Information Sciences 577 (2021) 353–378

Contents lists available at ScienceDirect

Information Sciences

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

BARF: A new direct and cross-based binary residual feature fusion with uncertainty-aware module for medical image classification

Moloud Abdar [a,*], Mohammad Amin Fahami [b], Satarupa Chakrabarti [c], Abbas Khosravi [a], Paweł Pławiak [d,e], U. Rajendra Acharya [f,g,h], Ryszard Tadeusiewicz [i], Saeid Nahavandi [a]

[a] Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Geelong, Australia
[b] Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran
[c] School of Computer Engineering, KIIT University, Bhubaneswar, India
[d] Department of Computer Science, Faculty of Computer Science and Telecommunications, Cracow University of Technology, Warszawska, Krakow, Poland
[e] Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Bałtycka, Gliwice, Poland
[f] Department of Electronics and Computer Engineering, Ngee Ann Polytechnic, Singapore
[g] Department of Biomedical Engineering, School of Science and Technology, SUSS University, Singapore
[h] Department of Biomedical Informatics and Medical Engineering, Asia University, Taichung, Taiwan
[i] Department of Biocybernetics and Biomedical Engineering, AGH University of Science and Technology, Krakow, Poland

ARTICLE INFO

Article history:
Received 5 March 2021
Received in revised form 25 June 2021
Accepted 3 July 2021
Available online 06 July 2021

Keywords:
Medical image classification
Fusion model
Deep learning
Early fusion
Uncertainty quantification
Monte Carlo dropout

ABSTRACT

Automatic medical image analysis (*e.g.*, medical image classification) is widely used in the early diagnosis of various diseases. The computer-aided diagnosis (CAD) systems enable accurate disease detection and treatment. Nowadays, deep learning (DL)-based CAD systems have been able to achieve promising results in most of the healthcare applications. Also, uncertainty quantification in the existing DL methods have not gained enough attention in the field of medical research. To fill this gap, we propose a novel, simple and effective fusion model with uncertainty-aware module for medical image classification called Binary Residual Feature fusion (BARF). To deal with uncertainty, we applied the Monte Carlo (MC) dropout during inference to obtain the mean and standard deviation of the predictions. The proposed model has two main strategies: direct and cross validated using four different medical image datasets. Our experimental results demonstrate that the proposed model is efficient for medical image classification in real clinical settings.

1. Introduction

In the past few years deep learning (DL) [1–3], one of the subfields of machine learning (ML), has seen a histrionic renaissance that is mainly due to computational supremacy and accessibility to huge datasets. The area has seen outstanding advancement in terms of capability of machines to recognize and manipulate data (*i.e.*, images [4], language and text analysis [5,6], gene selection [7], etc.). The two fields that stand to gain vastly from the resurgence of DL are healthcare and medicine.

* Corresponding author.
E-mail addresses: m.abdar1987@gmail.com, mabdar@deakin.edu.au (M. Abdar), mafahami@gmail.com (M.A. Fahami), chakrabartisatarupa@gmail.com (S. Chakrabarti), abbas.khosravi@deakin.edu.au (A. Khosravi), plawiak@pk.edu.pl (P. Pławiak), aru@np.edu.sg (U.R. Acharya), rtad@agh.edu.pl (R. Tadeusiewicz), saeid.nahavandi@deakin.edu.au (S. Nahavandi).

https://doi.org/10.1016/j.ins.2021.07.024
0020-0255/

Among the various types of computer programming, ML is characteristically distinct as it has the ability to transform input into outputs with the help of rules that are derived from set of training examples. Essential requirements for developing a ML system are proficiency in the domain and engineering concepts that aid in building specific feature extractors for transforming data and perceiving suitable patterns. On the other hand, DL has the competence of developing and learning patterns automatically from raw data or feature learning from data with the help of successively arranged layers comprising of primitive non-linear operations. Determining features and performing a task are combined into one framework. Highly complex functions are learned with iterative warping of data through the layers. The scalability of DL model pertaining to large dataset enables them to outperform traditional ML models.

Distinct features of DL models have proved to be useful in image analysis, specifically in medical domain. Image acquisition devices have seen incredible advancement and thus data generated is also large. This swift growth in medical images also required extensive interpretation by human expert which is prone to error because of intricacy of images, difference in explanation approaches, subjectivity, accuracy and throughput [8]. Identifying objects from natural images has been made easier by empowering machines with the use of DL networks. Advancement in DL technology is gradually getting incorporated across the healthcare sector through imaging based medical diagnosis and data processing. Image diagnosis deals with identification of abnormalities, quantification of measurement and detecting changes over time. DL methods have shown potential in providing state-of-the-art unbiased and automated interpretation of different types of medical images that are useful for information processing and accurate diagnosis. DL techniques have opened new avenues in healthcare sector addressing wide range of problems from screening of carcinomas to monitoring of diseases and treatment suggestions.

1.1. Fusion-based machine learning and deep learning methods

Imaging technology is a vital aspect of medical diagnosis but suffers from information limitation due to single modal medical images. Due to this reason fusion of medical images is one of the sought-after research areas. There are two major types of medical image fusion – single mode fusion and multimodal fusion [9]. As single-mode lacks in providing sufficient information, multimodal image fusion is studied as it covers an extensive range of methods that address complex medical related issues [10]. Amalgamation of images with exhaustive spectral and anatomical information from either a single modality or multiple modalities into a single image is known as image fusion. The chief emphasis of image fusion lies in improving the quality of a particular image while safeguarding the pertinent characteristics of image such that it can be used for diagnosis. Thus, robust and self-learning techniques are required for medical image analysis using fusion method. Traditionally there are two types of fusion – early fusion and late fusion [11]. Illustration of features from multiple modalities is formed in early fusion which is followed by learning of correlation and connections between features of individual modality. Late fusion allows use of unimodal decision values along with a fusion mechanism. Usage of different modules on diverse modalities creates flexibility and makes it easier to handle missing modality. In multimodal approach domain-specific deep neural networks (DNNs) are used to create the individual representations that are finally combined using addition or concatenation methods [12].

Late fusion for convolutional networks has been studied along with two other fusion strategies, namely information exchange at an intermediate layer and simultaneous linking of information using cross-stitch technique at different layers, to provide a detailed comparison among the three fusion techniques for DL [13]. Apart from the mentioned fusion technique, for biometric recognition system two types of fusion methods: pre-classification and post-classification are presented. In pre-classification, fusion occurs before classification and procedures used are feature-level and sensor-level techniques. Pre-classification fusion generally suffers from redundant data, noise and problems pertaining to multi-environment acquisition of images. The post-classification fusion takes place after the classification procedure, thus it is devoid of noise and enhances the recognition performance of the system [14]. Spatial and temporal based medical image fusion methods have shortcomings while the extracting features and to overcome the defects, convolutional neural network (CNN) has been used for image fusion. Fusion branch has been used in deep CNN to integrate the classification results of two decoder branches such that the alpha values are obtained as the result of soft segmentation [15].

It can be noted from previous studies that fusion techniques are time consuming and failed in utilizing the temporal information. A progressive fusion network has been used to extract intra-frame temporal as well as spatial correlation among several low-resolution frames. Therefore, medical image fusion varies from spatial domain to DL and researchers have proposed various fusion methods with their own set of advantages and shortcomings. Most of the fusion methods are open to modification and address the problems related to it. DL has played a crucial part in improving the effect of fusion in medical images [16]. Despite all these, the important point is the level of trust in machine and DL models in their outcomes. In other words, the certainty of the methods regarding the results obtained remains a very important and key area.

1.2. Uncertainty quantification in deep learning and machine learning

In application of UQ techniques play an important role to enhance the predictability and validity of ML and DL methods [17]. A comprehensive review on UQ methods in ML and DL is presented in [17]. It is used to evaluate the reliability of a ML or DL model. During the development of DL algorithms, efforts are mainly directed to boost the performance of the model instead of risk management associated with it. Images are important in medical imaging and application because they help in diagnosis and provide clinical intervention options. As a result, such important applications necessitate a method that can

enumerate the risk of failure, quantify uncertainties and identify the source of the uncertainty. In DL systems predictive failure occurs due to innate ambiguity of task or failure of trained model to define the data. Ambiguity in the data results in intrinsic uncertainty while model uncertainty stems from vagueness in model specifications. The contributing factors to the model uncertainty are parameter uncertainty and bias of the model. Hence, parameter and intrinsic uncertainty mainly contribute to the prognostic failure of various DL models. Precise estimation of uncertainties is necessary as it will help us to understand the limitations of learning models and alert unsure predictions. Across a wide-ranging spectrum of applications, DL-based models have attained commendable predictive precision but meticulous quantification of their predictive uncertainty is a challenging task specially in the medical field. Over the years, researchers have addressed this problem with different techniques and approaches.

1.3. Uncertainty quantification in medical data analysis

Medical image interpretation is an exciting task as it is intricate and often marks the presence of unwanted artifacts, obstructions, limited contrast and so on. Use of DL methods for analyzing medical data comes with its advantages as well as disadvantages in form of uncertainty. Predictions lacking UQ are not considered reliable. Therefore, UQ is an important aspect in DL and more evidently in medical data analysis as decisions taken or diagnosis made based on the outcome of the different methods will have direct impact in real life situation. Uncertainty modeling is based on two main uncertainties namely – epistemic or model uncertainty and aleatoric or data uncertainty. Studies have shown the use of Bayesian techniques, Gaussian methods and ensemble techniques for the quantification of uncertainty [17]. As discussed in the recently published review paper, classification uncertainty is associated with DL models [17]. The presence of noise is estimated by classification uncertainty. Bayesian DL (BDL) method depended on both uncertainty and calibration [18] helped in increasing the classification accuracy of the network. Raczkowsk et al. [19] used accurate, reliable and active (ARA) framework (image classification) with Bayesian CNN (BCNN) for the classification of histopathological images of colorectal carcinoma. The proposed work calculated the uncertainty associated with each tested image. Authors also showed that using variational dropout-based entropy measure of uncertainty increased the learning process of DL network. Therefore, it can be analyzed from different studies that evolving DL techniques have the potential to transform the biomedical image processing. But it comes the challenge of reliability of DL prediction and hence, the issue of uncertainty emerges. With advancement in research, different techniques have been developed to quantify the uncertainty such that reliability and prediction capability of DL methods can be enhanced.

1.4. Concluding remarks and organization

This study aims to propose a new fusion mechanism for greater productivity of medical image features. But at the same time, we have a glimpse of whether our proposed model is aware of the uncertainty of the obtained results or not. As mentioned earlier, UQ plays an important role in increasing the trust in the results obtained by various machine and DL methods. Therefore, it can be argued that one of the key motivations of this study is to propose intelligent-based approaches for classifying medical images. In addition, another motivation is that, our proposed models should be aware of their certainty and uncertainty during predictions. Furthermore, presenting models with outstanding performance and promising outcomes is also one of our main motivations in this study. Finally, we seek to increase trust in the results obtained by clinicians, physicians and patients by considering uncertainties and maximizing the uncertainty of predictions. This is very important challenge in medical applications. Taking into account the differences and gaps in the previous studies, we propose two novel, simple yet efficient fusion models called direct-based **B**in**A**ry **R**esidual feature **F**usion (**direct-based BARF**) and **cross-based BARF** for medical image classification. To do so, the proposed fusion model uses both the binary tree combination (BTC) and residual combination techniques. Meanwhile, our proposed fusion model is able to deal with its uncertainties using Monte Carlo (MC) dropout. Furthermore, the proposed model benefits from the combination of MC dropout with standard dropout. Along with dealing with uncertainty, providing a comprehensive model that can perform well on a variety of data is one of the important gaps in the medical domain. Therefore, it can be summarized that having impressive performance along with considering uncertainty is a key goal in medical studies using ML and DL methods. Hence, we considered this important goal of our study. In summary, the main contributions of the study are listed as follows:

- Proposed *two* new fusion techniques (direct and cross-based BARF) to extract features from different DL methods and used for medical image classification.
- Computed the performance of all applied fusion models for each class separately.
- Applied both BTC and residual combination techniques to develop the proposed fusion models.
- Quantified prediction uncertainties for the proposed fusion models using MC dropout.
- Combined MC and standard dropouts are employed to optimize the performance of the proposed fusion models;
- Considered data size (big and small) and data type (*i.e.*, grayscale or color).
- Compared the performance of our proposed fusion models with different DL methods when tested on four different medical image datasets.

The rest of this study includes the following sections. Section 2 provides main preliminaries linked to our proposed model. In Section 3, the proposed uncertainty-aware BARF models are discussed in detail. The experimental results and discussions are presented in Sections 4 and 5, respectively. Finally, we conclude the study and list several open research directions in Section 6.

2. Preliminaries

In this section, we first explained the main differences between fusion models. Then, we briefly explained CNNs (also called ConvNet) and Residual Neural network (ResNet). Then, we discussed the working of binary tree combination (BTC). The detailed application of Monte Carlo (MC) dropout, difference between MC dropout and standard dropout are discussed.

2.1. Fusion models

A broad range of fusion models have exhibited extraordinary performance to optimize the results of single models [20,21]. Fig. 1 shows four different well-known fusion models including early or data-level fusion, intermediate fusion, slow fusion and late or decision-level fusion. Fusion models fist train individual modalities and then joined by using different strategies such as voting, score averaging, Canonic Correlation Analysis (CCA) and many more [13]. A significant number of previous studies on supervised learning tasks have broadly relied on ensembling of feature embeddings of separately extracted trained deep models (called late or decision-level fusion) for a wide variety of applications. But the early and data-level fusion have also obtained promising outcomes. But each of these models have strengths and weaknesses. Meanwhile, fusing diverse modalities of DL-based medical image analysis is conducted using various fusion models [22]. These fusion models can be considered as an interdisciplinary research field which combine and correlate disparate homogeneous and heterogeneous data modalities to deal with a wide variety of difficult prediction tasks in different research areas such as human–computer interaction, computer vision, biomedical informatics, medical data analysis and many more. In other words, depending on the medical task/problem, multimodal fusion approaches can range from fusion of multi-view data of the same modality or fusing heterogeneous data modalities. The previous studies have shown that application of diverse types of fusion models for medical data analysis have yielded extraordinary results. Based on these advantages, we employed a novel the early feature-level fusion model for medical image classification task.

2.2. CNNs and ResNet

Convolutional neural networks (CNNs or ConvNet) are a type of DNNs developed based on their shared-weights structure and translation invariance which are an alternative class of classical artificial neural networks (ANNs) [23]. CNNs have achieved remarkable success in a board range of applications such as computer vision [24], text analysis [5] and many more. There are various pre-trained architectures based on CNNs. Residual-based NN is a pre-trained deep CNN model which uses the short-cut concept introduced in 2015 [25]. Fig. 2 compares the building diagram of CNNs and ResNet. Residual connections are a class of skip connections which permit gradients to flow using a network directly.

As indicted in Fig. 2, residual connections allow the flow of information from initial layers to other layers except their next layer. Inspired by this strategy, we proposed a novel, simple and efficient residual-based fusion model for medical image classification. It is worth noting that, we have only one skip connection. In other words, each layer of our proposed model also connects to its second higher layer. The more details about the proposed model is provided in the following subsections.

