasets but also to showcase the effectiveness and adaptability of the proposed method.

To provide a comprehensive overview of the models employed, it is important to outline their specific attributes. DenseNet, characterized by a robust deep CNN architecture, involves a total of 121 layers, contributing to its capacity for complicated feature extraction and representation. In contrast, MobileNet features a more streamlined architecture with approximately 28 layers, resulting in an efficient computational performance while maintaining good accuracy. The MobileNetv2 model, an advanced version of MobileNet, takes this efficiency of the original version a step further, with an extended architecture of approximately 53 layers, maintaining a balance between model complexity and computational cost. On the other hand, utilizing VGG models, the VGG16 is characterized by an architecture comprising 16 layers, while VGG19 is an extended version with 19 layers. This design variation facilitates a deeper representation of hierarchical features.

To quantify the uncertainty of each model, MC-Dropout was used as a Bayesian approximation method to derive samples from the posterior distribution. A total of 100 samples were drawn from the posterior distribution for each instance of the model. The method

Table 1
The performance of individual models using ALL dataset.

Models	Base Model Accuracy	Bayesian version		
		Accuracy	Mean	S. deviation
DenseNet	0.9815	0.9661	0.9784	0.0780
MobileNet	0.9861	0.9861	0.9946	0.0392
MobileNetV2	0.9768	0.9661	0.9842	0.0705
VGG16	0.9876	0.9876	0.9974	0.0240
VGG19	0.9738	0.9738	0.9878	0.0567

was utilized in two different ways to evaluate the effectiveness of the top K uncertainty ranking. The first method utilized the top K on the samples themselves, and the second method applied the same method to the mean of the classes after sampling. The detailed results from these methodologies are presented in Tables 1–4 for both datasets used.

Table 1 presents an exhaustive compilation of various models deployed along with their corresponding performances within the ALL dataset. Each model was tested in two distinct versions, and their performances were captured in terms of accuracy: the base model and the Bayesian version of the model. Both versions of the models were very similar in terms of the hyperparameters. Some Bayesian versions of models lack a bit behind non-Bayesian versions counterparts. Table 2 illustrates the performance of the ensemble of all five models, with the top three models referring to those models that used the difference (discrepancy) equation and ranking technique to select only the top three models with the highest difference rate between positive and negative predicted classes per instance. In general, the non-Bayesian model ensemble exhibited comparable or marginally lesser performance compared to Bayesian models. Intriguingly, Bayesian models that adhered to the top k = 3 in the tables performed better in terms of accuracy than Bayesian models that used all models to calculate the performance of the ensemble. Two distinct scenarios were explored: the first entailed conducting sampling on a single image, followed by the derivation of the posterior distribution for that image, and then using the mean to test the performance of the ensemble. In the second scenario, the difference (discrepancy) per sample was calculated to find the probability of each class, and then determine the posterior distribution of each class for an image. The ensemble performed better in both scenarios, particularly when the top three models were chosen, compared to models were all models used to test performance.

Table 3 presents the performance of the models used in this study on the breast cancer dataset. The table provides accuracy regarding both the base models and their corresponding Bayesian versions. Importantly, the configurations of both models were similar in terms of the hyperparameters used. Table 4 illustrates the performance of the ensemble for all five models across three distinct setup environments. In the first setup, an ensemble of non-Bayesian models was tested. When non-Bayesian models were used, the ensemble of all models demonstrated superior performance compared to the application of the difference (discrepancy) equation with the top k = 3. However, the performance of the non-Bayesian ensemble lags behind the Bayesian versions of the ensemble. In the second setup, Bayesian versions of models were utilized, where the posterior distribution of each class for an image was calculated, and then the difference (discrepancy) equation was applied to the mean of these classes with k = 3 to select the top three models with the highest difference between the predicted negative and positive classes. In this setting, ensembles with all models performed less compared to those comprised of the top three models. In the last setup, the discrepancy equation applied to each sample of an image for the prediction of each class was calculated. These samples were then used to draw the posterior distribution. Evidently, the ensembles with the top three models outperformed the other ensembles encompassing all models, manifesting enhanced accuracy. Moreover, the accuracy of Bayesian ensembles demonstrated a noteworthy enhancement of 1.52% for the top k = 3 models when the mean was used, and 0.76% for ensembles where samples were used. Overall, the results revealed that Bayesian ensembles, particularly those using the top-k models, outperformed non-Bayesian ensembles and ensembles using all models in terms of accuracy.

