A modified Bayesian Convolutional Neural Network for Breast Histopathology Image Classification and Uncertainty Quantification

Pushkar Khairnar [a,1], Ponkrshnan Thiagarajan [a,2], Susanta Ghosh [b,3]

[a] *Department of Mechanical Engineering-Engineering Mechanics, Michigan Technological University,MI,USA*
[b] *Mechanical Engineering-Engineering Mechanics and the Center for Data Sciences at the Institute of Computing and Cybersystems, Michigan Technological University, MI, USA*

Abstract

Convolutional neural network (CNN) based classification models have been successfully used on histopathological images for the detection of diseases. Despite its success, CNN may yield erroneous or overfitted results when the data is not sufficiently large or is biased. To overcome these limitations of CNN and to provide uncertainty quantification Bayesian CNN is recently proposed. However, we show that Bayesian–CNN still suffers from inaccuracies, especially in negative predictions. In the present work we extend the Bayesian–CNN to improve accuracy and the rate of convergence. The proposed model is called modified Bayesian–CNN. The novelty of the proposed model lies in an adaptive activation function that contains a learnable parameter for each of the neurons. This adaptive activation function dynamically changes the loss function thereby providing faster convergence and better accuracy. The uncertainties associated with the predictions are obtained since the model learns a probability distribution on the network parameters. It reduces overfitting through an ensemble averaging over networks, which in turn improves accuracy on the unknown data. The proposed model demonstrates significant improvement by nearly eliminating overfitting and remarkably reducing (about 38%) the number of false negative predictions. We found that the proposed model predicts higher uncertainty for images having features of both the classes. The uncertainty in the predictions of individual images can be used to decide when further human–expert intervention is needed. These findings have the potential to advance the state–of–the–art machine learning–based automatic classification for histopathological images.

Keywords: Bayesian Convolutional Neural Networks, Breast Cancer, Histopathological Imaging, Machine Learning, Uncertainty Quantification.

1. Introduction

1.1. Histopathological imaging for Breast cancer

Breast cancer is the most common cause of cancer in women [1, 2, 3]. It has the highest incidence (43.3 per 100 000 population) than any other cancer and the highest mortality rate (15%) of all cancer deaths in women in 2012 [1]. Thus detection and diagnosis of breast cancer are vital in reducing the impact of the disease. Histopathological imaging is considered as the gold standard for breast cancer detection and diagnosis [4]. Histopathology is a diagnostic technique that involves microscopic examinations of tissues to study the sign of a disease. This method preserves the underlying architecture of the tissues thereby providing a significant contribution to the diagnosis of diseases. It is the only way to detect some of the diseases like lymphocytic infiltration of cancer. Histopathological images have a lot of information and structure which makes it highly reliable in the diagnosis of diseases especially in almost all types of cancer [5].

1.2. State–of–the–art machine learning algorithms for Histopathological imaging

With the advancement in digital imaging, computer-aided diagnosis (CAD) is focused in the recent past

[1] email: ppkhairn@mtu.edu
[2] email: thiagara@mtu.edu
[3] email: susantag@mtu.edu

This paper is to be submitted to IEEE for peer review.

[5, 6, 7, 8, 9, 10, 11]. Machine learning algorithms such as conventional neural networks and deep neural networks have shown tremendous potential in CAD applications. Feature extraction in a conventional neural network played a significant role in the predictions of the neural network. Various methods of feature extraction followed by classifiers such as support vector machine, K-nearest neighbor, decision tree, Naive Bayes amongst others were explored in the literature for classification and segmentation of breast cancer images [12, 13, 14, 15, 16, 17, 18]. Though these methods performed reasonably well in terms of accuracy, manual feature extraction was a bottleneck in improving the results.

Owing to their advantage of automatic feature extraction, deep learning algorithms have become powerful tools in medical imaging and diagnosis applications. Convolutional neural networks (CNNs) were successful in learning advanced features from the input images and were demonstrated to perform well in classification, both binary [19, 20, 21] and multiclass classifications [22, 4], and metastasis detection [23] on breast cancer images. Apart from being able to perform classification, CNNs were also successfully implemented in problems involving segmentation [24, 25, 26] and detecting regions of interest which contain in-depth discriminatory information for classification in large whole-slide images [27, 28, 29]. CNNs also provided the scope to implement transfer learning, in which the knowledge gained from training one network is applied to another network to solve a similar problem. Thus, transfer learning was implemented by researchers to improve CNNs performance in breast cancer detection [30, 31, 32]. A grand challenge on breast cancer histology images was conducted to advance the state–of–the–art in classifying these images. It was concluded that CNN was the most successful method to classify these breast cancer images [33]. The top performers [34, 35, 36] in this challenge used the architecture of an existing network such as Resnet [37], Densenet [38], Inception [39], VGG16 [40], etc and pre-trained these networks using ImageNet [41].