2.3. Binary Tree Combination (BTC)

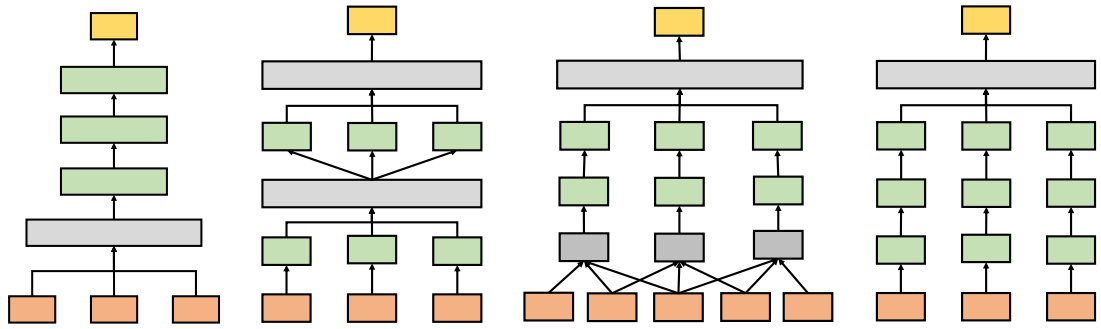
A binary tree is a tree-based structure in which each node has only two sub-trees (also called children). There is a difference between a normal tree and binary tree. There is no limit on the degree of nodes in a normal tree while the degree of each node in a binary tree is not more than two. In this study we employed a complete binary tree which has the following characteristics:

- All higher (external) nodes in our model has two internal children;
- In the proposed binary tree, the depth is same from the leaves to the root in the entire process.

In the normal routine of trees, we start from the root and reach the leaves, but we employed the reverse system in our work. We connected the leaf nodes in pairs step by step to form the root. Generally, the total nodes (N) of a tree with height h and tree's degree d can be calculated as

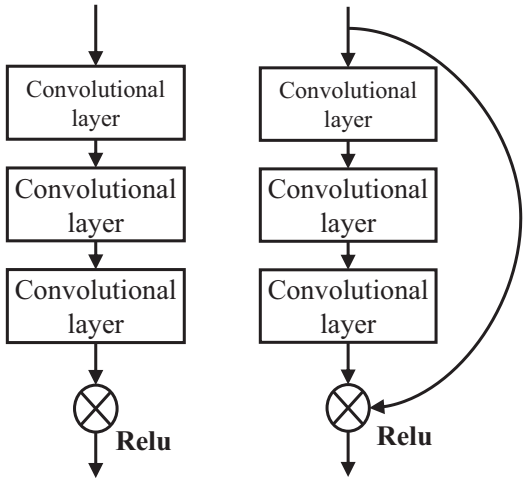
$$N = \sum_{i=0}^{h-1} d_i = \frac{d^h - 1}{d - 1}, \tag{1}$$

where h is the height of a tree and d is the tree's degree. Therefore, for a complete binary tree ($d = 2$), we have



(a) Early or data-level fusion (b) Intermediate fusion (c) Slow fusion (d) Late or decision-level fusion

Fig. 1. Illustration of various fusion-based architectures [20,21].



(a) Traditional CNNs (b) ResNet

Fig. 2. Traditional architecture of: (a) CNN and (b) ResNet [26]. *Note*: ReLU is the rectified linear unit.

$$N = \sum_{i=0}^{h-1} 2_i = \frac{2^h - 1}{2 - 1} = 2^h - 1. \tag{2}$$

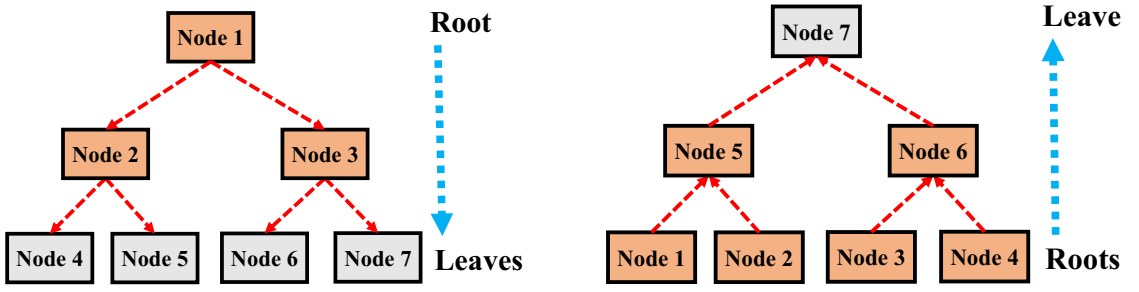
In summary, for a complete binary tree with height h, the tree has $2^h - 1$ nodes (total nodes) and number of leaves is 2^h. In this study, we employed reverse binary tree strategy (Fig. 3.b) as our combination approach called Binary Tree Combination (BTC).

2.4. Uncertainty quantification module

Nowadays, ML and DL methods have a significant ability to ideally learn powerful representations and then map the high dimensional data samples to an array of outputs. Due to this ability, ML and DL methods are able to achieve promising predictive results. However, they poorly perform in measuring uncertainties based on the obtained results. Basically, almost all ML and DL methods "*do not know*" what exactly "*they know*" and for this reason, they may certainly classify one sample which they have never seen before. The uncertainty estimate of such methods help to understand and explore what exactly these models do not know. Furthermore, UQ plays a fundamental role in dealing with noise structures. This unique feature provides exceptional conditions for the ML and DL models to better understand their limits and acknowledge uncertain predictions.

2.4.1. Uncertainty in NN/DL models

Generally, there are two major views on types of uncertainty. But, Tagasovska and Lopez-Paz [27] suggested three important types: approximation, aleatoric and epistemic uncertainties. While, more studies considered just aleatoric and epistemic



(a) Traditional binary tree (b) Reverse binary tree

Fig. 3. Traditional binary tree (left) vs reverse binary tree (right). As can be seen, for the reverse binary tree (right), we combine roots to reach the final leaf. This approach is used in our BTC method.

uncertainties as main ones [28–31]. Based on the previous studies, we considered *aleatoric* and *epistemic* uncertainties as two main uncertainties in ML and DL. The exact definition of aleatoric and epistemic uncertainties can be found in [28,32]. There are many UQ methods to deal with uncertainties such as Monte Carlo (MC) dropout [32], deep ensemble [33], BDL [34,35], and so on. In this work, we quantified uncertainties using MC dropout [32].

2.4.2. Monte Carlo (MC) dropout as uncertainty measure

MC dropout introduced by Gal and Ghahramani [32] is a powerful approach used to perform VI on BNNs. The main purpose of Bayesian technique is to find the most appropriate posterior distribution which is often intractable from a computational point of view. To solve this problem, sampling methods are widely being used. As shown in [32] MC sampling scheme is used to sample the posterior of BNNs (or BDL) from the prior parameter by applying multiple stochastic forward passes during test time.

As explained in [32] MC dropout is another approach which performs VI on BNNs. Based on excellent performance of MC dropout in the review literature [36], we employed it as our UQ method. A brief details of MC dropout is given in the following. Suppose $\hat{y}$ be the final output of an ANN model having L layers and loss function $E(.,.)$ (*e.g.*, the softmax loss). Let y_i be the observed output of x_i for $i = 1$ to N data points and **X,Y** be the input and output sets, respectively. To obtaine a minimization objective (also called as cost), we can frequently use L_2 regularisation which weighted by different weight decay λ:

$$\mathscr{L}_{dropout} := \frac{1}{N}\sum_{i=1}^{N} E(y_i, \hat{y}_i) + \lambda \sum_{i=1}^{L}\left(||\mathbf{W}_i||_2^2 + ||\mathrm{b}_i||_2^2\right). \tag{3}$$

Based on [32], MC dropout in different ANN and DL models can be used as an uncertainty measure. To show the predictive distribution of the model, we used $p(y|x,\mathscr{D})$ which y is the target (class), x is the input and $\mathscr{D}$ includes whole training data with N samples $\mathscr{D} = (x_i, y_i)_{i=1}^{N}$. The predictive distribution can inspect the variance to indicate the uncertainty. In order to calculate the predictive distribution, a distribution over the functions or the parameters as the posterior distribution (*i.e.*, $p(\Theta|\mathscr{D})$) should be used. The MC dropout [32] can provide a scalable solution to learn such predictive distributions. Each dropout is relevant to different sample obtained from the approximate parametric posterior distribution as $\Theta_t \sim q(\Theta|\mathscr{D})$ where Θ_t is a dropout configuration or a simulation obtained by $q(\Theta|\mathscr{D})$. Therefore, the approximate predictive distribution is given by

$$q(y^*|x^*) = \int p(y^*|x^*,\omega)q(\omega)d\omega, \tag{4}$$

where q(ω) is the variational distribution and $\omega = \{\mathbf{W}_i\}_{i=1}^{L}$ is the random variable set related to a model including L layers. Afterwards, the predictive variance of model can be calculated by

$$\mathrm{Var}_{q(\mathbf{y}^*|\mathbf{x}^*)}\left((\mathbf{y}^*)^T(\mathbf{y}^*)\right) \approx \tau^{-1}\mathrm{I}_D + \frac{1}{N}\sum_{t=1}^{T}\hat{\mathbf{y}}^*\left(\mathrm{b}^*, \mathbf{W}_1^t, \ldots, \mathbf{W}_L^t\right)^T\hat{\mathbf{y}}^*\left(\mathrm{x}^*, \mathbf{W}_1^t, \ldots, \mathbf{W}_L^t\right) - \mathbb{E}_{q(\mathbf{y}^*|\mathbf{x}^*)}(\mathbf{y}^*)^T\mathbb{E}_{q(\mathbf{y}^*|\mathbf{x}^*)}(\mathbf{y}^*). \tag{5}$$

For simplicity, the MC integration of the model's likelihood can uncover the predictive distribution as follows:

$$\begin{aligned} q(\mathrm{y}|\mathrm{x}) &\overset{\mathrm{VI}}{\approx} \int_{\Omega} \underbrace{p(\mathrm{y}|\mathrm{x},\Theta)}_{\text{Likelihood}}\underbrace{q(\Theta|\mathscr{D})}_{\text{Posterior}}d\Theta, \\ &\overset{\mathrm{MC}}{\approx} \frac{1}{N}\sum_{t=1}^{T} p(\mathrm{y}|\mathrm{x},\Theta_t), \mathrm{w.r.t.}\Theta_t \sim q(\Theta|\mathscr{D}), \end{aligned} \tag{6}$$

where VI represents variational inference. In order to further simplify these equations, the likelihood can be considered as a Gaussian distribution:

$$p(\mathrm{y}|\mathrm{x},\Theta) = \mathcal{N}\left(\mathbf{f}(\mathrm{x},\Theta), \mathrm{s}^2(\mathrm{x},\Theta)\right), \tag{7}$$

where the Gaussian function $\mathcal{N}$ can be specified by using the mean $\mathbf{f}(\mathrm{x},\Theta)$ as well as the variance $\mathrm{s}^2(\mathrm{x},\Theta)$ corresponding to the output of the MC dropout. As a summary we have:

$$\mathrm{MCdropout}(\mathrm{x}) \sim \mathbf{f}(\mathrm{x},\Theta), \mathrm{s}^2(\mathrm{x},\Theta). \tag{8}$$

In Fig. 4, we illustrated the working of MC dropout[1] in NNs.

2.4.3. MC dropout vs standard dropout

Let's start with standard dropout which is applied only during training time. Standard dropout is a regularization approach used to deal with overfitting. It should be noted that standard dropout is not used during the test time. Instead, there are all the connections and nodes, but the weights are accordingly adjusted. It may be noted that for standard dropout, the predictions during the test time is deterministic. The MC dropout can be applied during both training and test time. Unlike standard dropout, the predictions are not deterministic, but it depends on the nodes or links are randomly chosen. Hence, the applied model with MC dropout predicts diverse values each time. Also, the fundamental goal of MC dropout is to acquire some random predictions and then interpret those random predictions as samples of a probabilistic distribution.

3. Proposed uncertainty-aware BARF models

In this section we introduced our proposed fusion models in two sub-sections. First, we showed a schematic view of BARF model to show its working. We then explained the direct-based BARF model which is a prerequisite to understand our second proposed architecture called cross-based BARF. Both models contain uncertain (probabilistic) modules which help to measure the uncertainty of the whole model. A general view of the proposed BARF fusion model based on BTC is illustrated in Fig. 5. (Fig. 6).

3.1. Proposed direct-based BinAry Residual feature Fusion (BARF)

The early fusion techniques generate joint consensus of different input features obtained from several modalities (models). The proposed *BARF* is type of ensemble of different methods. The early feature-level fusion includes many features and hence the running time increases. However, such large-scale feature vectors along with appropriate learning methods provide better performance in the end. Hence, having outstanding performance justifies the uniqueness of a ML/DL method. In addition, better performance and accurate diagnosis in medicine is inevitable because of the close connection between the diagnosis and people's lives. Hence, we have provided a new early feature-level fusion model which can classify medical images accurately.

In this study, two new feature level fusion models (named direct and cross-based *BARF* models) are implemented by a simple yet efficient concatenation of different feature sets obtained from multiple information sources (pre-trained DL models). The direct-based BARF model utilizes different useful properties of recent modern state-of-the-art pre-trained DL models, including the DenseNet-201, VGG-19, EfficientNet-B7, InceptionResNetV2, InceptionV3, MobileNetV2, ResNet152V2, and Exception methods. Let $X = \langle X_1, X_2, X_3, \ldots, X_8 \rangle$, where X_i is the output embedding of ith pre-trained model. Let $X_i = \left\langle X_{i_1}, X_{i_2}, X_{i_3}, \ldots, X_{i_j} \right\rangle$ be the flattened output of the eight pre-trained models. Suppose, X_i is a one-dimensional vector which contains j real numbers in $\mathbb{R}$.

In this study, we used three types of Dense layers in our proposed *BARF* model: Dense 64, Dense 32, and Dense 16 at first, second, and third levels of our model, respectively. The formulation of each node at Dense layers (Dense 64) used in the first level of our BARF model is defined as follows:

$$D_{64}(X_i) = \mathrm{ReLU}\left(\sum_j R_{i_j} X_{i_j}\right), \tag{9}$$

where D_{64} is the output of each node in the Dense 64, ReLU is its activation function, and R_i is the coefficient vector to create the linear combination of vector X_i.

It should be noted that there are four Dense layers (Dense 32) at the second level of the BARF model. Here we present the formula of the first Dense layer at the second level of BARF model which is the same for the other three Dense layers. To obtain this formula, we replace the vector X_i in Eq. 9 by:

$$Y = concat(Dropout_{0.3}(D_{64}(X_1)), Dropout_{0.3}(D_{64}(X_2))), \tag{10}$$

[1] Source: https://docs.aws.amazon.com/prescriptive-guidance/latest/ml-quantifying-uncertainty/mc-dropout.html

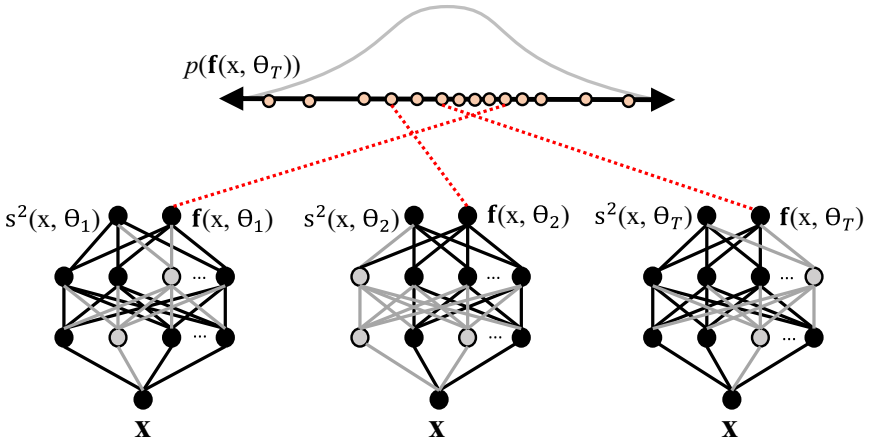


Fig. 4. A general view of MC dropout in NNs. The gray circles show some randomly switching neurons off whereas MC dropout uses the black circles to show the neurons in forward propagation.