Figs. 4, 5, 6, and 7 show the probability of distribution of three images for all datasets, using ensemble samples. Fig. 4 indicates a case where the predicted image was misclassified. The early Pre-B class was erroneously identified as the predominant class despite not being so. In this particular image, four out of five models had misclassified predictions, with three of them incorrectly labeling it as early Pre-B. Notably, only one model classified it successfully; however, it also exhibited high uncertainty in the predicted class compared to the model uncertainty in general. Fig. 5 shows a case with low uncertainty and a correctly classified image. In this particular example, three models correctly classified prediction with low uncertainty, while two models classified it correctly but with moderate and high uncertainty in prediction. However, when utilizing the top three models, the predicted ensemble had low uncertainty and correctly classified images. Fig. 6 depicts an illustration of high uncertainty paired with accurate classification. In this example, three Bayesian models misclassified the true class in the prediction, whereas the other two models correctly classified true classes. Nonetheless, both ensemble versions, those picking the top three models or utilizing all models to predict the class, effectively classified input images with high uncertainty. The last illustration, Fig. 7, indicates the case where ensembles witnessed a shift in model predictions. In this particular case, the three Bayesian models made incorrect predictions, whereas the other two models correctly classified the outputs. Interestingly, the top three ensemble versions appropriately classified the input, whereas the ensemble version, whereas the ensemble version employing all models for prediction encountered misclassification of the input image.

In summary, the Bayesian deep-learning-based ensemble can leverage the performance of the Bayesian model while being able to quantify uncertainty. Bayesian ensembles generally perform better than non-Bayesian models, even though some individual Bayesian models may have lower accuracy rates. This is because the Bayesian model can generalize the prediction better than conventional models. Empirical evidence derived from this study indicates that the top K model in differences, according to these research findings, can perform better than averaging all models to calculate the prediction of ensembles, or at the very least, perform similarly with an

Table 2
The performance of different ensembles utilizing ALL dataset.

Type of ensemble	Accuracy	Mean	S. Deviation
Base model (non-Bayesian) (top 3)	0.9953		
Base model (non-Bayesian) (all models)	0.9953		
Bayesian per mean (top 3)	0.9969	0.9924	0.0484
Bayesian per mean (all models)	0.9953	0.9742	0.0731
Bayesian per samples (top 3)	0.9969	0.9924	0.0449
Bayesian per samples (all models)	0.9953	0.9742	0.0731

Table 3
The performance of individual models using the breast cancer dataset.

Models	Base Model Accuracy	Bayesian version		
		Accuracy	Mean	S. deviation
DenseNet	0.7948	0.7692	0.9244	0.1261
MobileNet	0.8076	0.7948	0.9567	0.1010
MobileNetV2	0.8012	0.8141	0.9382	0.1238
VGG16	0.8141	0.8141	0.9292	0.1129
VGG19	0.7884	0.7884	0.9558	0.0942

Table 4
The performance of different ensembles utilizing the breast cancer dataset.

Type of ensemble	Accuracy	Mean	S. Deviation
Base model (non-Bayesian) (top 3)	0.8269		
Base model (non-Bayesian) (all models)	0.8397		
Bayesian per mean (top 3)	0.8525	0.9281	0.1395
Bayesian per mean (all models)	0.8397	0.8690	0.1593
Bayesian per samples (top 3)	0.8461	0.9216	0.1445
Bayesian per samples (all models)	0.8397	0.8690	0.1593