Although CNN perform better for image classification than other machine learning techniques in terms of accuracy, their parameters are deterministic and thus can not provide any measure of uncertainty in predictions. Uncertainty is a measure of trust in machine decisions and this metric cannot be neglected in medical diagnosis applications such as breast cancer detection as these involve decisions that affect the lives of human beings [42]. Neural networks in general are prone to overfitting especially when they are learned over a small data set. In addition, predictions based on deterministic estimates might provide incorrect results with high confidence. In order to understand the confidence of predictions and to avoid overfitting problems, uncertainty quantification is important [42]. The Bayesian neural network is an efficient state–of–the–art machine learning technique to quantify uncertainties [43, 44, 45].

1.3. Present work: Modified Bayesian–CNN–based classification model for Histopathological images

In a Bayesian–CNN, the weights and biases are random variables as opposed to deterministic variables used in the CNN. Through the stochastic nature of the parameters, Bayesian–CNN captures the variability in the data set and computes the uncertainties in its predictions. Obtaining probability distributions for the weights can be seen as training an ensemble of networks instead of a single network, where each network has its weights sampled from the learned probability distributions. Towards this, Blundell et al. [43] introduced an efficient algorithm called “Bayes by backprop”. This algorithm can be used to learn the probability distribution on weights and it was compatible with the conventional backpropagation scheme. They have also shown that it is possible to regularise the weights by minimizing a loss function known as the variational free energy and demonstrated that this method of regularisation showed performance comparable with dropout on MNIST classification. Shridhar et al. [44, 45] extended the “Bayes by backprop” algorithm to convolutional neural networks. The proposed Bayesian-CNN architecture was implemented for image classification, image super-resolution, and generative adversarial networks. This method has shown performance as good as conventional CNN and in addition provided uncertainty measures and regularisation. Kendall et al. [46] provided a Bayesian deep learning framework combining aleatoric and epistemic uncertainties. The authors improved the model's performance by 1 to 3% over its deterministic counterpart by reducing the effect of noisy data. The trade–offs between modeling aleatoric and epistemic uncertainties were also studied. Kwon et al. [47] proposed a method of quantifying uncertainties in classification using Bayesian neural networks.

Bayesian–CNNs are gaining popularity due to their ability to provide uncertainties associated with the predictions. Yet, there has been no known work to the authors' knowledge that demonstrates the advantages of Bayesian–CNN over CNN to classify histopathological images. In this work, at first we show the advantages of Bayesian–CNN for classification of breast histopathological images. We found that the Bayesian–CNN improves accuracy and reduces overfitting over CNN in addition to quantifying uncertainties. Our results show that the Bayesian–CNN remarkably reduces the false negative predictions. However, we found there is still a significant number of false negatives. Reducing false negative predictions is singularly important in histopathology (or in any medical imaging applications). Therefore there is a significant scope for improvement in reducing the number of false negative predictions. In the present work, we propose a new model to extend the Bayesian–CNN with the objective of further reducing the number of false negative predictions and to improve its accuracy and convergence. The proposed model is henceforth referred as modified Bayesian–CNN.

The key novelty of the proposed model are two fold: 1) we introduced an adaptive activation function with learnable parameters to replace the non adaptive activation function of the Bayesian–CNN, 2) we have demonstrated the use of uncertainty measures to improve the accuracy. The proposed model significantly decreases the number of false negative predictions over Bayesian–CNN. The proposed model nominally improves the accuracy and rate of convergence. A performance–comparison of the three networks, namely CNN and Bayesian–CNN and the proposed modified Bayesian–CNN, is presented.

The rest of the paper is organized as follows: in section 2 the data set and the methods used in this work are described; in section 3 results obtained in the work and their analysis are presented followed by conclusions in section 4.

2. Data set and methods

2.1. Data acquisition

The breast histopathological images used in this work are from a publicly available data set [48]. These are images containing regions of Invasive Ductal Carcinoma (IDC) which is the most common subtype of all breast cancers [48]. These regions are separated from the whole slide images by pathologists. The original data set consisted of 162 whole mount slide images of Breast Cancer specimens scanned at 40x. From the original whole slide images, 277,524 patches of size 50 x 50 were extracted (198,738 IDC negative and 78,786 IDC positive) and provided as a data set for classification.