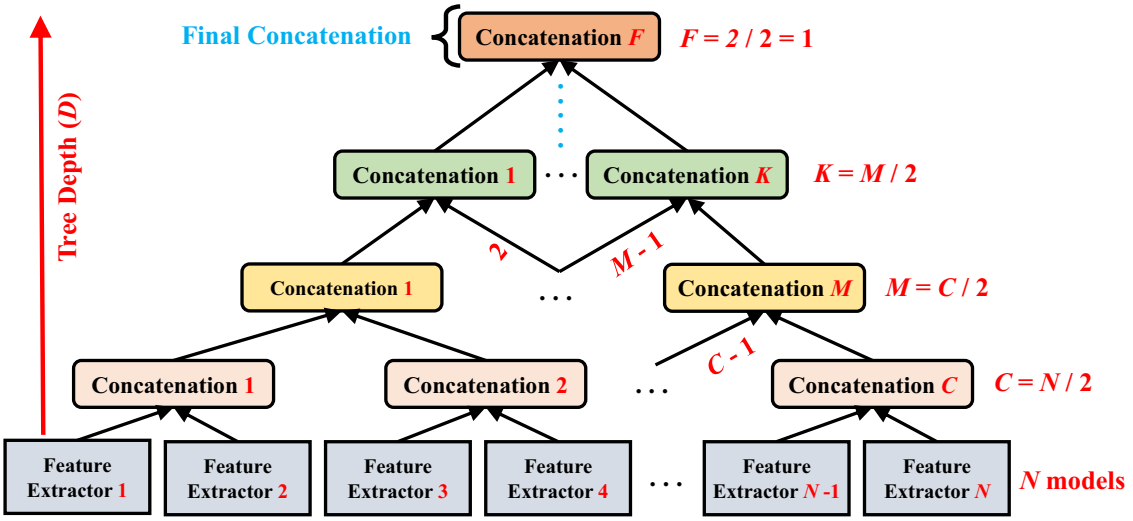


Fig. 5. A general view of BTC strategy used in the proposed BARF fusion models. In this work, we have N models for feature extraction and the depth of tree is D. As mentioned earlier, this model follows the idea of BTC which means at each step forward half of the base models of the previous level remains. For example, we can see that C is equal to $N/2$. This means that the number of base methods (N) must be a power of 2 ($N = 2^j$), where $j = \{1, 2, 3, \ldots, \mathbb{N}\}$. *Note*: N is the number of base models while $\mathbb{N}$ is the symbol of the set of natural numbers. The depth D is equal to j.

$$D_{32}(Y) = \text{ReLU}\left(\sum_j R_j Y_j\right), \tag{11}$$

where j the length of vector Y.

Finally, we present the formula for the Dense layers at the third level of BARF model. There are two Dense layers (Dense 16) at the third level. Here, we present the formula for the first one. Consider:

$$Z = concat(D_{32}(1), D_{32}(2), MCDropout_{0.3}(D_{64}(X_1)), MCDropout_{0.3}(D_{64}(X_2)), MCDropout_{0.3}(D_{64}(X_3)), MCDropout_{0.3}(D_{64}(X_4))), \tag{12}$$

Then:

$$D_{16}(Z) = \text{ReLU}\left(\sum_j R_j Z_j\right), \tag{13}$$

where $D_{32}(1)$ and $D_{32}(2)$ are the first and second Dense 32 layers at the second level and j the length of vector Z.

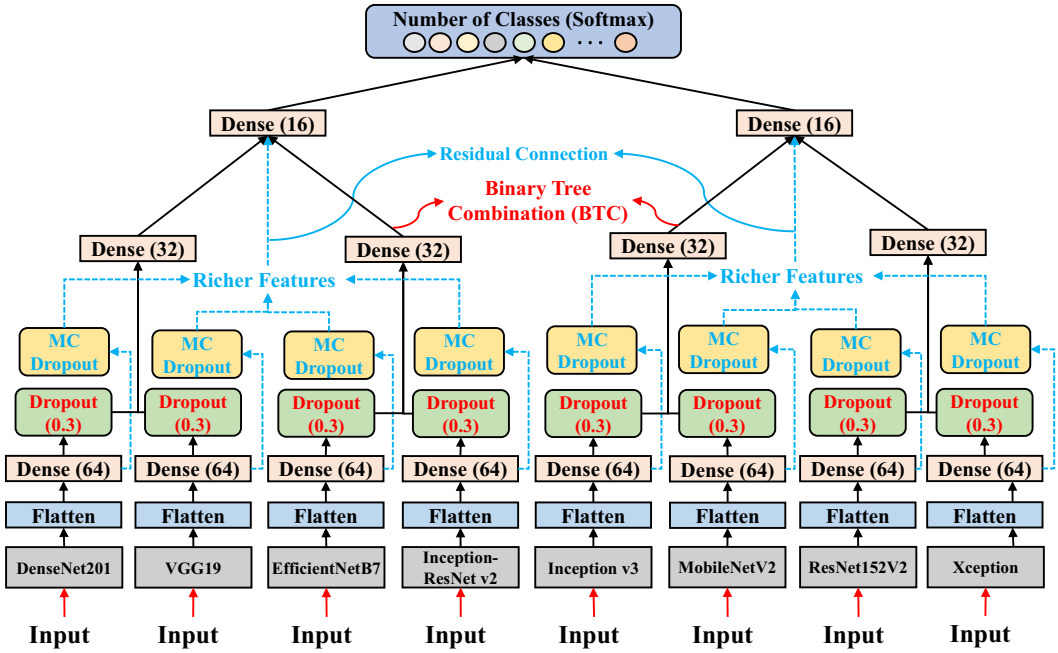


Fig. 6. A schematic view of the proposed direct-based BARF with same input images for all pre-trained models.It may be noted that the activation function in all dense layers is ReLU function.

In the following, we have provided the main characteristics of our proposed direct-based BARF model.

- Pre-trained models on ImageNet dataset: pre-trained models on ImageNet dataset have always been one of the best options to extract the features hidden in different kinds of images. We have used 8 different pre-trained models in the proposed direct/cross based BARF models. Each pre-trained model extracts one feature map from the given input medical image. The feature maps of different models are fused in the remainder of the architecture. Since the ImageNet weights of pre-trained models are achieved by 3-channel of colored images, we need to convert the gray-scale images to 3-channel images. This approach helps to remain the same weights of pre-trained models meaningful. By putting a gray-scale image three times together, we achieved one 3-channel image. This technique is used to convert gray-scale images to 3-channel images [37].
- Feature fusion as ensemble of different deep neural networks (DNNs): One of the best ways to take advantage of different DNN components is to combine them creatively. We have designed a feature fusion-based model as an ensemble architecture in which we have combined the properties of eight different deep neural networks. It should be noted that this step is the first key point in UQ in our proposed model as it acts like an ensemble.
- Binary Tree Combination (BTC): The general architecture of our deep ensemble model is inspired by the structure of binary trees. The tree structure of the model allows us to utilize different kinds of CNN combinations. The key advantages in designing our BARF model are: (i) using the properties of double and quadruple combinations of pre-trained models at the same time, and (ii) taking advantage of residual blocks (connections). It should be noted that the main backbone of the proposed BARF model follows the BTC approach. BTC also allows us to employ both standard dropout (applying dropout just at the training time) and probabilistic dropouts (applying dropout at the training and inference time) at the same time at the same level of tree-based model.
- Direct residual fusion components: We have added two residual fusion blocks to our model. The advantages of residual modules are proven in the context of DNNs. This is the main and only difference in our proposed direct and cross models. In the direct model there are two residual layers in the two sides of the tree. In the "direct-based BARF" model, each of these residual layers connects the output of the probabilistic dropout components of one side of the tree to two levels higher at the same side in the tree structure. Other features at these higher levels come from non-probabilistic (standard) dropout components. Therefore, the residual components help to combine the certain and uncertain results of the dropout level in the tree-based structure of the model.
- Uncertainty awareness module: As mentioned before, the model has a dropout level which contains two kinds of dropouts. The normal dropouts just run at the training time and the probabilistic dropouts run during both the training and inference steps. These probabilistic dropout components are the second key points in quantifying uncertainty of our proposed direct and cross-based BARF models.

3.2. Proposed Cross-based BARF

The only difference between the second proposed fusion model and the first one is the residual components. As mentioned in the previous sub-section (see 3.1), in both direct and cross models, there is one residual layer in each main side of the tree (there are two main sides). In the direct-based BARF model each layer connects the output of probabilistic dropout components of one side of the tree to two levels higher at the same side of the tree. In the cross-based BARF model, each residual layer connects the output of probabilistic dropout components of one side of the tree to two levels higher at the **another** side of the tree. It helps the destination components of residual layers to take the advantages of based models' properties in both sides of the tree. Using this approach, we mix the features achieved by the pre-trained models in both sides of the tree using the residual connection components. A schematic view of the proposed cross-based BARF is shown in Fig. 7.

4. Experimental outcome

In this section, the experimental outcomes of applied methods using different medical image datasets is presented. We first explain the datasets used and then discuss the experiment setup of current study. Finally, the results obtained using each medical dataset are presented..

4.1. Datasets

In this study, we used four medical image datasets: coronavirus (COVID-19) CT scans[2], chest X-ray images[3], Optical Coherence Tomography (OCT) images[3], and skin cancer dermoscopic images[4]. Table 1 presents the details of each dataset used in our study. As three out of four datasets used in our study are grayscale images, we pre-processed [37] these images before feeding to the pre-trained models. We feed gray-scale image three times along with together to achieve a 3-channel model. More information about the medical datasets is presented in Table 1. Sample image of each dataset is shown in Fig. 8.

4.2. Experiment Setup

The experiments are performed on a Windows-based system with GeForce RTX 2080 Ti. The four datasets described in sub-Section 4.1 are divided into two main groups: training and validation. The coronavirus (COVID-19) CT scans are not officially split to train and validation sets, 80% of the total images are used as training and the rest (20%) for validation. However, we directly used other three datasets as they are originally separated into training and validation categories (official train/-validation split). The performance parameters used to evaluate the prediction are as follows: *recall* (Eq. 14), *precision* (Eq. 15), *F1-score* (Eq. 16), *accuracy* (Eq. 17), and the *area under the curve (AUC)*. In this study, as we used the grayscale images, preprocessing is highly recommended to improve the proposed fusion-based medical image classification model [37].

$$Recall = \frac{TP}{TP + FN}, \tag{14}$$

$$Precision = \frac{TP}{TP + FP}, \tag{15}$$

$$F1 - score = 2 * \frac{Recall * Precision}{Recall + Precision} = \frac{2TP}{2TP + FP + FN}, \tag{16}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{17}$$

where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.

In this study, we employed eight well-known pre-trained models for ex tracing features of raw medical images: DenseNet201, VGG19, EfficientNetB7, Inception-ResNet v2, Inception v3, MobileNetV2, ResNet152V2 and Xception to extract the features from the raw medical images. More details about these pre-trained models are provided in Sections 3.1 and 3.2.

4.3. Results

In this section, we investigate the impact of different deep feature fusion models used for medical image classification with various evaluation metrics. Deep feature extractions are important techniques not only used for medical image classification but for many other applications also. The main reason is that all the extracted features have their own advantages which can be used in a specific domain. Hence, we proposed a simple, efficient yet powerful fusion model. The proposed

[2] Source: https://www.kaggle.com/hgunraj/covidxct
[3] Source: https://data.mendeley.com/datasets/rscbjbr9sj/3
[4] Source: https://www.kaggle.com/fanconic/skin-cancer-malignant-vs-benign

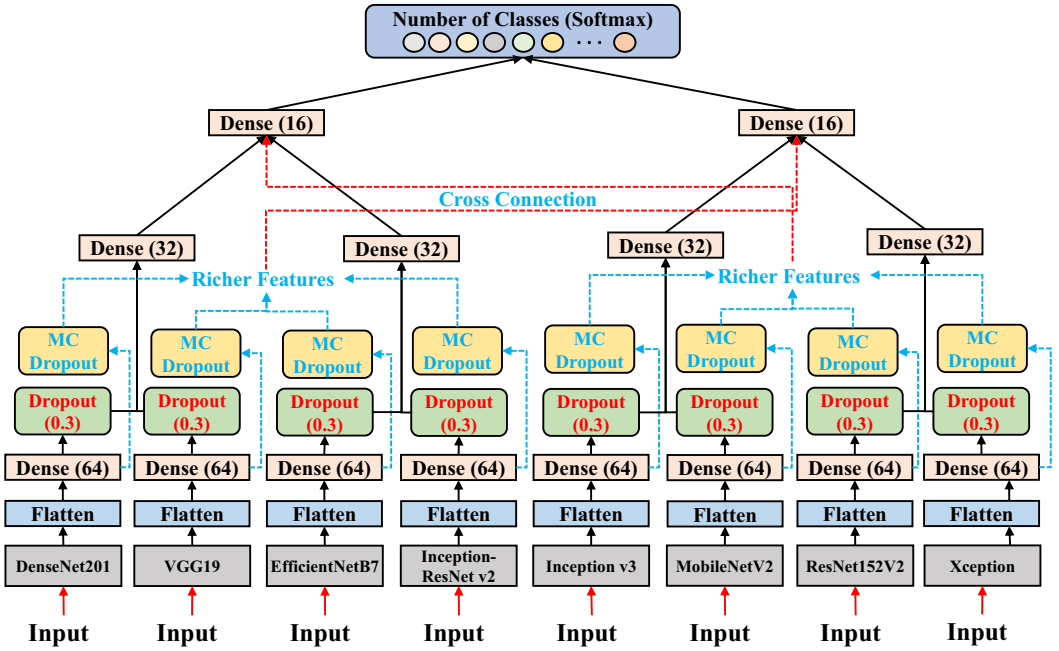


Fig. 7. A schematic view of the proposed cross-based BARF with same input images for all pre-trained models. It may be noted that the activation function in all dense layers is ReLU function.

Table 1
Details of datasets used in this study.

Dataset	Disease/Cancer	# of Samples	# of Classes	Type	Category
CT scan	COVID-19	104009	3	Grayscale	Big data
X-ray	Pneumonia	5856	2	Grayscale	Small data
OCT	Retinal Structural Changes	84484	4	Grayscale	Big data
Dermoscopic	Skin cancer	3297	2	Colory	Small data

BARF model includes two main strategies: direct and cross (see 3.1 and 3.2). To show their performance, we applied early and late fusions of eight pre-trained models and then compared their performances with our proposed direct and cross-based BARF models. All fusion models have included uncertainty quantification module. However, as discussed earlier, the proposed direct and cross-based BARF models benefited from the simultaneous use of standard and MC dropouts, the residual connection helped to provide this option.