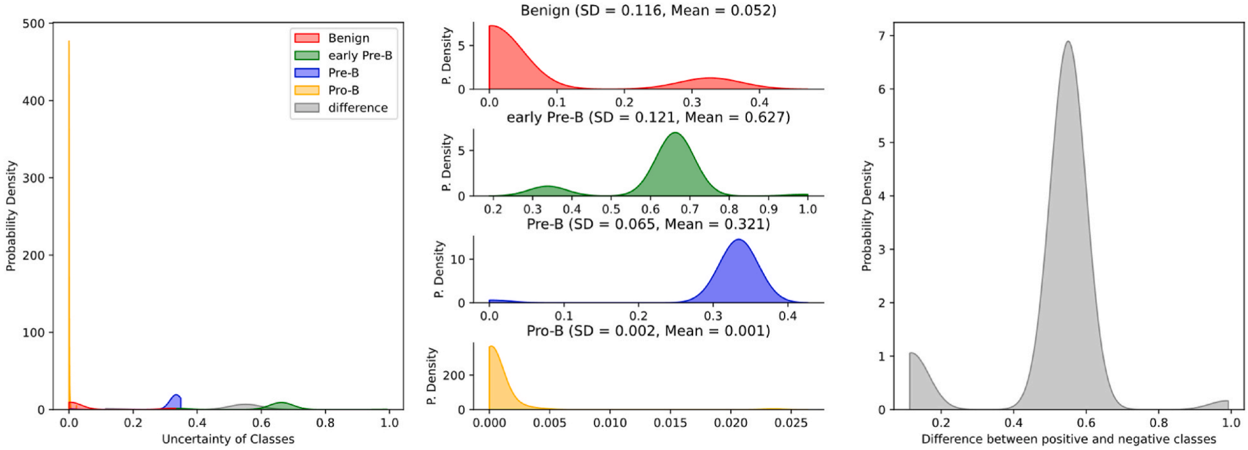

Fig. 4. Posterior distribution of misclassified ALL classes of ensemble output.

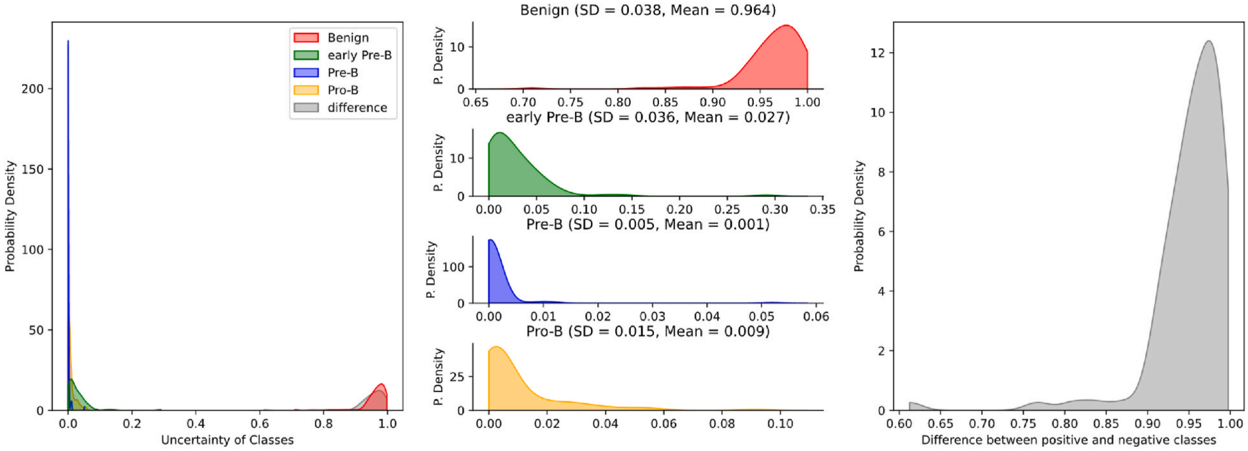

Fig. 5. Posterior distribution of correctly classified ALL classes of ensemble output with low uncertainty.