2.2. Data Preparation

The classification of the data was carefully done by experienced pathologists providing the ground truth for training. The entire data set was divided into a training set and a testing set (80%-20% split) in our study. The training set and the testing set consists of two classes (IDC positive (1) and IDC negative (0)) each. Data training was carried out on the training set and testing set was used to evaluate the classification and uncertainty quantification performance. The training data is shuffled before training which ensures each data item creates an independent and unbiased change on the model. The image size is 3 x 50 x 50 (D x H x W), where D is the depth (color channels), H is the height, and W the width. The images were stored in the form of arrays compatible with the software and network architecture. The images were converted from uint8 to float format for normalizing, as uint8 type arrays are compatible only with integers. For most of the images, the majority of the pixel values are greater than 200. This makes the computation extremely expensive especially with the large size of data. Further, it leads to problems such as singularity/gradient explosion during the evaluation of loss function. In addition, the original images are such that the information that needs to be highlighted has low pixel values. To solve these problems, we computed the complement of all the images (training and testing) and then used as inputs to the neural network. Fig. 1 shows the comparison of original and processed images.

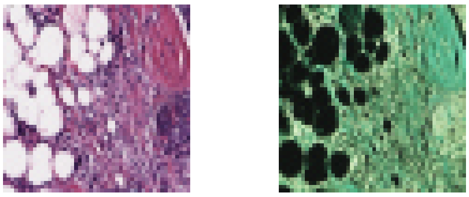

(a) (b)

Figure 1: (a) Original image of training data ; (b) complementary image of training data

2.3. Methods

An overview of methods used in Bayesian Neural Networks and Bayesian Convolution neural networks that are developed in [43] and [44] respectively are provided in this section. Bayesian convolution neural network presented in [44] is the extension of the concepts of Bayesian neural network presented by Blundell et al. in [43]. The mathematical background in this section is revisited for completeness.

The neural network can be defined as a probabilistic model $P(y|x, w)$, where $x \in \mathbb{R}^p$ is an input to the network, $y \in \Upsilon$ are each possible outputs for which the neural network assigns probability, trained with the set of parameters w.

This set of parameters w is learned using a complete Bayesian approach. In this approach given training data D, the posterior distribution of weights $P(w|D)$ is calculated using Bayesian inference for neural networks which involves marginalization over all possible values of w. Once the posterior distribution $P(w|D)$ is obtained, predictions on the unseen data are obtained by taking expectations on the predictive distributions. The predictive distribution of an unknown label $\widehat{y}$ of a test data item $\widehat{x}$ is given by $P(\widehat{y}|\widehat{x}) = \mathbb{E}_{P(w|D)}[P(\widehat{y}|\widehat{x}, w)] = \int_{\Omega_w} P(\widehat{y}|\widehat{x}, w)P(w|D)dw$.
To estimate the posterior distribution we use Bayes' Rule which gives:

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)} \tag{1}$$

The term $P(D|w)$ the likelihood of the training data (D) given a parameter setting (w). Assuming each training data to be independent and identically distributed, the above term becomes the product of likelihood, $P(D|w) = \prod_{n=1}^{N} P(y^n|w, x^n)$ where, (x^n, y^n) are the training data item and known label respectively.
The prior $P(w)$ is our belief about the distribution of weights without seeing the data. The term $P(D)$ in the equation (1) is intractable which makes the posterior distribution $P(w|D)$ intractable . The term $P(D)$ involves marginalization over the weight distribution: $P(D) = \int_{\Omega_w} P(D|w)P(w)dw$.

For variational inference, the posterior distribution $P(w|D)$ which is intractable is approximated with a tractable simpler distribution over the model weights $q(w)$, with variational parameters θ, where $\theta = (\mu, \sigma)$ if we assume our weight distributions to be Gaussian. Therefore, each of the weight (parameter) of the neural network will be parameterized by two parameters, the mean μ and the standard deviation σ, doubling the number of parameters. We fit the variational parameters θ such that $q(w|\theta) \approx P(w|D)$. The variational posterior $q(w|\theta)$ is used instead of the intractable posterior $P(w|D)$ for the inference, $P(\widehat{y}|\widehat{x}) = \mathbb{E}_{q(w|\theta)}[P(\widehat{y}|\widehat{x}, w)]$. Estimating $q(w|\theta) \approx P(w|D)$ we can say that we have learned the distribution of weights given training data.

To make the variational posterior $q(w|\theta)$ and the true posterior $P(w|D)$ similar, we minimize the KL divergence between them:

$$\begin{aligned} KL[q(w|\theta)||P(w|D)] &= \mathbb{E}_{q(w|\theta)}\left[\log\frac{q(w|\theta)}{P(w|D)}\right] \\ = \int q(w|\theta)\log\frac{q(w|\theta)}{P(w)}dw &+ \int q(w|\theta)\log(P(D))dw \\ - \int q(w|\theta)\log(P(D|w))dw & \end{aligned}$$

The term $\log(P(D))$ makes the above equation intractable. Although intractable, it is constant. Therefore, minimizing the KL divergence can be defined as:

$$\begin{aligned} \min\,(KL[q(w|\theta)||P(w|D)]) = \min(KL[q(w|\theta)||P(w)] \\ - \mathbb{E}_{q(w|\theta)}[\log(P(D|w))] \end{aligned}$$

The term $\left(KL[q(w|\theta)||P(w)] - \mathbb{E}_{q(w|\theta)}[\log(P(D|w))]\right)$ is called the Variational Free Energy (VFE) which is to be minimized.