In the following, we report the obtained results for each dataset separately. We have first shown the results obtained for COVID-19 (CT scan) dataset at the validation stage in Table 2. It can be noted from the table that, both direct and cross-based BARF fusion models have outperformed the early and late fusion models. However, we noticed that the direct-based BARF fusion model performed slightly better than the cross-based BARF fusion model. Hence, we have chosen direct-based BARF fusion model as the best fusion model for COVID-19 classification in this study [38]. Another important feature of our study is examining the results for each class individually. This advantage allows us to carefully investigate the performance of models for each class. Figs. 9–12 provide the detailed performance metrics (*i.e.*, recall, precision, and F1-score) of the proposed fusion models obtained using CT scans (COVID-19) for both training and validation phases.

In our second experiment, we tested the developed fusion models using X-ray dataset and the obtained results are shown in Table 3. The CT scans (COVID-19) are applied to all four fusion models with UQ module. It can be noted from the table that, the proposed cross-based BARF achieved the best performance followed by early and direct-based BARF fusion models. Figs. 13–16 show the performance of all fusion models tested using X-ray dataset for both training and validation phases using early fusion, late fusion, direct-based BARF, and cross-based BARF, respectively.

Our third grayscale and final medical data is the OCT dataset are used to show the superiority of our both proposed BARF models. The obtained results using OCT dataset is presented in Table 4. The results clearly show that our direct and cross-based BARF models are superior to the other two applied fusion models (Early and late fusion models). Although the performance of our both fusion models are promising, our proposed cross-based BARF model performed better (with the accuracy of 92.50%) than direct-based BARF model with an accuracy of 91.40%. Figs. 17–20 show the performance of all fusion models

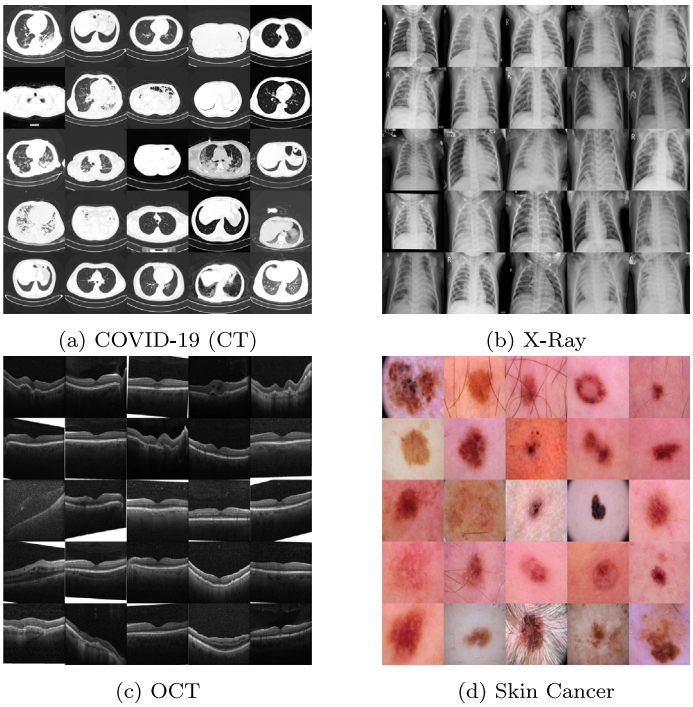


Fig. 8. Sample image of each dataset: (a) COVID-19 (CT), (b) X-ray, (c) OCT, and (d) skin cancer, respectively.

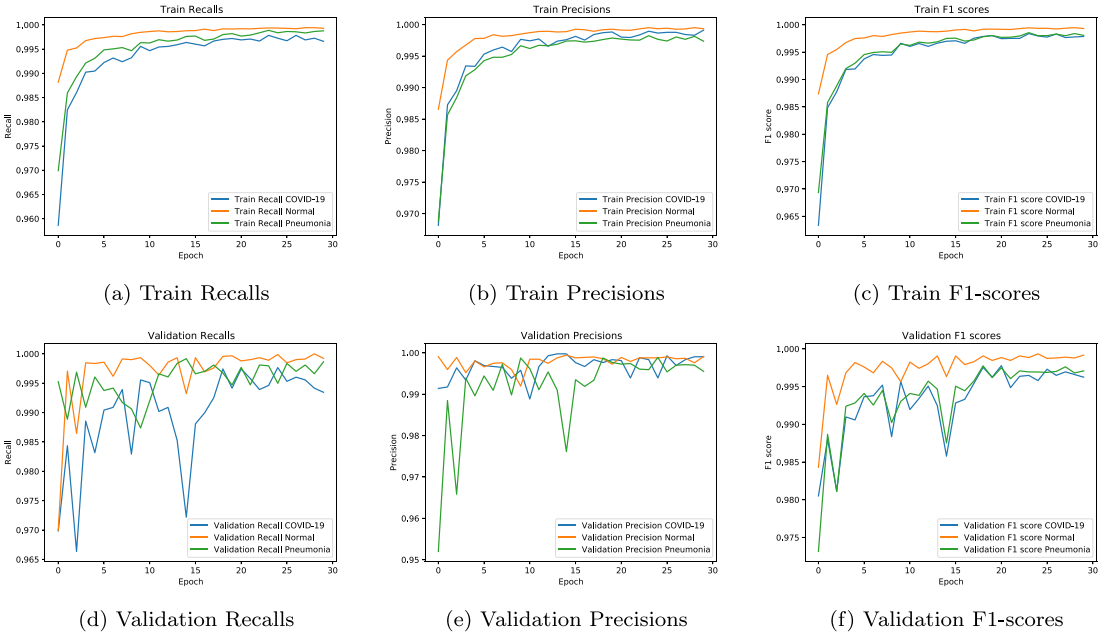
Table 2
Performance comparison with various fusion models obtained using COVID-19 dataset at the validation stage.

Method	Class	Performance				
		Recall (%)	Precision(%)	F1-score (%)	Accuracy (%)	AUC
Early fusion	COVID-19	99.35	99.91	99.62	-	-
	Normal	99.92	99.91	99.91	-	-
	Pneumonia	99.86	99.55	99.70	-	-
	Average	99.71	99.79	99.74	99.78	0.9994
Late fusion	COVID-19	83.48	100	90.99	-	-
	Normal	99.99	96.74	98.33	-	-
	Pneumonia	94.13	99.81	96.88	-	-
	Average	92.53	98.85	95.40	97.62	0.9936
Direct-based BARF (ours)	COVID-19	99.84	99.95	99.89	-	-
	Normal	99.99	99.92	99.95	-	-
	Pneumonia	99.92	99.93	99.92	-	-
	Average	99.91	99.93	99.92	99.93	0.9997
Cross-based BARF (ours)	COVID-19	99.81	99.98	99.89	-	-
	Normal	99.99	99.91	99.94	-	-
	Pneumonia	99.85	99.89	99.86	-	-
	Average	99.88	99.92	99.89	99.92	0.9997

tested on OCT dataset for both training and validation phases using early fusion, late fusion, direct-based BARF, and cross-based BARF, respectively.

Finally, we investigated the performance of all fusion models using colored medical images (*i.e.*, skin cancer [39]). As discussed above, unlike previous three datasets we have a 3-channel model for this colored image dataset. The performance of all applied fusion models including UQ module is reported in Table 5. The obtained outcomes show that the direct and cross-based BARF models have achieved the same accuracy of 89.24%, precision (89.11%) and F1-score (89.18%). However, cross-based BARF has performed with better AUC of 0.9422. For this reason, we chose the cross-based BARF as the best model for skin cancer classification dataset. Figs. 21–24 provide the performance of all fusion models tested with skin cancer images dataset for both training and validation phases using early fusion, late fusion, direct-based BARF, and cross-based BARF, respectively.

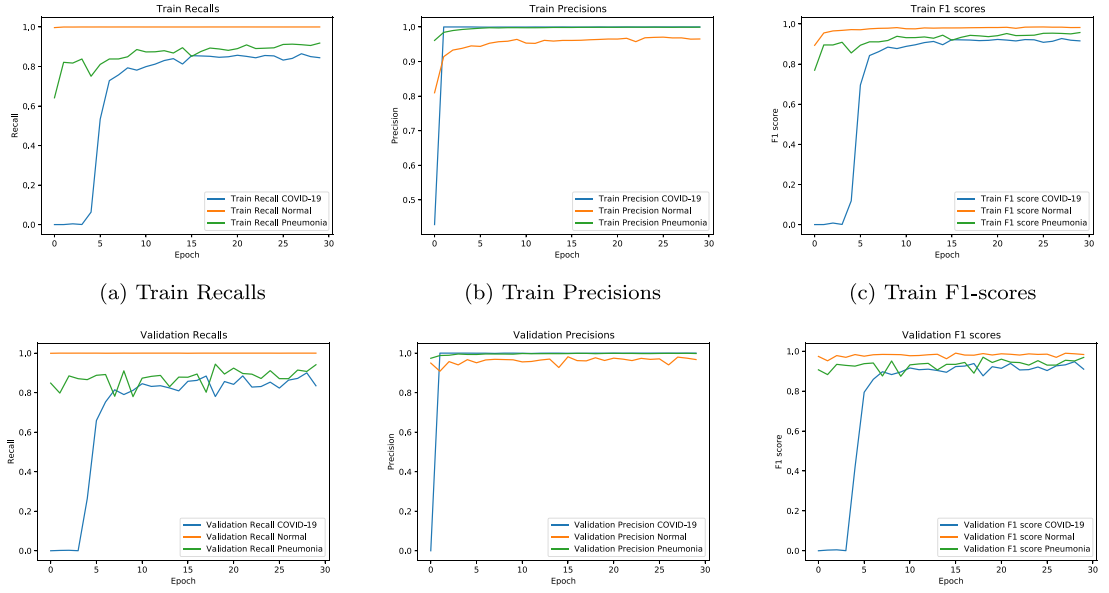
In summary, our results obtained indicate that both proposed BARF models achieved outstanding results. The outcomes indicate that the cross-based BARF model has outperformed the other applied fusion models for three out of four medical



(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 9. Detailed performance metrics of early fusion model tested on CT scans (COVID-19).

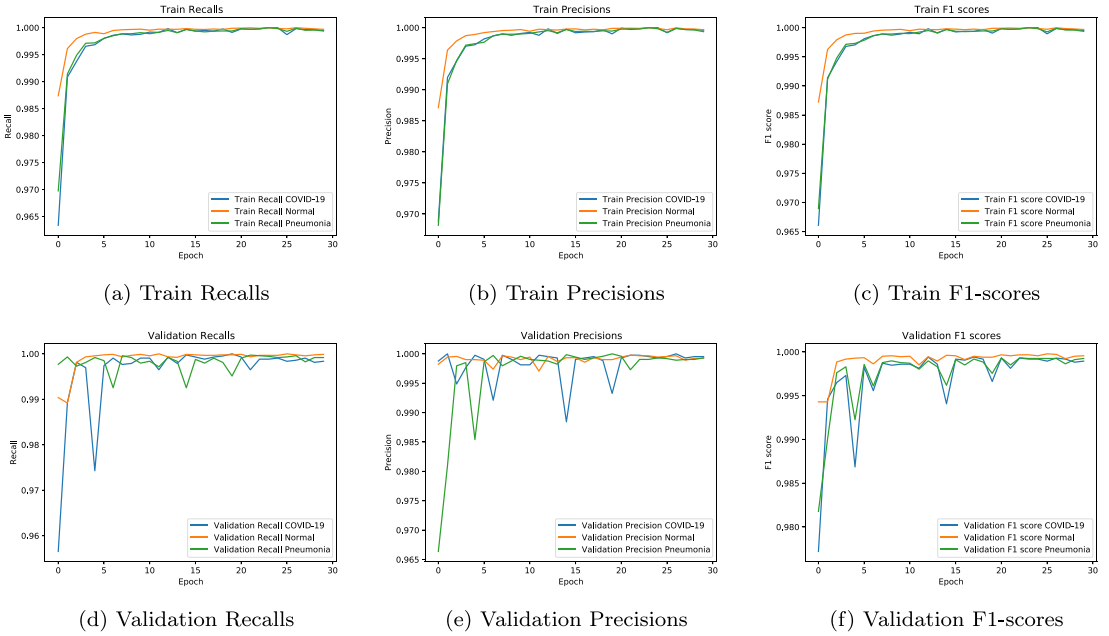


(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 10. Detailed performance metrics of late fusion model tested on CT scans (COVID-19).

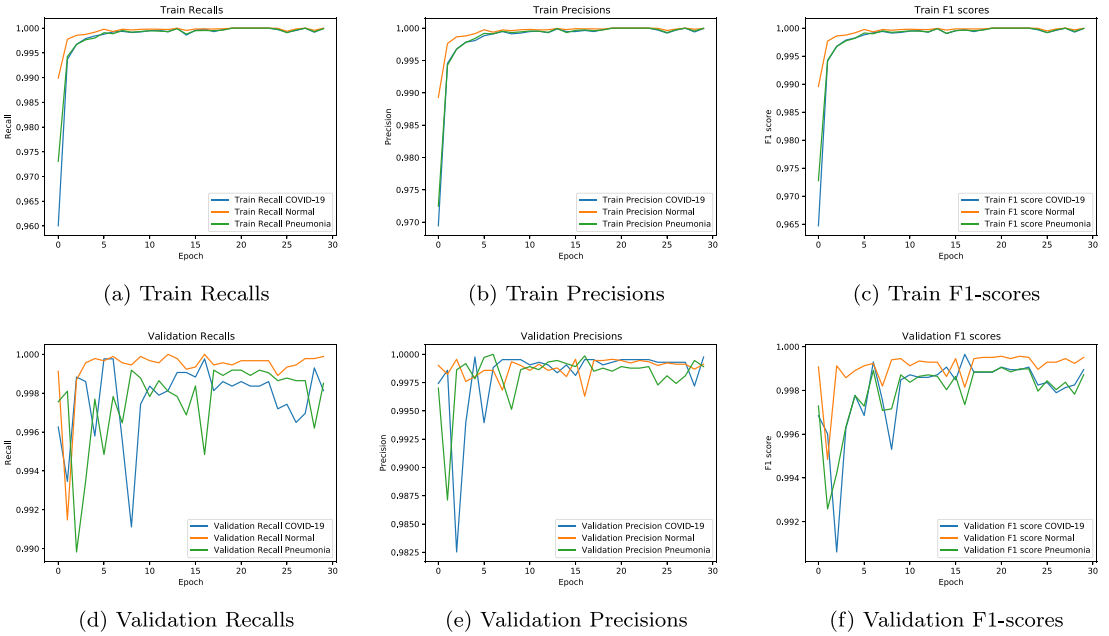
datasets (X-ray, OCT, and skin cancer) used in this study. However, our results reveal that direct-based BARF model obtained slightly better performance for CT scan (COVID-19) classification task. These results confirm the strength of our both proposed fusion models. Our results obtained justifies that the proposed approach is capable of simultaneously use both standard and MC dropouts. This means that the UQ method can be included in the BARF model for many healthcare applications and real scenarios.