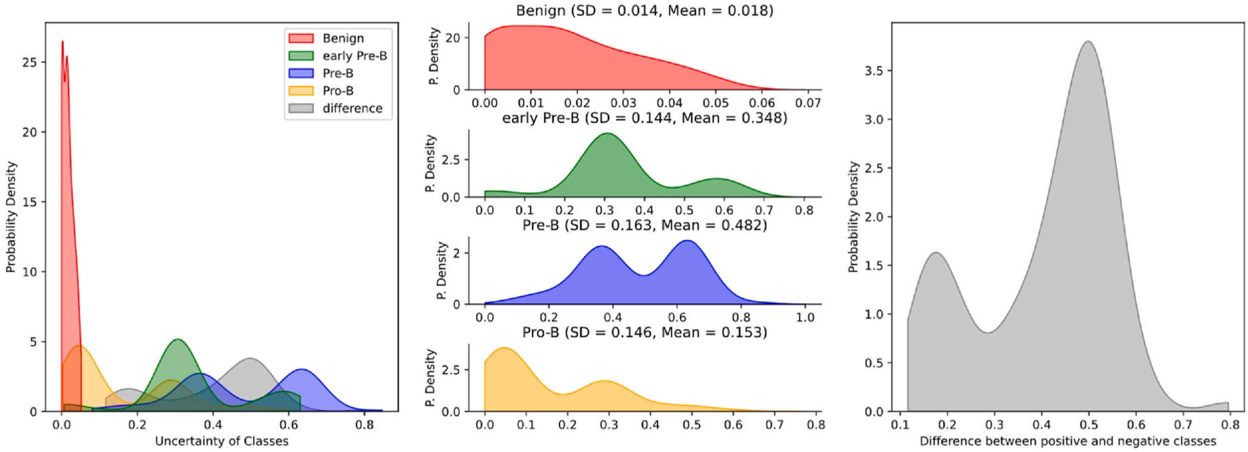

Fig. 6. Posterior distribution of correctly classified ALL classes of ensemble output with high uncertainty.

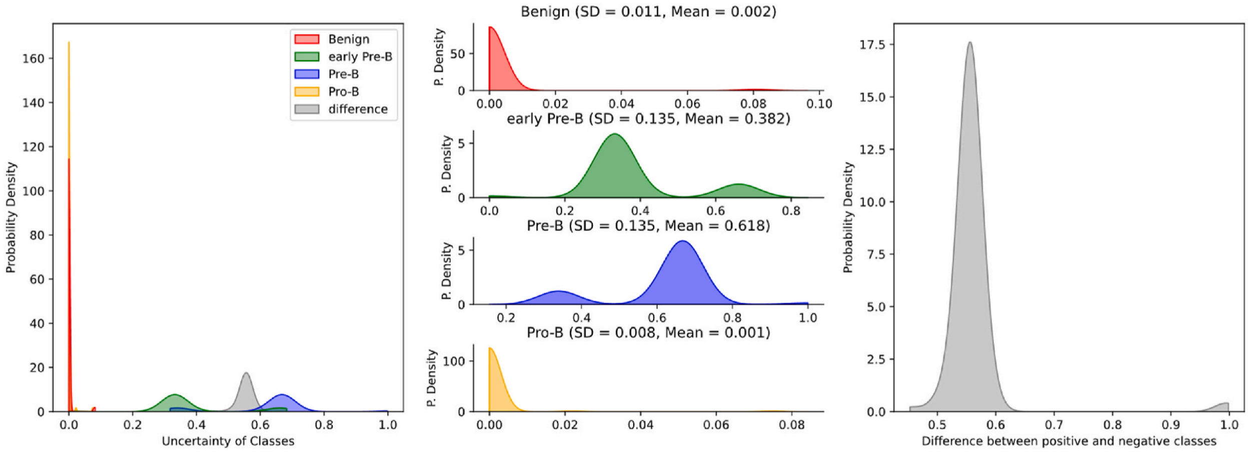

Fig. 7. Posterior distribution of correctly classified only for k = 3 ensemble for ALL classes with high uncertainty.

appropriate number of models. Further investigations are needed, particularly on larger datasets, to demonstrate the general impact of choosing only some models based on predefined criteria, such as the differences (discrepancy) proposed within this study.

5. Conclusion

Bayesian deep learning emerges as a potent approach for assessing uncertainty in image classification tasks. However, there are cases where the performance of Bayesian models in uncertain classification legs behind conventional classification models. To bridge this performance disparity, a promising way involves harnessing the power of Bayesian model ensembles. These ensembles have exhibited the capability to outshine individual Bayesian models and even marginally outperform non-Bayesian ensembles across various scenarios.

In this study, a novel technique has been introduced to leverage the ability of Bayesian ensembles to enhance uncertainty quantification. The proposed method utilizes the difference (discrepancy) between predicted positive and negative classes, employing this metric to effectively rank models. By selecting the top-k models for each sample or instance, the ensemble generates outputs for positive and negative classes. The experimental results illustrate that this method enhances the performance of the predicted outputs or, at the very least, yields a comparable impact in terms of accuracy. Notably, the Bayesian ensemble's accuracy witnessed an enhancement of up to 1.52% across various test datasets.