Exact minimization of the cost function is computationally impossible, therefore gradient descent and various other approximations are used. The cost function given above can be approximated through a Monte Carlo sampling procedure as follows:

$$\mathcal{F}(D, \theta) \approx \sum_{i=1}^{n} [\log q\left(w^{(i)}|\theta\right) - \log P\left(w^{(i)}\right) - \log P\left(D|w^{(i)}\right)]$$

Where, $w^{(i)}$ denotes the i^{th} Monte Carlo sample drawn from the variational posterior $q(w^{(i)}|\theta)$. Each term in the cost function has weights that are drawn from the variational posterior. To implement and compute the above cost we need three terms: 1) log of variational posterior, 2) log of prior (Gaussian or scale mixture), 3) log-likelihood of the data.
The operation on the cost function works well with mini-batch optimization. The exact loss form on mini batch i is as follows [49]:

$$\begin{aligned} \mathcal{F}\,(D_i, \theta) &= \frac{1}{M} KL[q(w|\theta)|P(w)] - \mathbb{E}_{q(w|\theta)}\,[P\,(D_i|w)] \\ &\approx \frac{1}{M} \log q(w^{(i)}|\theta) - \log P(w^{(i)}) - \log P(D_i|w^{(i)}) \end{aligned}$$

Where, M corresponds to the number of batches, and $\mathcal{F}(D, \theta) = \sum_{i=1}^{M} \mathcal{F}\,(D_i, \theta)$

The variational posterior (q) is composed of independent Gaussian distribution for each parameter. The sample of weights are obtained by sampling the unit Gaussian, shifting and scaling by the mean μ and a standard deviation σ respectively. To ensure that the standard deviation is always non-negative, it is expressed as $\sigma = \text{softplus}(\rho) = \log(1 + \exp(\rho))$, pointwise. Parameters, w, of the variational posterior, q, can be re-written in terms of a parameter–free noise as $[w = \mu + \log(1 + \exp(\rho)) \circ \varepsilon]$ or $[w = \mu + \sigma \circ \varepsilon]$. Where, $\circ$ is point-wise multiplication, ε is defined below [43]. The steps for optimization are as follows:

Optimization steps:

1. Sample $\varepsilon_i^j \sim \mathcal{N}(0, 1)$ $\{i = 1, ..., N$ and $j = 1, ..., M\}$
   N: Number of parameters in the network,
   M: Number of samples drawn from the variational posterior.
2. Let $w_i^j = \mu_i + \log(1 + \exp(\rho_i)) \circ \varepsilon_i^j$.
3. Let $f(\{w_i^j(\mu_i, \rho_i), \mu_i, \rho_i\}_{i=1...N, j=1...M})$
   $= \sum_{j=1}^{M} \sum_{i=1}^{N} \left[\log q(w_i^j|\mu_i, \rho_i) - \log P(w_i^j) P(D|w_i^j) \right]$
4. Calculate the gradient w.r.t the mean μ_i
   $$\frac{\partial f(w_i^j(\mu_i, \rho_i), \mu_i, \rho_i)}{\partial \mu_i} = \sum_{j=1}^{M} \frac{\partial f}{\partial w_i^j} + \frac{\partial f}{\partial \mu_i}$$
5. Calculate the gradient w.r.t the standard deviation parameter ρ_i
   $$\frac{\partial f(w_i^j(\mu_i, \rho_i), \mu_i, \rho_i)}{\partial \rho_i} = \sum_{j=1}^{M} \frac{\partial f}{\partial w_i^j} \frac{\varepsilon_i^j}{1 + \exp(-\rho_i)} + \frac{\partial f}{\partial \rho_i}$$
6. Update the variational parameters:
   $$\mu_i \leftarrow \mu_i - \alpha \frac{\partial f(w_i^j(\mu_i, \rho_i), \mu_i, \rho_i)}{\partial \mu_i}$$
   $$\rho_i \leftarrow \rho_i - \alpha \frac{\partial f(w_i^j(\mu_i, \rho_i), \mu_i, \rho_i)}{\partial \rho_i}$$
   where α is the learning rate

2.3.1. *Bayesian Convolution Neural Networks*

The parameters of CNNs are filters or kernels which are to be learned during training. In the case of a Bayesian–CNN these kernels are represented by probability distributions as shown in Fig. 2.