(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 11. Detailed performance metrics of the proposed direct-based BARF model tested on CT scans (COVID-19).



(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

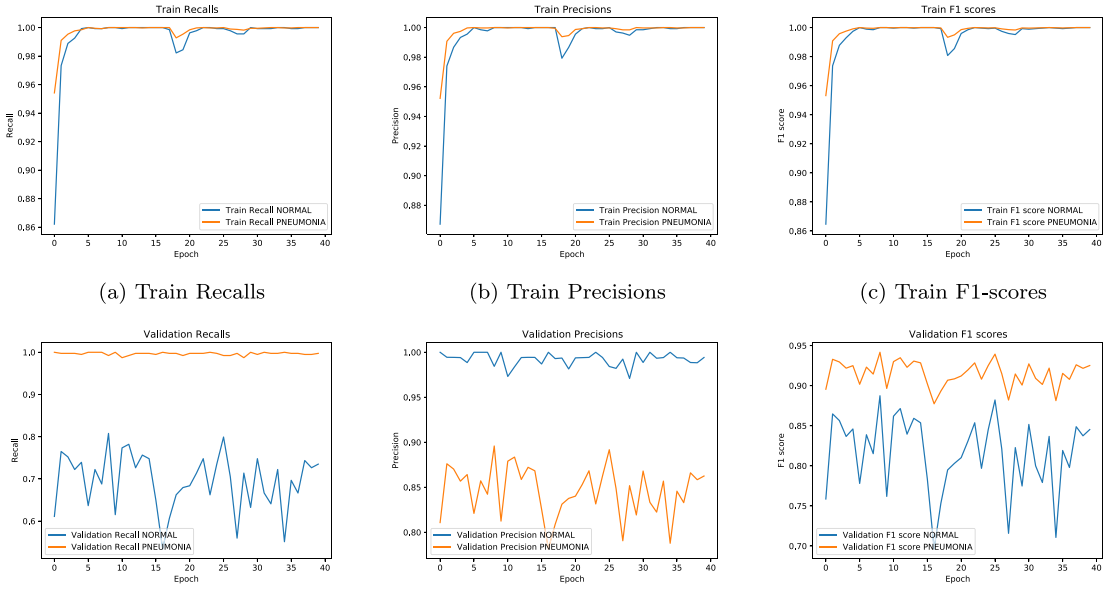
Fig. 12. Detailed performance metrics of the proposed cross-based BARF model tested on CT scans (COVID-19).

5. Discussion

In this study, we introduced a novel and efficient fusion model for medical image classification named BARF using direct and cross strategies. One of the important strengths of BARF is the use of fusion model in feature level like an ensemble approach. This emphasizes the power of collective decision-making rather than individual decision-making to choose the most important and effective features. Various studies have proven (for more information please see [17]) that, hybrid frameworks perform much better than individual methods. Along with better performance, these combined methods are

Table 3
Performance comparison with various fusion models obtained using X-ray dataset at the validation stage.

Method	Class	Performance				
		Recall (%)	Precision(%)	F1-score (%)	Accuracy (%)	AUC
Early fusion	Normal	73.50	99.42	84.51	-	-
	Pneumonia	99.74	86.25	92.50	-	-
	Average	86.62	92.83	88.50	89.90	0.9371
Late fusion	Normal	50.85	100	67.41	-	-
	Pneumonia	100	77.23	87.15	-	-
	Average	75.42	88.61	77.28	81.57	0.8954
Direct-based BARF (ours)	Normal	72.22	100	83.86	-	-
	Pneumonia	100	85.71	92.30	-	-
	Average	86.11	92.85	88.08	89.58	0.9262
Cross-based BARF (ours)	Normal	73.93	100	85.01	-	-
	Pneumonia	100	86.47	92.74	-	-
	Average	86.96	93.23	88.87	90.22	0.9368



(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

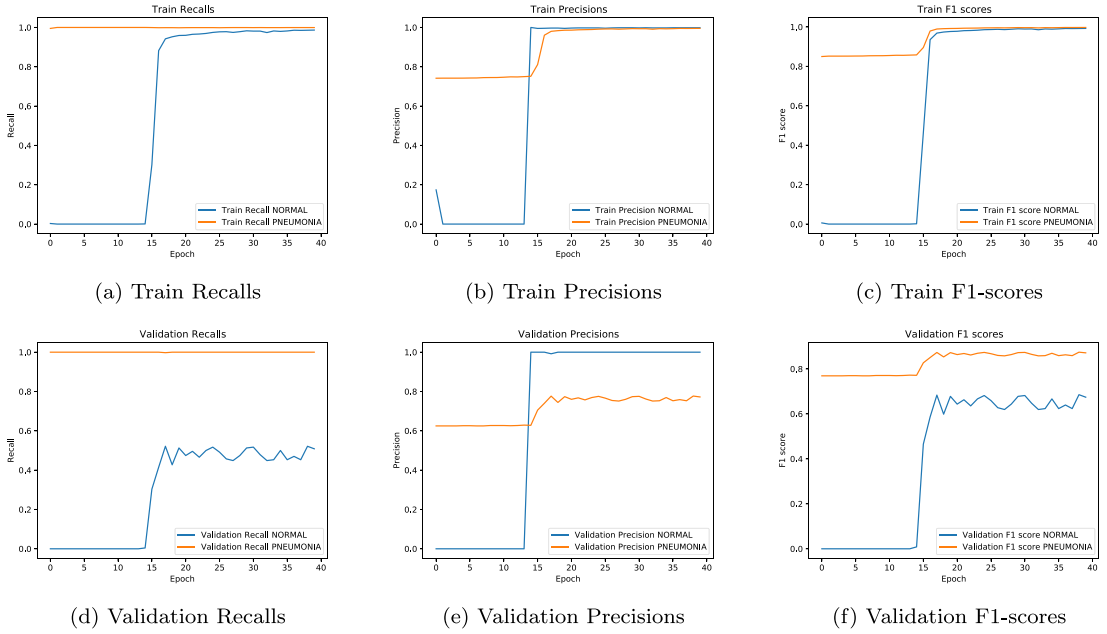
Fig. 13. Detailed performance metrics of early fusion model tested on X-ray images.

also appropriate to deal with uncertainties. These two advantages persuaded us to propose a hybrid system for medical image classification. Uncertainty in the predicted results can put patients and even healthy people's life at risk. This study payed special attention to two important points: (1) getting impressive performance in medical image classification, and (2) determining uncertainties in the predictions obtained using our proposed fusion model.

As shown in Section 4, the proposed direct and cross-based BARF models are tested on four different medical image datasets. The obtained outcomes reveal the outstanding performance by these fusion models. To highlight the superiority of our BARF model, we applied two well-known fusion models namely early and late fusions. In order to provide a fair comparison, since both proposed direct-based BARF and cross-based BARF models included MC dropout, we also added the uncertainty module (*i.e.*, MC dropout) to early fusion and late fusion. The obtained results reported in Section 4 clearly reveal that our both fusion models have obtained impressive performance for various medical data classification.

To highlight the promising results of our proposed fusion models, we provided a detailed comparison with the existing methods reported in the literature (see Table 6). We humbly emphasize that the purpose of this comparison is not to diminish the value of previous studies or methods. Our main goal is to propose a robust and accurate classification system for medical image classification. Table 6 provides a comparative performance obtained using our proposed fusion methods with the previous studies for each dataset.

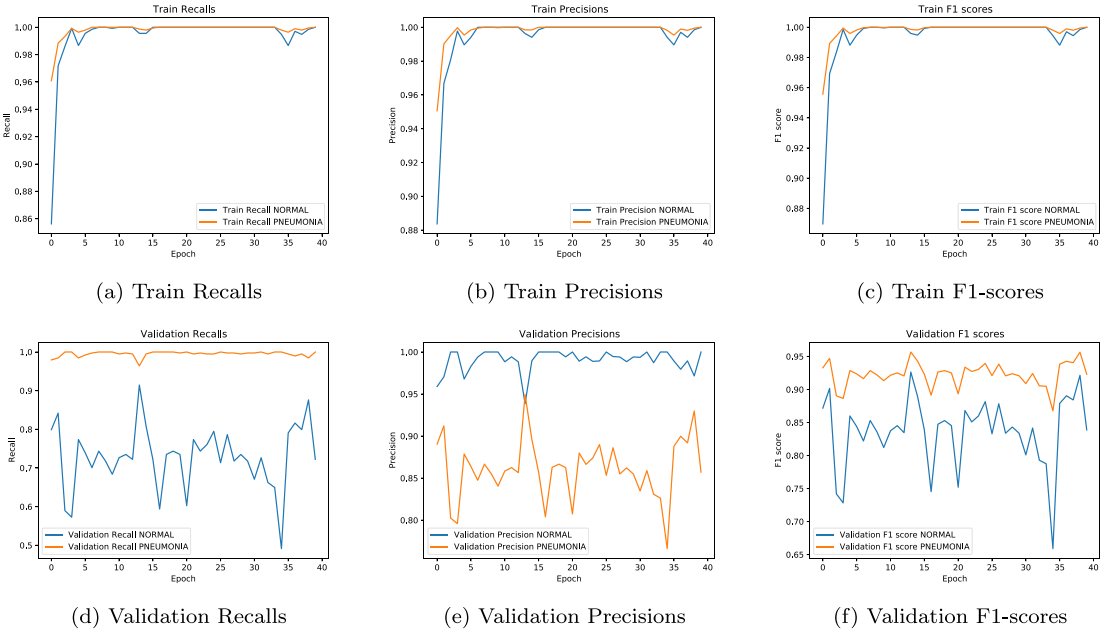
It can be noted from Table 6 that our proposed fusion models have been able to perform reasonably well. We used big datasets of COVID-19 (CT scan) and OCT to highlight the impact of our proposed fusion models. On the other hand, smaller



(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 14. Detailed performance metrics of late fusion model tested on X-ray images.

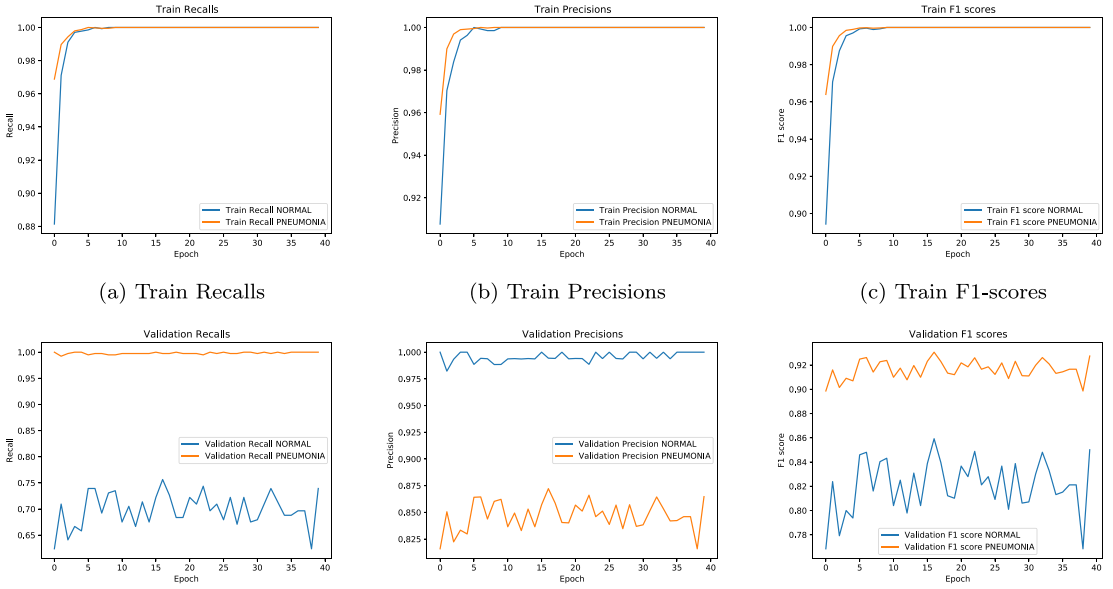


(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 15. Detailed performance metrics of the proposed direct-based BARF model tested on X-ray images.

X-ray and skin cancer datasets are used to evaluate the performance of our direct and cross-based BARF models with smaller medical data. The next important point to be noted in Table 6 is that most of the past studies have avoided considering uncertainty. As we mentioned earlier, dealing with uncertainty is a key point in medical data analysis. Therefore, it can be argued that considering the uncertainty in this study has tremendous advantage. We have obtained good performance (Table 6) using our fusion models for medical image classification. As shown in Table 6, the cross-based BARF model has achieved better performance in 3 out of 4 datasets. Thus, cross-based BARF model performed better than direct-based BARF model.



(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

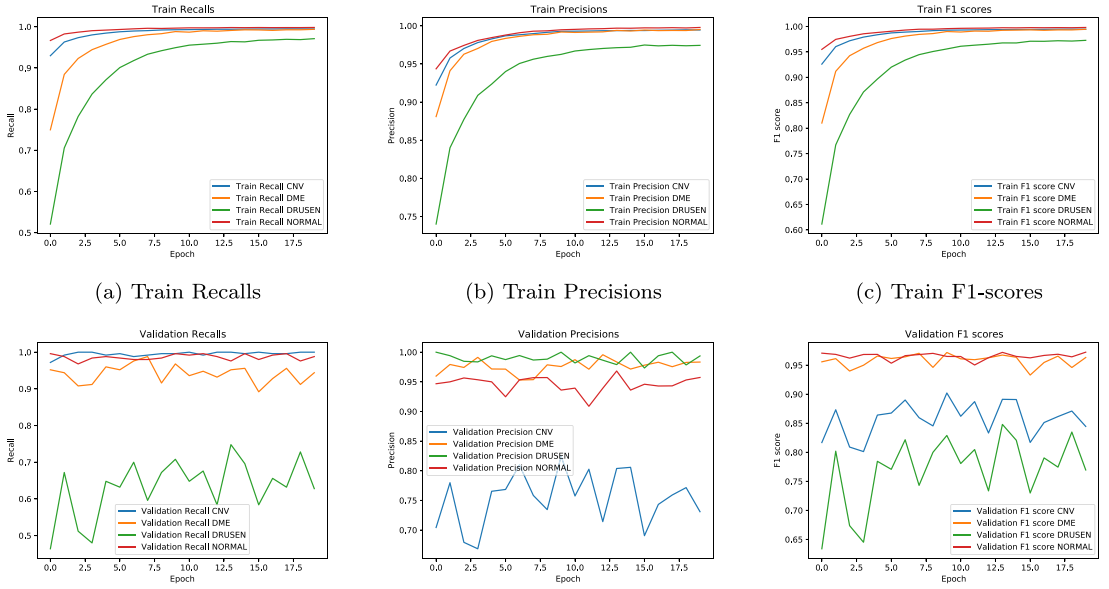
Fig. 16. Detailed performance metrics of the proposed cross-based BARF model tested on X-ray images.

Table 4
Performance comparison with various fusion models obtained using OCT dataset at the validation stage.