Additional information

No additional information is available for this paper.

Data availability statement

Data associated with this study are available through these sources:

1. M. Aria, M. Ghaderzadeh, D. Bashash, H. Abolghasemi, F. Asadi, A. Hosseini, Acute Lymphoblastic Leukemia (ALL) image dataset, Kaggle. (2021). https://doi.org/10.34740/KAGGLE/DSV/2175623.
2. W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data Brief. 28 (2020) 104,863. https://doi.org/10.1016/j.dib.2019.10486

CRediT authorship contribution statement

Abdullah A. Abdullah: Writing - review & editing, Writing - original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Masoud M. Hassan:** Writing - review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Yaseen T. Mustafa:** Writing - review & editing, Supervision, Methodology, Conceptualization.

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.

References

[1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, V. Makarenkov, S. Nahavandi, A review of uncertainty quantification in deep learning: techniques, applications and challenges, Inf. Fusion 76 (2021) 243–297, https://doi.org/10.1016/j.inffus.2021.05.008.

[2] A.A. Abdullah, M.M. Hassan, Y.T. Mustafa, A review on Bayesian deep learning in healthcare: applications and challenges, IEEE Access 10 (2022) 36538–36562, https://doi.org/10.1109/ACCESS.2022.3163384.

[3] 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), IEEE, 2020, pp. 7344–7350, https://doi.org/10.1109/ICRA40945.2020.9196844.

[4] P. Li, F.Z. Duraihem, A.U. Awan, A. Al-Zubaidi, N. Abbas, D. Ahmad, Heat transfer of hybrid nanomaterials base maxwell micropolar fluid flow over an exponentially stretching surface, Nanomaterials 12 (2022) 1207, https://doi.org/10.3390/nano12071207.

[5] W. Shatanawi, N. Abbas, T.A.M. Shatnawi, F. Hasan, Heat and mass transfer of generalized fourier and Fick's law for second-grade fluid flow at slendering vertical Riga sheet, Heliyon 9 (2023) e14250, https://doi.org/10.1016/j.heliyon.2023.e14250.

[6] A.U. Awan, N.A. Ahammad, W. Shatanawi, S.A. Allahyani, E.M. Tag-ElDin, N. Abbas, B. Ali, Significance of magnetic field and Darcy–Forchheimer law on dynamics of Casson-Sutterby nanofluid subject to a stretching circular cylinder, Int. Commun. Heat Mass Tran. 139 (2022) 106399, https://doi.org/10.1016/j.icheatmasstransfer.2022.106399.

[7] M.A. Almessiere, Y. Slimani, N.A. Algarou, M.G. Vakhitov, D.S. Klygach, A. Baykal, T.I. Zubar, S.V. Trukhanov, A.V. Trukhanov, H. Attia, M. Sertkol, İ.A. Auwal, Tuning the structure, magnetic, and high frequency properties of Sc-doped $Sr_{0.5}Ba_{0.5}Sc_xFe_{12-x}O_{19}/NiFe_2O_4$ hard/soft nanocomposites, Adv Electron Mater 8 (2022), https://doi.org/10.1002/aelm.202101124.

[8] D.I. Shlimas, A.L. Kozlovskiy, M.V. Zdorovets, Study of the formation effect of the cubic phase of LiTiO2 on the structural, optical, and mechanical properties of Li2±xTi1±xO3 ceramics with different contents of the X component, J. Mater. Sci. Mater. Electron. 32 (2021) 7410–7422, https://doi.org/10.1007/s10854-021-05454-z.

[9] D.A. Vinnik, A.Yu Starikov, V.E. Zhivulin, K.A. Astapovich, V.A. Turchenko, T.I. Zubar, S.V. Trukhanov, J. Kohout, T. Kmječ, O. Yakovenko, L. Matzui, A.S. B. Sombra, D. Zhou, R.B. Jotania, C. Singh, Y. Yang, A.V. Trukhanov, Changes in the structure, magnetization, and resistivity of $BaFe_{12-x}Ti_xO_{19}$, ACS Appl. Electron. Mater. 3 (2021) 1583–1593, https://doi.org/10.1021/acsaelm.0c01081.