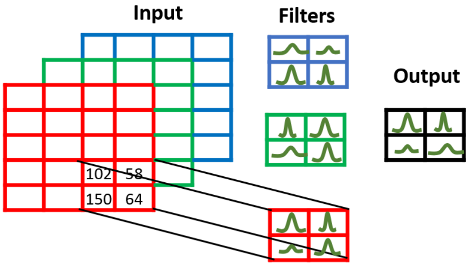


Figure 2: Parameters of the Bayesian–CNN representing probability distributions instead of deterministic values.

During the training of Bayesian–CNN the reparameterization trick is applied on these filters or kernels, which are of shape h×w×d. They are sampled from the variational posterior $q(w|\theta)$ using the following equation:

$$w_{h,w,d} = \mu_{h,w,d} + \log\left(1 + \exp\left(\rho_{h,w,d}\right)\right) \circ \varepsilon_{h,w,d}$$

where h is the height, w is the width and d is the depth of the filter and $\circ$ represents point-wise multiplication. After the sampling from the variational posterior, similar optimization steps as explained in the previous section are followed.

2.3.2. *Uncertainty Quantification*

The uncertainty quantification becomes extremely important when dealing with the applications related to autonomous vehicles, medical imaging, etc. Bayesian deep learning makes it possible to quantify the uncertainties in the prediction as we have probability distribution over weights. Taking an expectation of the predictive posterior probability distribution: $\mathbb{E}_{q(w|\theta)}[P(\widehat{y}|\widehat{x}, w)]$ gives us the most probable prediction of the unknown data $\widehat{x}$. The variance of the predictive posterior probability distribution: $\text{Var}_{q(w|\theta)}[P(\hat{y}|\hat{x}, w)]$ quantifies the uncertainties. There are two types of uncertainties: the Aleatoric and the Epistemic uncertainty. The variance of the predictive posterior probability distribution is the sum of both these uncertainties:

$$\text{Var}_{q(w|\theta)}[P(\widehat{y}|\widehat{x}, w)] = \text{aleatoric} + \text{epistemic}$$

The aleatoric uncertainty corresponds to the noise in the data set whereas the epistemic uncertainty corresponds to the variability of the model developed[44]. One of the promising approaches to quantify the uncertainties is explained in [47]. The uncertainties are obtained from the variance of the predictive posterior probability distribution,

$$\text{Var}_{q(w|\theta)}[P(\widehat{y}|\widehat{x}, w)] = \mathbb{E}_{q(w|\theta)}\left[(y - \mathbb{E}[y])^2\right]$$
$$= \int_{\Omega_w} \left[\left[\text{diag}\left(\mathbb{E}_{P(\widehat{y}|\widehat{x},w)}[\widehat{y}]\right) - \mathbb{E}_{P(\widehat{y}|\widehat{x},w)}[\widehat{y}]\mathbb{E}_{P(\widehat{y}|\widehat{x},w)}[\widehat{y}]^T\right]\right.$$
$$q(w|\theta)dw] + \int_{\Omega_w} \left[\left[\mathbb{E}_{P(\widehat{y}|\widehat{x},w)}[\widehat{y}] - \mathbb{E}_{q_\theta(\widehat{y}|\widehat{x})}[\widehat{y}]\right]\right.$$
$$\left.\left[\mathbb{E}_{P(\widehat{y}|\widehat{x},w)}[\widehat{y}] - \mathbb{E}_{q_\theta(\mathcal{Y}|x)}[\widehat{y}]\right]^T q(w|\theta)dw\right]$$

The above expression is the sum of aleatoric and the epistemic uncertainties which is derived from the variant of the law of the total variance [47]

The first term of the above equation (variance if the predictive posterior probability) is defined as the aleatoric uncertainty and the second term is the epistemic uncertainty. Due to the integral term in the above equations, it is intractable and requires approximations as follows. The equation of the aleatoric uncertainty is defined as: $\frac{1}{N}\sum_{n=1}^{N}\operatorname{diag}\left(\widehat{p}_n\right) - \widehat{p}_n\widehat{p}_n^T$ where, $\widehat{p}_n = p\left(\widehat{w}_n\right) = \operatorname{softmax}\left\{f^{\widehat{w}_t}(\widehat{x})\right\}$. And the equation of the epistemic uncertainty is defined as: $\frac{1}{N}\sum_{n=1}^{N}\left(\widehat{p}_n - \overline{p}\right)\left(\widehat{p}_n - \overline{p}\right)^T$ where, $\widehat{p}_n = p\left(\widehat{w}_n\right) = \operatorname{softmax}\left\{f^{\hat{W}_t}(\widehat{x})\right\}$ and $\overline{p} = \sum_{n=1}^{N}\frac{\widehat{p}_n}{N}$. The overall equation for the variance is:

$$\operatorname{Var}_{q(w|\theta)}[P(\widehat{y|x}, w)] = \left(\frac{1}{N}\sum_{n=1}^{N}\operatorname{diag}\left(\widehat{p}_n\right) - \widehat{p}_n\widehat{p}_n^T\right) + \left(\frac{1}{N}\sum_{n=1}^{N}\left(\widehat{p}_n - \bar{p}\right)\left(\widehat{p}_n - \bar{p}\right)^T\right)$$

The variance of the predictive distribution can be calculated by the above equations which provide us with the confidence of the network in making predictions for a given image.