Method	Class	Performance				
		Recall (%)	Precision(%)	F1-score (%)	Accuracy (%)	AUC
Early fusion	CNV	100	73.10	84.45	-	-
	DME	94.40	98.33	96.32	-	-
	DRUSEN	62.80	99.37	76.96	-	-
	Normal	98.80	95.74	97.24	-	-
	Average	89.00	91.63	88.74	89.00	0.9736
Late fusion	CNV	100	82.78	90.57	-	-
	DME	95.20	99.58	97.34	-	-
	DRUSEN	41.20	100	58.35	-	-
	Normal	100	91.91	95.78	-	-
	Average	84.10	93.56	85.51	88.30	0.9833
Direct-based BARF (ours)	CNV	100	80.39	89.12	-	-
	DME	95.20	97.14	96.16	-	-
	DRUSEN	72.80	99.45	84.06	-	-
	Normal	97.60	93.85	95.68	-	-
	Average	91.40	92.70	91.25	91.40	0.9821
Cross-based BARF (ours)	CNV	100	81.70	89.92	-	-
	DME	93.20	99.15	96.08	-	-
	DRUSEN	77.20	99.48	86.93	-	-
	Normal	99.20	93.94	96.59	-	-
	Average	92.40	93.56	92.38	92.50	0.9806

For more clarity, sample output posterior distributions of COVID-19, X-ray, OCT, and skin cancer with their corresponding uncertainty estimates using the proposed cross-based BARF are shown in Figs. 25 (COVID-19), 26 (X-ray), 27 (OCT), and 28 (skin cancer), respectively. In these figures, we have shown the posterior distribution of correctly classified (in blue) and misclassified or incorrect (in orange) classes. For example, if the output posterior distributions of correctly classified and misclassified overlap, it means that the model has lot of uncertainty. However, if the output posterior distributions of correctly classified and misclassified do not have any overlap, then, it indicates certainty in predictions.It may be noted that the number of MC samples, *i.e.*, forward passes from dropout which can be used to calculate the predictive mean and uncertainty is 500.

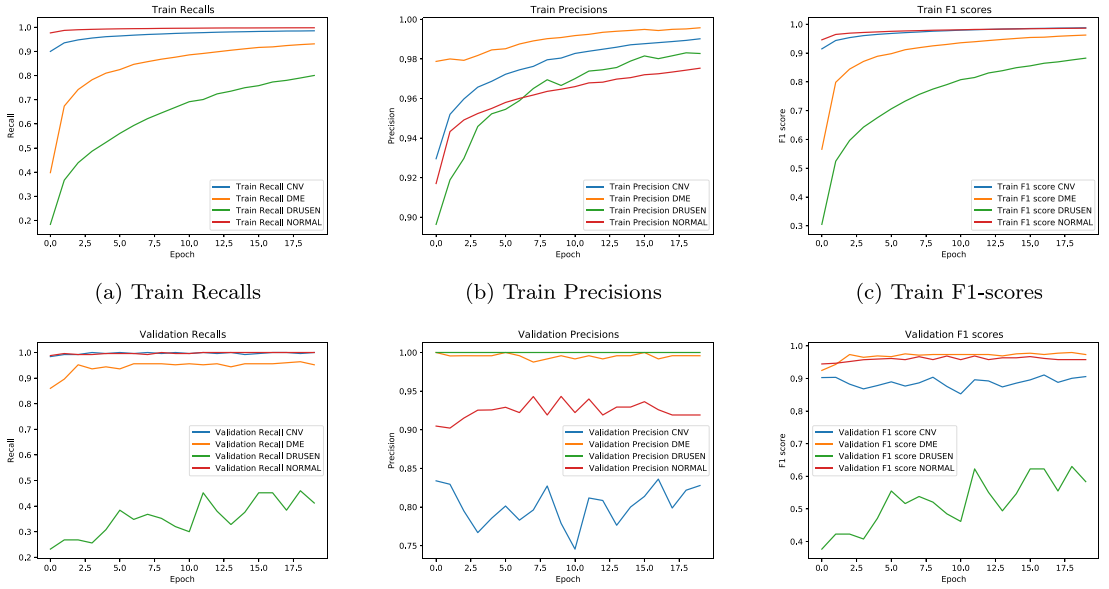
A very motivating point to consider is the ubiquity of ML and DL models for classifying the medical data. The results presented in Tables 2–5 clearly reveal that both early and late fusion models exhibited variable behaviors using different medical data. In other words, the results indicate that in some medical image data early fusion resulted in worst performance



(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 17. Detailed performance metrics of early fusion model tested on OCT images.

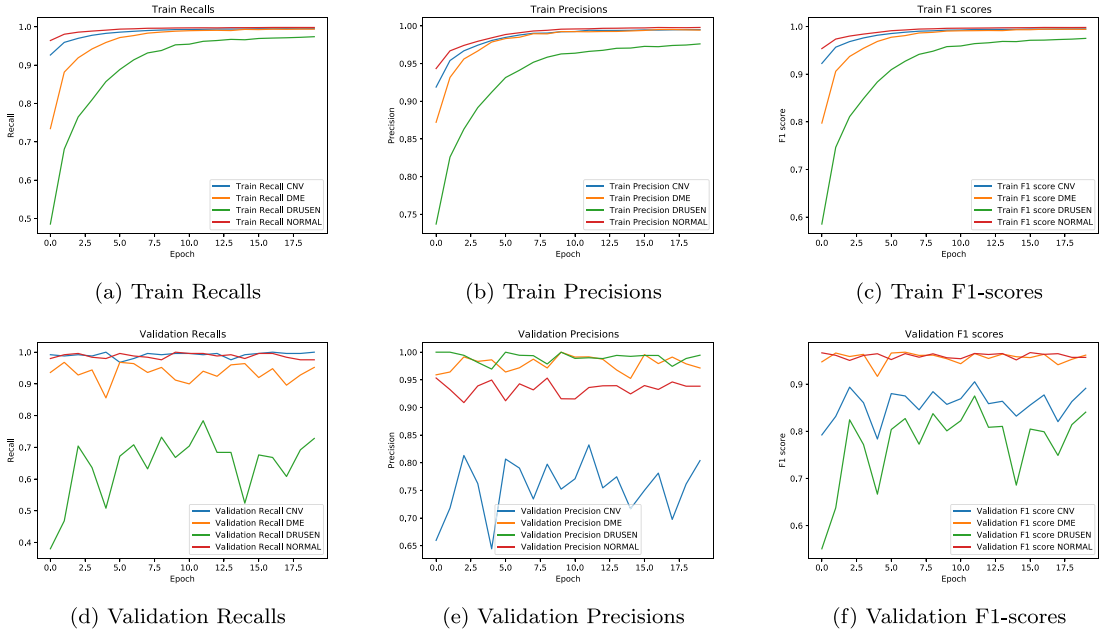


(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 18. Detailed performance metrics of late fusion model tested on OCT images.

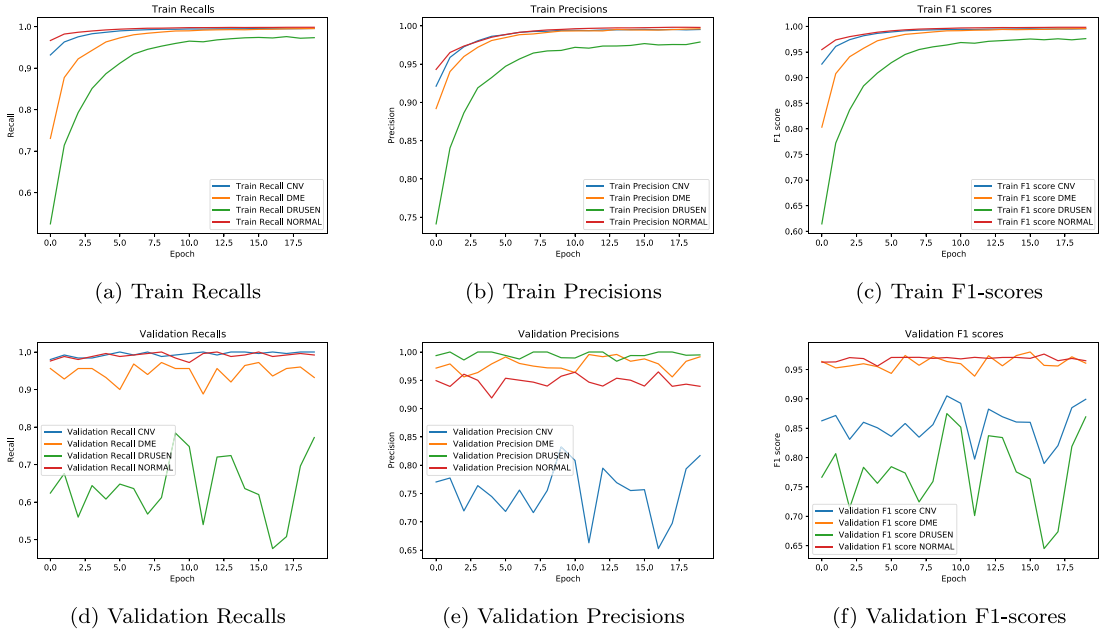
while with other medical image data, late fusion did not perform adequately. However, our both fusion models (*i.e.*, direct and cross-based BARF) showed highest performance and similar behavior using different medical image datasets. It may be noted that our proposed fusion models are more general and comprehensive than the reported previous methods. Also, we have equipped our fusion models with UQ module. Moreover, it can be noted from Table 6 most of the previous methods aimed to achieve better performance and not facing the uncertainty quantification in the models and their predictions. In



(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 19. Detailed performance metrics of the proposed direct-based BARF model tested on OCT images.



(a) Train Recalls (b) Train Precisions (c) Train F1-scores

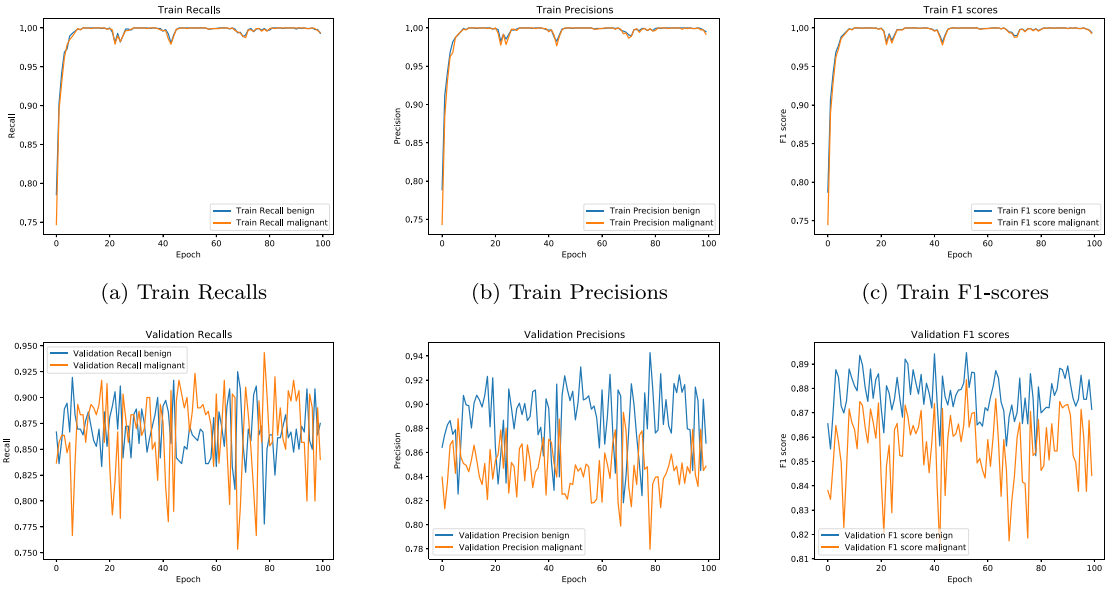
(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 20. Detailed performance metrics of the proposed cross-based BARF model tested on OCT images.

this study, we have focused on both (i) good performance and able to handle uncertainty in the ML and DL models. Our both proposed fusion models can be used as a clinical decision support system (CDSS) for medical image classification in real life scenario. Finally, we have proposed an accurate medical image classification system with high classification performance with UQ, which may be more beneficial for the clinicians and patients.

Table 5
Performance comparison of various fusion models using skin cancer dataset at validation stage.

Method	Class	Performance				
		Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)	AUC
Early fusion	Benign	87.50	86.78	87.13	-	-
	Malignant	84.00	84.85	84.42	-	-
	Average	85.75	85.81	85.77	85.91	0.9053
Late fusion	Benign	86.11	92.81	89.33	-	-
	Malignant	92.00	84.66	88.17	-	-
	Average	89.05	88.73	88.75	88.79	0.9492
Direct-based BARF (ours)	Benign	88.06	91.88	89.92	-	-
	Malignant	90.67	86.35	88.45	-	-
	Average	89.36	89.11	89.18	89.24	0.9217
Cross-based BARF (ours)	Benign	88.61	91.40	89.98	-	-
	Malignant	90.00	86.82	88.38	-	-
	Average	89.30	89.11	89.18	89.24	0.9422



(a) Train Recalls (b) Train Precisions (c) Train F1-scores

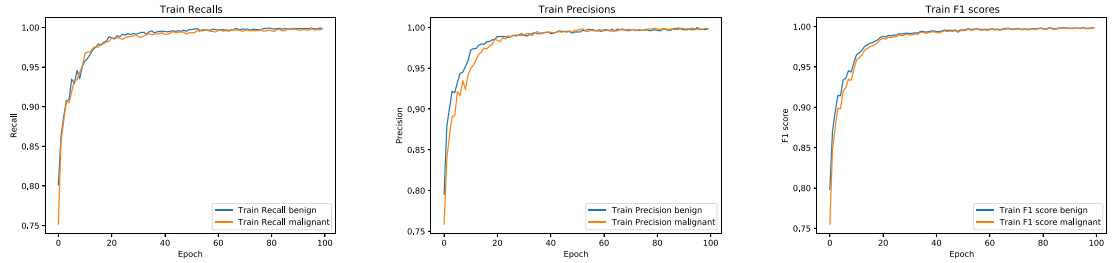
(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 21. Detailed performance metrics of early fusion model tested on skin cancer images.

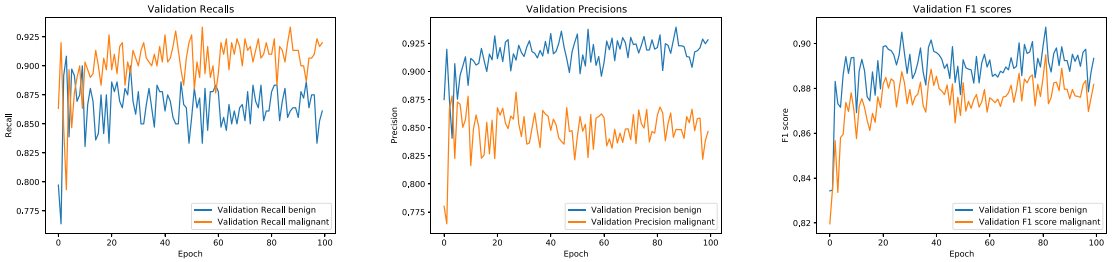
5.1. Major advantages and disadvantages

Each ML and DL method has its own strengths and weaknesses. The main salient features of our developed fusion methods are given below:

- Achieved superior performance for medical image classification.
- Prediction of model uncertainty more accurately;
- Introduction of hybrid system to adopt a multilateral decision-making.
- Proposed a dynamic clinical decision support system for medical image classification with flexibility to choose various base methods and tree depth.
- Evaluation of performance of the proposed model with various medical data.
- Used both binary and multiple classes of medical data;
- Analyzed of results accurately in each class of datasets to ensure the transparency in the final outcomes.
- Generated model is tested with small, big, gray and colored medical image data.
- Evaluated the stability of the model behavior using various medical image data.