[10] I.V. Korolkov, N. Zhumanazar, Y.G. Gorin, A.B. Yeszhanov, M.V. Zdorovets, Enhancement of electrochemical detection of Pb2+ by sensor based on track-etched membranes modified with interpolyelectrolyte complexes, J. Mater. Sci. Mater. Electron. 31 (2020) 20368–20377, https://doi.org/10.1007/s10854-020-04556-4.

[11] B. Song, S. Sunny, S. Li, K. Gurushanth, P. Mendonca, N. Mukhia, S. Patrick, S. Gurudath, S. Raghavan, I. Tsusennaro, S.T. Leivon, T. Kolur, V. Shetty, V. R. Bushan, R. Ramesh, T. Peterson, V. Pillai, P. Wilder-Smith, A. Sigamani, A. Suresh, moni A. Kuriakose, P. Birur, R. Liang, Bayesian deep learning for reliable oral cancer image classification, Biomed. Opt Express 12 (2021) 6422, https://doi.org/10.1364/BOE.432365.

[12] S. yadav, Bayesian deep learning based convolutional neural network for classification of Parkinson's disease using functional magnetic resonance images, SSRN Electron. J. (2021), https://doi.org/10.2139/ssrn.3833760.

[13] 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, https://doi.org/10.1016/J.COMPBIOMED.2021.104418.

[14] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: representing model uncertainty in deep learning, in: 33rd International Conference on International Conference on Machine Learning, ICML, JMLR.org, New York, 2016, pp. 1050–1059.

[15] O. Dürr, B. Sick, E. Murina, Probabilistic Deep Learning: with Python, Keras and TensorFlow Probability, Manning Publications Company, 2020.

[16] M.-H. Chen, Q.-M. Shao, J.G. Ibrahim, Monte Carlo Methods in Bayesian Computation, Springer New York, New York, NY, 2000, https://doi.org/10.1007/978-1-4612-1276-8.

[17] D.M. Blei, A. Kucukelbir, J.D. McAuliffe, Variational inference: a review for statisticians, J. Am. Stat. Assoc. 112 (2017) 859–877, https://doi.org/10.1080/01621459.2017.1285773.

[18] A.A. Abdullah, M.M. Hassan, Y.T. Mustafa, Uncertainty quantification for MLP-mixer using bayesian deep learning, Appl. Sci. 13 (2023) 4547, https://doi.org/10.3390/app13074547.

[19] S.S. Olofintuyi, E.A. Olajubu, D. Olanike, An ensemble deep learning approach for predicting cocoa yield, Heliyon 9 (2023) e15245, https://doi.org/10.1016/j.heliyon.2023.e15245.

[20] P.L. McDermott, C.K. Wikle, Deep echo state networks with uncertainty quantification for spatio-temporal forecasting, Environmetrics 30 (2019), https://doi.org/10.1002/env.2553.

[21] D. Althoff, L.N. Rodrigues, H.C. Bazame, Uncertainty quantification for hydrological models based on neural networks: the dropout ensemble, Stoch. Environ. Res. Risk Assess. 35 (2021) 1051–1067, https://doi.org/10.1007/s00477-021-01980-8.

[22] J. Tang, J. Hu, J. Heng, Z. Liu, A novel Bayesian ensembling model for wind power forecasting, Heliyon 8 (2022) e11599, https://doi.org/10.1016/j.heliyon.2022.e11599.

[23] S. Shin, Y. Her, R. Muñoz-Carpena, Y.P. Khare, Multi-parameter approaches for improved ensemble prediction accuracy in hydrology and water quality modeling, J. Hydrol. (Amst.) 622 (2023) 129458, https://doi.org/10.1016/j.jhydrol.2023.129458.

[24] O. Kisi, M. Alizamir, A. Docheshmeh Gorgij, Dissolved oxygen prediction using a new ensemble method, Environ. Sci. Pollut. Control Ser. 27 (2020) 9589–9603, https://doi.org/10.1007/s11356-019-07574-w.

[25] P. Thiagarajan, P. Khairnar, S. Ghosh, Explanation and use of uncertainty obtained by bayesian neural network classifiers for breast histopathology images, IEEE Trans. Med. Imag. (2021), https://doi.org/10.1109/TMI.2021.3123300.

[26] M. Gour, S. Jain, Uncertainty-aware convolutional neural network for COVID-19 X-ray images classification, Comput. Biol. Med. 140 (2022), https://doi.org/10.1016/j.compbiomed.2021.105047.