2.3.3. Adaptive activation

The weights and bias of a perceptron in a neural network performs a linear transformation of the inputs. The output of this linear transformation is passed to an activation function that determines if a particular perceptron is activated for a given input. Non-linear activation function is a key component of a neural network, which enables it to learn complex functions with a small number of perceptrons. However, this nonlinearity of the activation function is known to introduce problems such as exploding or vanishing gradients. Thus, there is a trade off between the learning capabilities vs training complexities of a perceptron, which can be optimised. To achieve this, we propose a formulation for Bayesian-CNN through introducing a learnable activation function that adapts to the training data. Such adaptivity in the context of other machine learning models has been reported in the literature [50, 51], but not for Bayesian-CNN.

This adaptive activation function contains a parameter that is learnt during the training of the neural network. To this end, a trainable hyperparameter (α) is introduced in the activation functions of a Bayesian–CNN and the resulting network is the modified Bayesian–CNN. The details of the proposed adaptive activation function is given below.

$$\sigma(\alpha f_k(x^{k-1}))$$

where,
$f_k(x^{k-1}) = w^k x^{k-1} + b^k$
σ is the activation function
α is the trainable hyperparamter
w and b are the weights and bias of the k^{th} layer and x^{k-1} is the output from the previous layer of the neural network.

The modified loss function $J(\Theta)$ has an additional parameter α which is to be optimised along with the parameters w and b, i.e $\Theta = (w, b, \alpha)$.
We seek to find

$$\alpha^* = \arg\min(J(\alpha))$$

.
The parameter α is updated by gradient descent as,

$$\alpha^{m+1} = \alpha^m - \eta\nabla_\alpha J^m(\alpha)$$

Where, η is the learning rate.

In this work, an adaptive sigmoid given by $\sigma(x) = \frac{1}{1+e^{-\alpha x}}$ is used.

3. Results and analysis

This section presents the development of an automatically regularized and unbiased classification model for histopathological data through Bayesian–CNN, which is described in the Section 2.3. The model is analyzed and juxtaposed with widely used CNN.

3.1. Training and testing of networks

The entire data set is divided into a training data set and a testing data set, which are used for the training and the testing of the networks respectively. The training is done using the images as inputs and the corresponding known classes (positive and negative) as outputs. The loss function, which is the cross-entropy for CNN or the variational free energy for Bayesian–CNN is evaluated for the training data during each epoch. The training is done repeatedly and the weights of the networks are updated until the loss converges. In the following the details of training and testing for CNN and Bayesian–CNN are described.

3.1.1. Convolutional neural networks

The accuracy for training and testing for each epoch is shown in Fig. 3 for CNN. The accuracy increases with epoch for the training data whereas the accuracy reduces for the testing data after the first few epochs. This problem is known as overfitting. Thus, CNN fails to generalize its predictions to unknown testing data and overfits the training data. The loss for each epoch during training is shown in Fig. 4. At around nine epoch, the testing accuracy starts to fall off compared to the training accuracy, which is long before the loss function has converged. The accuracy upon the convergence of the loss function is considered as the predictive capability of the network.

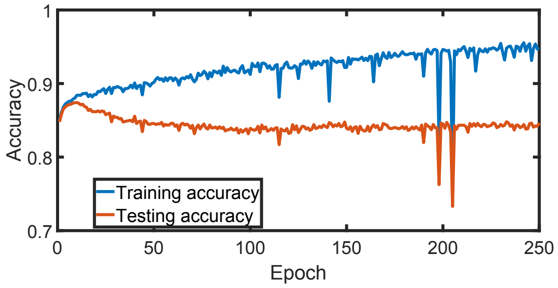


Figure 3: Training and testing accuracy of CNN.

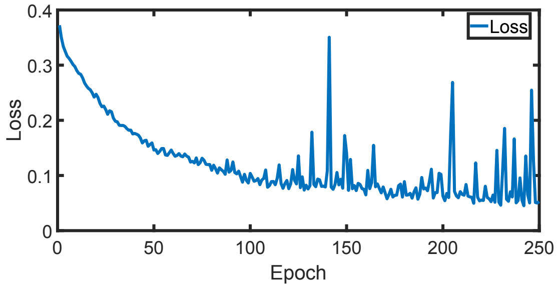


Figure 4: Loss of CNN.