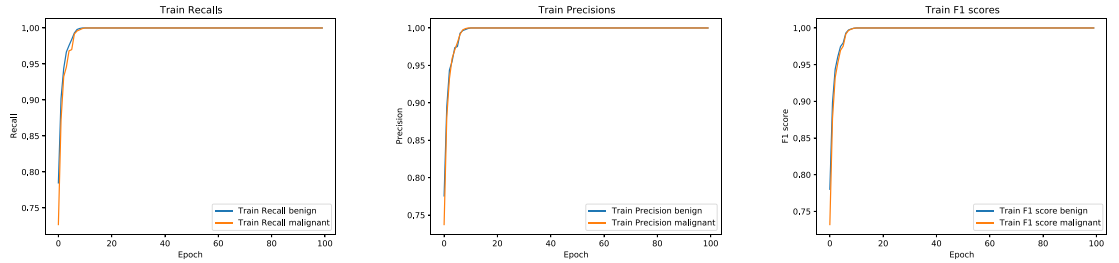


(a) Train Recalls (b) Train Precisions (c) Train F1-scores

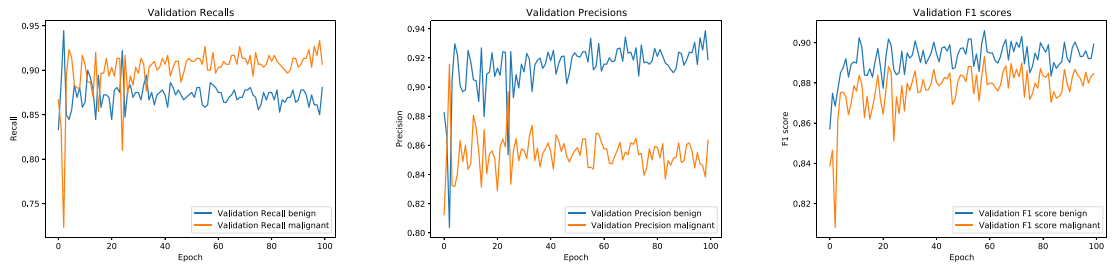


(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 22. Detailed performance metrics of late fusion model tested on skin cancer images.

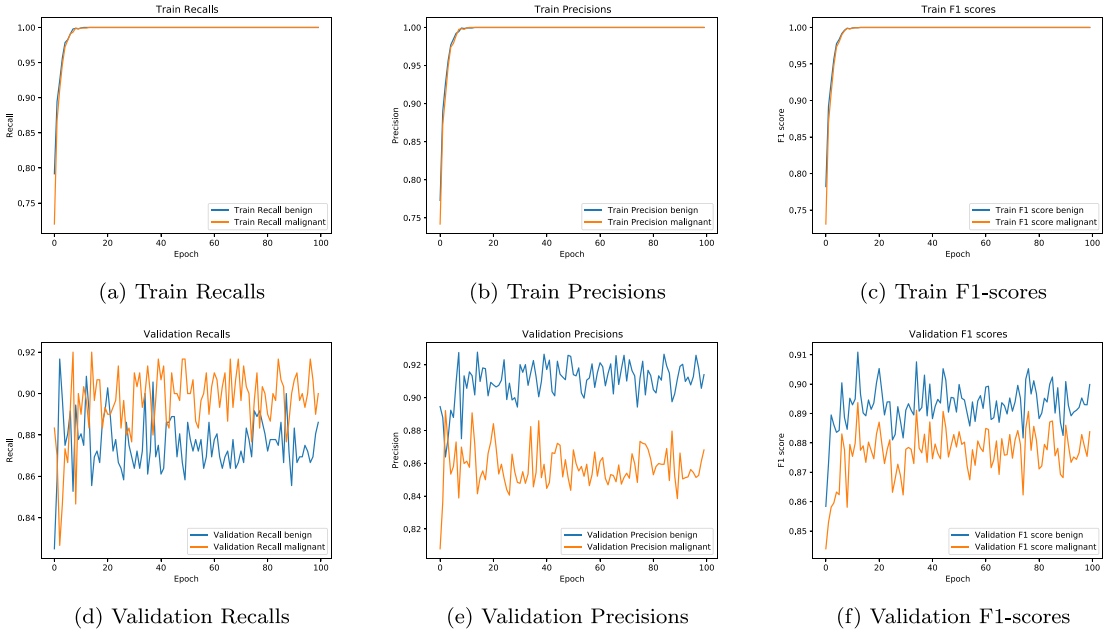


(a) Train Recalls (b) Train Precisions (c) Train F1-scores



(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

Fig. 23. Detailed performance metrics of the proposed direct-based BARF model tested on skin cancer images.



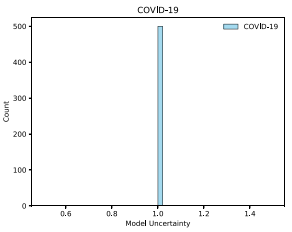
(a) Train Recalls (b) Train Precisions (c) Train F1-scores

(d) Validation Recalls (e) Validation Precisions (f) Validation F1-scores

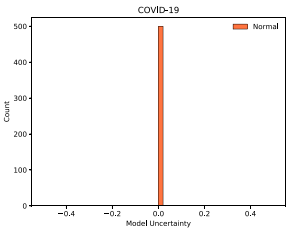
Fig. 24. Detailed performance metrics of the proposed cross-based BARF model tested on skin cancer images.

Table 6
Performance comparison of the proposed BARF model with other existing methods tested on the used datasets (%). Note: MIL: multiple instance learning, Both: Augmentation + Self-supervision, UD-MIL: Uncertainty-Driven Deep Multiple Instance Learning, TWDBDL: Three-Way Decision Bayesian Deep Learning.

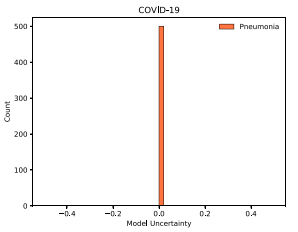
Dataset	Study	Method	# of images	Accuracy	Recall	Precision	F1-score	Uncertainty
COVID-19	Loey et al. [40] (2020)	AlexNet	11012	76.38	63.83	N/A	N/A	NO
	Loey et al. [40] (2020)	ResNet50	11012	81.41	80.85	N/A	N/A	NO
	Loey et al. [40] (2020)	VGGNet16	11012	78.89	62.77	N/A	N/A	NO
	Loey et al. [40] (2020)	VGGNet19	11012	73.87	71.28	N/A	N/A	NO
	Loey et al. [40] (2020)	GoogleNet	11012	77.39	71.28	N/A	N/A	NO
	Gunraj et al. [41] (2020)	ResNet-50	104009	98.70	98.26	98.73	98.49	NO
	Gunraj et al. [41] (2020)	NASNet-A-Mobile	104009	98.60	98.20	98.30	98.24	NO
	Gunraj et al. [41] (2020)	EfficientNet-B0	104009	98.30	97.80	98.30	98.04	NO
	Gunraj et al. [41] (2020)	COVIDNet-CT	104009	99.10	98.76	99.16	98.96	NO
	Bai et al. [42] (2020)	DL	1186	96.00	N/A	N/A	N/A	NO
	Li et al. [43] (2020)	Rubik's cube Pro	2675	N/A	97.70	84.00	90.30	NO
	Abdar et al. [38] (2021)	UncertaintyFuseNet	19685	99.08	99.08	99.08	99.08	YES
	Li et al. [44] (2021)	MIL + Both	229	95.80	93.60	85.74	89.50	NO
	Ours (2021)	BARF (direct)	104009	99.93	99.91	99.93	99.92	YES
X-ray	Liang and Zheng [45] (2020)	CNN	5856	90.50	96.70	89.10	92.70	NO
	Chhikara et al. [46] (2020)	Deep CNN	5866	90.10	95.70	90.70	93.10	NO
	Luján-García et al. [47] (2020)	Xception Network	5232	87.98	99.20	84.30	91.20	NO
	Ours (2021)	BARF (cross)	5856	90.22	86.96	93.23	88.87	YES
OCT	Kermany et al. [48] (2018)	CNN	84484	96.10	96.12	96.10	96.10	NO
	Huang et al. [49] (2019)	Layer Guided CNN	84484	89.90	87.15	87.80	87.47	NO
	Chetoui et al. [50] (2020)	CNN	84484	98.46	98.37	N/A	N/A	NO
	Sunija et al. [51] (2020)	CNN	83484	99.69	99.69	99.69	99.68	NO
	Wang et al. [52] (2020)	UD-MIL	4644	93.30	N/A	N/A	91.70	YES
	Ours (2021)	BARF (cross)	84484	92.50	92.40	93.56	92.38	YES
Skin cancer	Hekler et al. [53] (2019)	Fusion method	12336	N/A	89.00	N/A	N/A	NO
	Bologna and Fossati [54] (2020)	DIMLP-ensemble	3297	84.90	N/A	N/A	N/A	NO
	Lee and Renee [55] (2020)	CNN	6600	82.90	N/A	N/A	N/A	NO
	Abdar et al. [35] (2021)	TWDBDL	3297	88.95	N/A	N/A	89.00	YES
	Ours (2021)	BARF (cross)	3297	89.24	89.30	89.11	89.18	YES



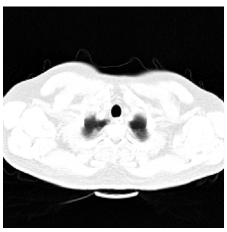
(a) Original image: COVID-19 (b) Correctly classified

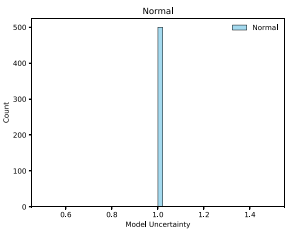


(c) Miscalssified

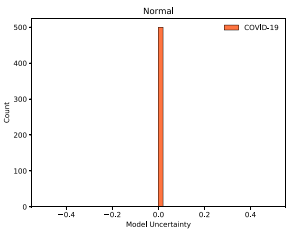


(d) Miscalssified

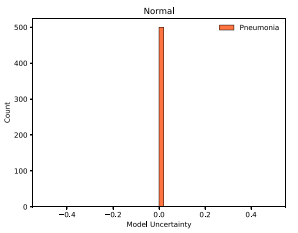




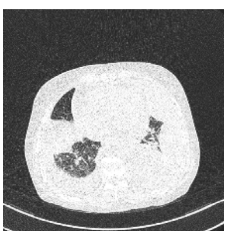
(e) Original image: Normal (f) Correctly classified

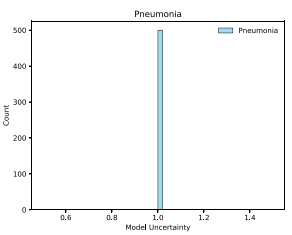


(g) Miscalssified

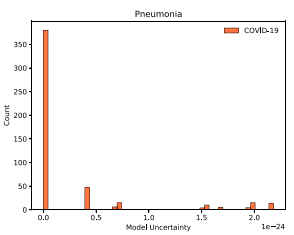


(h) Miscalssified

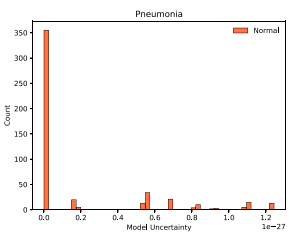




(i) Original image: Pneumonia (j) Correctly classified

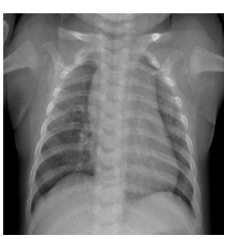


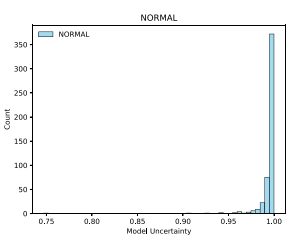
(k) Miscalssified



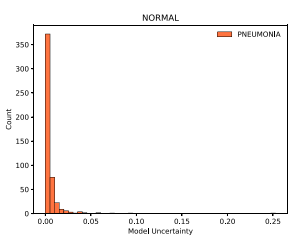
(l) Miscalssified

Fig. 25. Example output posterior distributions of COVID-19 and their corresponding uncertainty estimates using the proposed cross-based BARF.

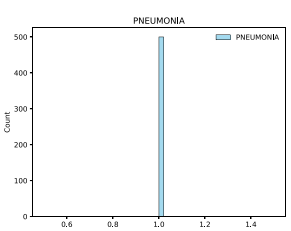




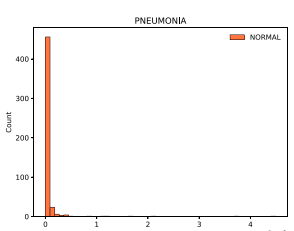
(a) Original image: Normal (b) Correctly classified



(c) Miscalssified

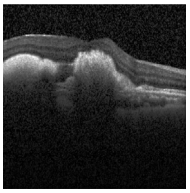
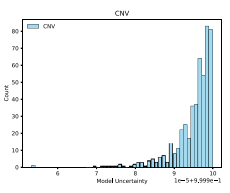


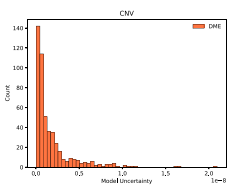
(d) Original image: Pneumonia (e) Correctly classified

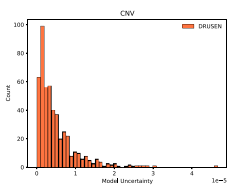


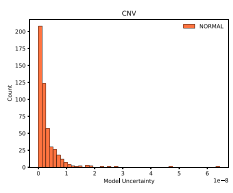
(f) Miscalssified

Fig. 26. Example output posterior distributions of X-ray and their corresponding uncertainty estimates using the proposed cross-based BARF.