[27] M. Subramanian, M.S. Kumar, V.E. Sathishkumar, J. Prabhu, A. Karthick, S.S. Ganesh, M.A. Meem, Diagnosis of retinal diseases based on bayesian optimization deep learning network using optical coherence tomography images, Comput. Intell. Neurosci. 2022 (2022) 8014979, https://doi.org/10.1155/2022/8014979.

[28] M. Loey, S. El-Sappagh, S. Mirjalili, Bayesian-based optimized deep learning model to detect COVID-19 patients using chest X-ray image data, Comput. Biol. Med. 142 (2022) 105213, https://doi.org/10.1016/j.compbiomed.2022.105213.

[29] R. Egele, R. Maulik, K. Raghavan, B. Lusch, I. Guyon, P. Balaprakash, AutoDEUQ: automated deep ensemble with uncertainty quantification, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022, pp. 1908–1914, https://doi.org/10.1109/ICPR56361.2022.9956231.

[30] L. Hoffmann, I. Fortmeier, C. Elster, Uncertainty quantification by ensemble learning for computational optical form measurements, Mach Learn Sci Technol 2 (2021) 035030, https://doi.org/10.1088/2632-2153/ac0495.

[31] T. Pearce, F. Leibfried, A. Brintrup, Uncertainty in neural networks: approximately bayesian ensembling, in: S. Chiappa, R. Calandra (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 234–244, in: https://proceedings.mlr.press/v108/pearce20a.html.

[32] B. He, B. Lakshminarayanan, Y.W. Teh, Bayesian deep ensembles via the neural tangent kernel, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Adv Neural Inf Process Syst, Curran Associates, Inc., 2020, pp. 1010–1022, in: https://proceedings.neurips.cc/paper_files/paper/2020/file/0b1ec366924b26fc98fa7b71a9c249cf-Paper.pdf.

[33] J. Zhang, F. Li, F. Ye, An ensemble-based network intrusion detection scheme with bayesian deep learning, in: ICC 2020 - 2020 IEEE International Conference on Communications (ICC), IEEE, 2020, pp. 1–6, https://doi.org/10.1109/ICC40277.2020.9149402.

[34] D. Li, L. Marshall, Z. Liang, A. Sharma, Hydrologic multi-model ensemble predictions using variational Bayesian deep learning, J. Hydrol. (Amst.) 604 (2022) 127221, https://doi.org/10.1016/j.jhydrol.2021.127221.

[35] F. Seligmann, P. Becker, M. Volpp, G. Neumann, Beyond Deep Ensembles: A Large-Scale Evaluation of Bayesian Deep Learning under Distribution Shift, 2023.

[36] M. Aria, M. Ghaderzadeh, D. Bashash, H. Abolghasemi, F. Asadi, A. Hosseini, Acute Lymphoblastic Leukemia (ALL) Image Dataset, Kaggle, 2021, https://doi.org/10.34740/KAGGLE/DSV/2175623.

[37] J. Rawat, A. Singh, H.S. Bhadauria, J. Virmani, Computer aided diagnostic system for detection of leukemia using microscopic images, Procedia Comput. Sci. 70 (2015) 748–756, https://doi.org/10.1016/j.procs.2015.10.113.

[38] J.A. Burger, Treatment of chronic lymphocytic leukemia, N. Engl. J. Med. 383 (2020) 460–473, https://doi.org/10.1056/NEJMra1908213.

[39] M. Ghaderzadeh, F. Asadi, A. Hosseini, D. Bashash, H. Abolghasemi, A. Roshanpour, Machine learning in detection and classification of leukemia using smear blood images: a systematic review, Sci. Program. 2021 (2021) 1–14, https://doi.org/10.1155/2021/9933481.

[40] R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer statistics, 2022, CA A Cancer J. Clin. 72 (2022) 7–33, https://doi.org/10.3322/caac.21708.

[41] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data Brief 28 (2020) 104863, https://doi.org/10.1016/j.dib.2019.104863.

[42] A.B. Nassif, M.A. Talib, Q. Nasir, Y. Afadar, O. Elgendy, Breast cancer detection using artificial intelligence techniques: a systematic literature review, Artif. Intell. Med. 127 (2022) 102276, https://doi.org/10.1016/j.artmed.2022.102276.