To alleviate the overfitting of the networks, regularization strategies are used, such as $L1$ or $L2$ regularization, cross-validation, early stoppage, and dropout [52, 53, 54]. However, these may require human intervention during training (such as parameter tuning) or may lead to loss of information due to drop out [53]. Herein we use the Bayesian–CNN to overcome the limitations of overfitting automatically through its probabilistic framework.

3.1.2. *Modified Bayesian–CNN*

The accuracy of predictions for both training and testing data of the Bayesian and Modified Bayesian-CNN is shown in Fig. 5 and Fig. 7 respectively. It evident that the problem of overfitting is eliminated by using a Modified Bayesian–CNN as the accuracy of predictions for both training and test data increases with an increase in the number of epochs. The value of the loss at different epochs is shown in Fig. 8. Since the problem of overfitting is eliminated, and the loss is decreasing with increasing epochs, we can use the Modified Bayesian–CNN to obtain a near–optimal solution by training the network for a higher number of epochs.

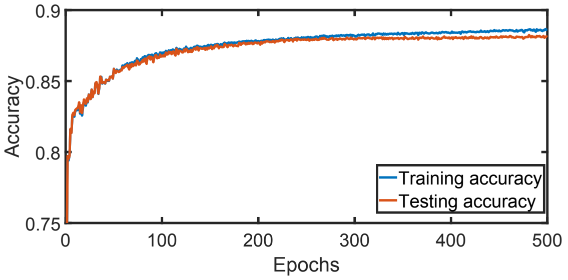


Figure 5: Training and testing accuracy of Bayesian–CNN

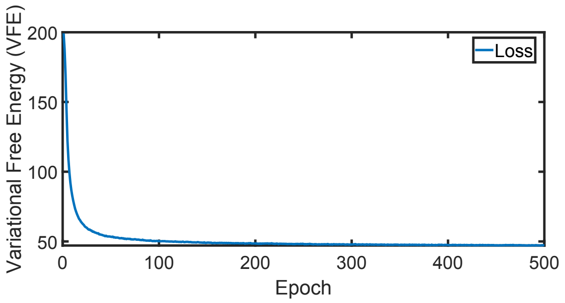


Figure 6: Loss for Bayesian–CNN

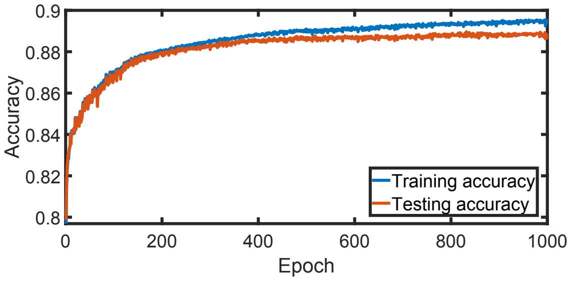


Figure 7: Training and testing accuracy of modified Bayesian–CNN

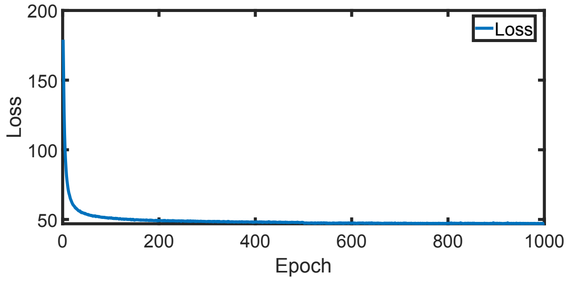


Figure 8: Loss for modified Bayesian–CNN

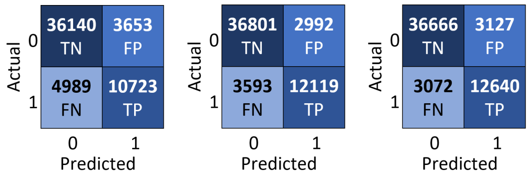
The values of the accuracy of training and testing data upon the convergence of the loss function are provided in Table 1. It is seen that the problem of overfitting drastically reduces the accuracy for test data in a CNN whereas this problem is not present for Bayesian–CNN and the Modified Bayesian–CNN. The training and test accuracy of the Modified Bayesian–CNN is higher than the Bayesian–CNN when the loss function is converged. It is also seen from Fig. 5 and Fig. 7 that the convergence of the Modified Bayesian–CNN is faster as compared to that of the Bayesian–CNN especially during early epochs of training. These are a significant improvement for a nominal increase in the computational cost for the Modified Bayesian–CNN. We found that the Modified Bayesian–CNN requires about twice the computational cost per epoch during the training as compared to CNN.