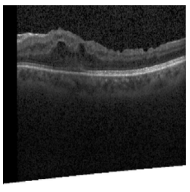
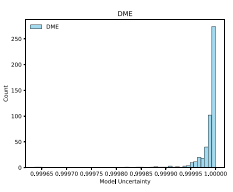



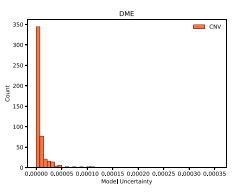


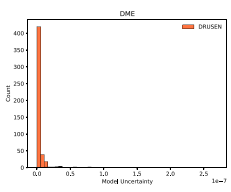


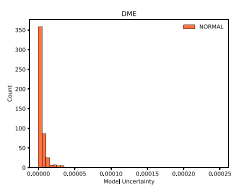


(a) Original image: CNV (b) Correctly classified (c) Miscalssified (d) Miscalssified (e) Miscalssified

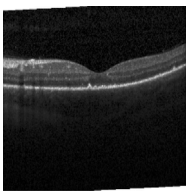
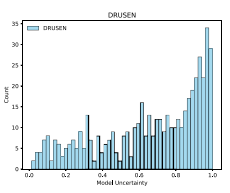



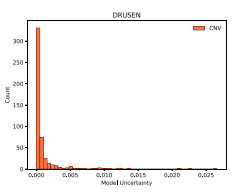


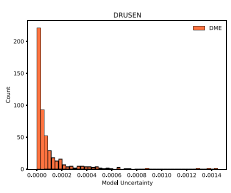


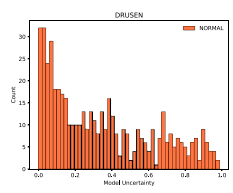


(f) Original image: DME (g) Correctly classified (h) Miscalssified (i) Miscalssified (j) Miscalssified

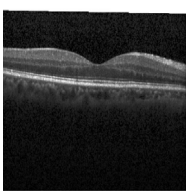
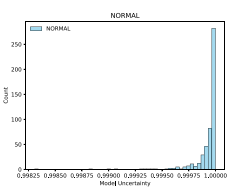



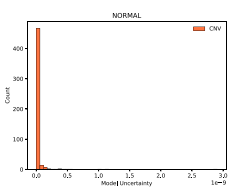


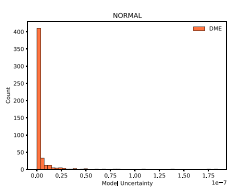


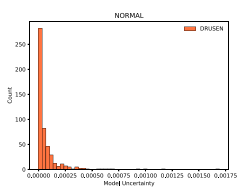


(k) Original image: DRUSEN (l) Correctly classified (m) Miscalssified (n) Miscalssified (o) Miscalssified







(p) Original image: Normal (q) Correctly classified (r) Miscalssified (s) Miscalssified (t) Miscalssified

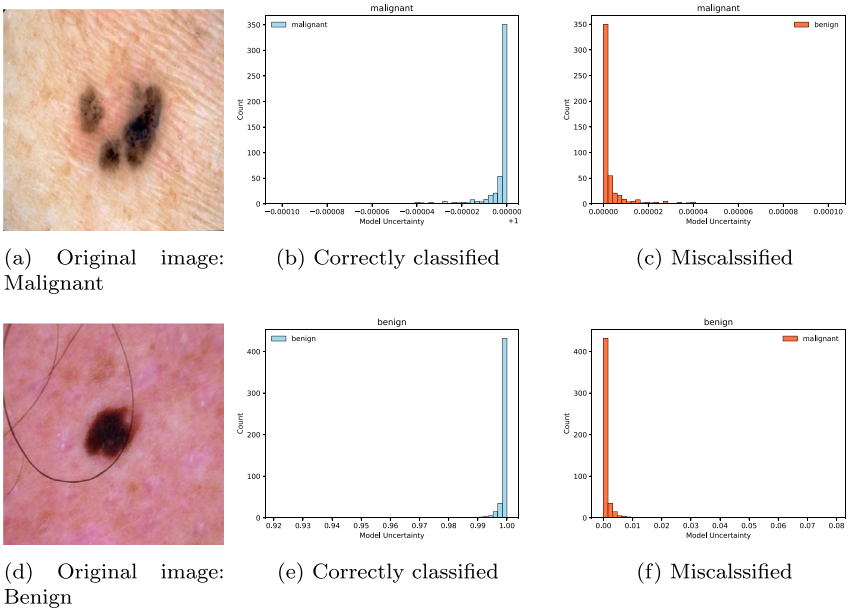
Fig. 27. Example output posterior distributions of OCT and their corresponding uncertainty estimates using the proposed cross-based BARF.

The limitations of our proposed method are also presented below:

- Time-consuming to obtain results using both direct and cross-based BARF methods.
- Unable to return uncertain samples to medical experts caused due to lack of cooperation caused medical team.
- Yields better performance using large medical data than for smaller data.

6. Conclusion

The most recent advancements in medical imaging and artificial intelligence technology has the paved the way for accurate diagnosis by developing novel robust CAD systems. Hence, we proposed a new feature fusion framework to extract the most effective features from various datasets for our proposed BARF medical image classification model using two strategies (direct and cross). Most of existing medical methods do not consider quantifying their uncertainty while predictions. To overcome this weakness, we employed an uncertainty-aware module using Monte Carlo (MC) dropout with our BARF model. Our proposed fusion models validated on all four medical datasets have obtained superior performance and demonstrated increased certainty in the obtained results. In the future, we plan to propose a novel UQ method based on decision making theory. Also, we aim to modify the proposed binary fusion model to have flexibility in choosing base feature extraction methods.



(a) Original image: Malignant (b) Correctly classified (c) Miscalssified

(d) Original image: Benign (e) Correctly classified (f) Miscalssified

Fig. 28. Example output posterior distributions of skin cancer and their corresponding uncertainty estimates using the proposed cross-based BARF.

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.

Acknowledgment

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

References

[1] X. Wang, Y. Zhao, F. Pourpanah, Recent advances in deep learning (2020).
[2] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C.P. Lim, X.-Z. Wang, A review of generalized zero-shot learning methods, arXiv preprint arXiv:2011.08641.
[3] Y. Luo, X. Wang, F. Pourpanah, Dual vaegan: A generative model for generalized zero-shot learning, Appl. Soft Comput. 107352 (2021).
[4] J. Zhang, Y. Xie, Q. Wu, Y. Xia, Medical image classification using synergic deep learning, Med. Image Anal. 54 (2019) 10–19.
[5] M.E. Basiri, M. Abdar, M.A. Cifci, S. Nemati, U.R. Acharya, A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques, Knowl.-Based Syst. 105949 (2020).
[6] M. Abdar, M.E. Basiri, J. Yin, M. Habibnezhad, G. Chi, S. Nemati, S. Asadi, Energy choices in alaska: Mining people's perception and attitudes from geotagged tweets, Renew. Sustain. Energy Rev. 124 (2020) 109781.
[7] J. Wang, Z. He, S. Huang, H. Chen, W. Wang, F. Pourpanah, Fuzzy measure with regularization for gene selection and cancer prediction, Int. J. Mach. Learn. Cybern. (2021) 1–17.
[8] L. Bote-Curiel, S. Munoz-Romero, A. Gerrero-Curieses, J.L. Rojo-Álvarez, Deep learning and big data in healthcare: A double review for critical beginners, Appl. Sci. 9 (11) (2019) 2331.
[9] T. Tirupal, B.C. Mohan, S.S. Kumar, Multimodal medical image fusion techniques–a review, Curr. Signal Transduct. Ther. 15 (1) (2020) 1–22.
[10] Y.-D. Zhang, Z. Dong, S.-H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, et al, Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation, Inform. Fusion 64 (2020) 149–187.
[11] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M.P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, NPJ Digital Med. 3 (1) (2020) 1–9.
[12] K. Liu, Y. Li, N. Xu, P. Natarajan, Learn to combine modalities in multimodal deep learning, arXiv preprint arXiv:1805.11730.
[13] A. Roitberg, T. Pollert, M. Haurilet, M. Martin, R. Stiefelhagen, Analysis of deep fusion strategies for multi-modal gesture recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[14] S. Umer, A. Sardar, B.C. Dhara, R.K. Rout, H.M. Pandey, Person identification using fusion of iris and periocular deep features, Neural Networks 122 (2020) 407–419.
[15] Y. Zhang, L. Gong, L. Fan, P. Ren, Q. Huang, H. Bao, W. Xu, A late fusion cnn for digital matting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7469–7478.
[16] B. Huang, F. Yang, M. Yin, X. Mo, C. Zhong, A review of multimodal medical image fusion techniques, Comput. Math. Methods Med. (2020).
[17] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion.
[18] G. Carneiro, L.Z.C.T. Pu, R. Singh, A. Burt, Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy, Med. Image Anal. 101653 (2020).

[19] Ł. Rczkowski, M. Możejko, J. Zambonelli, E. Szczurek, Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning, Scientific reports 9 (1) (2019) 1–12.
[20] J. Nie, J. Yan, H. Yin, L. Ren, Q. Meng, A multimodality fusion deep neural network and safety test strategy for intelligent vehicles, IEEE Transactions on Intelligent Vehicles.
[21] P. Yi, Z. Wang, K. Jiang, J. Jiang, T. Lu, J. Ma, A progressive fusion generative adversarial network for realistic and consistent video super-resolution, IEEE Transactions on Pattern Analysis and Machine Intelligence.
[22] S.-H. Wang, D.R. Nayak, D.S. Guttery, X. Zhang, Y.-D. Zhang, Covid-19 classification by ccshnet with deep fusion using transfer learning and discriminant correlation analysis, Inform. Fusion 68 (2020) 131–148.
[23] S. Albawi, T.A. Mohammed, S. Al-Zawi, Understanding of a convolutional neural network, in: 2017 International Conference on Engineering and Technology (ICET), IEEE, 2017, pp. 1–6.
[24] J. Qiu, J. Wang, S. Yao, K. Guo, B. Li, E. Zhou, J. Yu, T. Tang, N. Xu, S. Song, et al, Going deeper with embedded fpga platform for convolutional neural network, in: Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, 2016, pp. 26–35.
[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[26] F. Gao, T. Wu, J. Li, B. Zheng, L. Ruan, D. Shang, B. Patel, Sd-cnn: A shallow-deep cnn for improved breast cancer diagnosis, Comput. Med. Imaging Graph. 70 (2018) 53–62.
[27] N. Tagasovska, D. Lopez-Paz, Single-model uncertainties for deep learning, in: Advances in Neural Information Processing Systems, 2019, pp. 6417–6428.
[28] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in neural information processing systems, 2017, pp. 5574–5584.
[29] R. Tanno, D. Worrall, E. Kaden, A. Ghosh, F. Grussu, A. Bizzi, S.N. Sotiropoulos, A. Criminisi, D.C. Alexander, Uncertainty quantification in deep learning for safer neuroimage enhancement, arXiv preprint arXiv:1907.13418.
[30] T.S. Salem, H. Langseth, H. Ramampiaro, Prediction intervals: Split normal mixture from quality-driven deep ensembles, in: Conference on Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 1179–1187.
[31] J. Postels, H. Blum, C. Cadena, R. Siegwart, L. Van Gool, F. Tombari, Quantifying aleatoric and epistemic uncertainty using density estimation in latent space, arXiv preprint arXiv:2012.03082.
[32] 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.
[33] H.-I. Suk, S.-W. Lee, D. Shen, A.D.N. Initiative, et al, Deep ensemble learning of sparse regression models for brain disease diagnosis, Med. Image Anal. 37 (2017) 101–113.
[34] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, S. Udluft, Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 1184–1193.
[35] M. Abdar, M. Samami, S.D. Mahmoodabad, T. Doan, B. Mazoure, R. Hashemifesharaki, L. Liu, A. Khosravi, U.R. Acharya, V. Makarenkov, et al, Uncertainty quantification in skin cancer classification using three-way decision-based bayesian deep learning, Comput. Biol. Med. 104418 (2021).
[36] C. Stoean, R. Stoean, M. Atencia, M. Abdar, L. Velázquez-Pérez, A. Khosravi, S. Nahavandi, U.R. Acharya, G. Joya, Automated detection of presymptomatic conditions in spinocerebellar ataxia type 2 using monte carlo dropout and deep neural network techniques with electrooculogram signals, Sensors 20 (11) (2020) 3032.
[37] Y. Xie, D. Richmond, Pre-training on grayscale imagenet improves medical image classification, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018.
[38] M. Abdar, S. Salari, S. Qahremani, H.-K. Lam, F. Karray, S. Hussain, A. Khosravi, U.R. Acharya, S. Nahavandi, Uncertaintyfusenet: Robust uncertainty-aware hierarchical feature fusion with ensemble monte carlo dropout for covid-19 detection, arXiv preprint arXiv:2105.08590.
[39] J. Jaworek-Korjakowska, R. Tadeusiewicz, Determination of border irregularity in dermoscopic color images of pigmented skin lesions, in: 2014 36TH Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2014, pp. 6459–6462.
[40] M. Loey, G. Manogaran, N.E.M. Khalifa, A deep transfer learning model with classical data augmentation and cgan to detect covid-19 from chest ct radiography digital images, Neural Comput. Appl. (2020) 1–13.
[41] H. Gunraj, L. Wang, A. Wong, Covidnet-ct: A tailored deep convolutional neural network design for detection of covid-19 cases from chest ct images, Frontiers in Medicine 7.
[42] H.X. Bai, R. Wang, Z. Xiong, B. Hsieh, K. Chang, K. Halsey, T.M.L. Tran, J.W. Choi, D.-C. Wang, L.-B. Shi, et al, Ai augmentation of radiologist performance in distinguishing covid-19 from pneumonia of other etiology on chest ct, Radiology 201491 (2020).
[43] Y. Li, D. Wei, J. Chen, S. Cao, H. Zhou, Y. Zhu, J. Wu, L. Lan, W. Sun, T. Qian, et al, Efficient and effective training of covid-19 classification networks with self-supervised dual-track learning to rank, IEEE J. Biomed. Health Inform. 24 (10) (2020) 2787–2797.
[44] Z. Li, W. Zhao, F. Shi, L. Qi, X. Xie, Y. Wei, Z. Ding, Y. Gao, S. Wu, J. Liu, et al, A novel multiple instance learning framework for covid-19 severity assessment via data augmentation and self-supervised learning, Med. Image Anal. 101978 (2021).
[45] G. Liang, L. Zheng, A transfer learning method with deep residual network for pediatric pneumonia diagnosis, Computer Methods Programs Biomed. 187 (2020) 104964.
[46] P. Chhikara, P. Singh, P. Gupta, T. Bhatia, Deep convolutional neural network with transfer learning for detecting pneumonia on chest x-rays, in: Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals, Springer, 2020, pp. 155–168.
[47] J.E. Luján-García, C. Yáñez-Márquez, Y. Villuendas-Rey, O. Camacho-Nieto, A transfer learning method for pneumonia classification and visualization, Appl. Sci. 10 (8) (2020) 2908.
[48] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al, Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131.
[49] L. Huang, X. He, L. Fang, H. Rabbani, X. Chen, Automatic classification of retinal optical coherence tomography images with layer guided convolutional neural network, IEEE Signal Process. Lett. 26 (7) (2019) 1026–1030.
[50] M. Chetoui, M.A. Akhloufi, Deep retinal diseases detection and explainability using oct images, in: International Conference on Image Analysis and Recognition, Springer, 2020, pp. 358–366.
[51] A. Sunija, S. Kar, S. Gayathri, V.P. Gopi, P. Palanisamy, Octnet: A lightweight cnn for retinal disease classification from optical coherence tomography images, Comput. Methods Programs Biomed. 105877 (2020).
[52] X. Wang, F. Tang, H. Chen, L. Luo, Z. Tang, A.-R. Ran, C.Y. Cheung, P.-A. Heng, Ud-mil: uncertainty-driven deep multiple instance learning for oct image classification, IEEE J. Biomed. Health Inform. 24 (12) (2020) 3431–3442.
[53] A. Hekler, J.S. Utikal, A.H. Enk, A. Hauschild, M. Weichenthal, R.C. Maron, C. Berking, S. Haferkamp, J. Klode, D. Schadendorf, et al, Superior skin cancer classification by the combination of human and artificial intelligence, Eur. J. Cancer 120 (2019) 114–121.
[54] G. Bologna, S. Fossati, A two-step rule-extraction technique for a cnn, Electronics 9 (6) (2020) 990.
[55] K.W. Lee, R.K.Y. Chin, The effectiveness of data augmentation for melanoma skin cancer prediction using convolutional neural networks, in: 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), IEEE, 2020, pp. 1–6.