Table 1: A Comparison of CNN, Bayesian–CNN and Modified Bayesian-CNN

Network	**Training Accuracy**	**Testing Accuracy**
CNN	0.9555	0.8443
Bayesian–CNN	0.8855	0.8813
Modified Bayesian-CNN	0.8948	0.8883

3.2. *Quality assessment metrics*

Often the accuracy of prediction is not a sufficient descriptor to assess the performance of the classifier. So we assess the performance of the classifier through other popular quality assessment metrics, which are described in this section. These performance evaluation metrics use the testing data set which is unseen by the model during training. The confusion matrix is one such metric that presents the performance of a classification model through a tabular layout facilitating visualization of the performance. The four values of the confusion matrix are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).



(a) (b) (c)

Figure 9: Comparison of confusion matrices for a) CNN, b) Bayesian–CNN and Modified Bayesian–CNN

In addition to the improvement of accuracy, the modified Bayesian–CNN remarkably reduces the number of false negative predictions as seen in Fig. 9, which is a significant achievement. CNN is known to perform poorly for biased data set [55]. Given the data set used in this work is highly biased to the IDC negative class, the CNN model fails to overcome this bias yielding more negative predictions. We demonstrate that the modified Bayesian–CNN, overcomes this limitation of the CNN and provides an unbiased result. The modified Bayesian–CNN reduces the false negative predictions by about 38% (10% higher than the Bayesian–CNN) as seen in Fig. 9, demonstrating unbiased prediction. This is a remarkable advancement given that the medical imaging data sets are usually biased towards negative labels. To quantitatively assess the performance of the proposed binary classification the following metrics are used in addition to the confusion matrix: Accuracy, Precision, Recall, F1-score, Cohen's-kappa (CK) coefficient [56]. These metrics can be derived from the confusion matrix, which contains four values based on the actual label and the model prediction. The above metrics are defined by the following equations:

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{Number of images correctly predicted}}{\text{Total number of images}} \\
&= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\
\text{Precision} &= \frac{\text{IDC positive images correctly predicted}}{\text{Total IDC images predicted positive}} \\
&= \frac{\text{TP}}{\text{TP} + \text{FP}}
\end{aligned}$$

$$\begin{aligned}
\text{Recall} &= \frac{\text{IDC positive images correctly predicted}}{\text{Total IDC positive images}} \\
&= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
F1_{\text{score}} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\
&= \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \\
\text{CK Coefficient} &= \frac{p_0 - p_e}{1 - p_e}
\end{aligned}$$

Where,

$$\begin{aligned}
p_0 &= \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}} \\
p_e &= p_\alpha + p_\beta \\
p_\alpha &= \frac{\text{TN} + \text{FP}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}} \times \frac{\text{TN} + \text{FN}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}} \\
p_\beta &= \frac{\text{FN} + \text{TP}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}} \times \frac{\text{FP} + \text{TP}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}}
\end{aligned}$$

The performance of the neural network classifiers is evaluated based on all of the above metrics and the results are provided in table 2. It is seen that the modified Bayesian–CNN over-performs Bayesian–CNN on all metrics except precision and it over-performs CNN on all the performance metrics for this data set.

Table 2: Performance metrics of CNN and BCNN

Network	**Recall**	**Precision**	**F1-score**	**CK**
CNN	0.6824	0.7458	0.7126	0.6062
Bayesian–CNN	0.7713	0.8019	0.7863	0.7042
Modified Bayesian–CNN	0.8045	0.8017	0.8031	0.7251

3.3. Uncertainty quantification

Uncertainty quantification is a measure of how confident the network is in making predictions for the input images. Especially in medical applications that involve machine-assisted decision making, uncertainty quantification can help in building the trust needed to make the right decisions. Since the parameters of CNN are deterministic, it can not provide any measure of uncertainty in predictions for individual images. Thus along with other metrics described in the previous section, uncertainty quantification becomes an important measure that is estimated through the modified Bayesian–CNN in this work.

The parameters of the modified Bayesian–CNN are probability distributions whose variance provides an estimate of the uncertainties associated with the predictions. The convergence of the variational probability distribution of a modified Bayesian–CNN parameter with its actual mean and shifted mean (to zero) is shown for different epochs in Fig. 10.

We see from Fig. 10 that the mean along with the standard deviation changes with increasing epochs. The changes in the mean of the distribution are the same as the frequentist approach when point estimates are used for the parameters. The convergence of variance of the distribution decides the certainty or confidence of the network. The wider (large variance) the distribution is, the more uncertain the model is. To study the change in the model parameter's variance, we shifted the parameter's distribution to zero mean. As the epoch increases, the standard deviation decreases making the parameter more certain about the prediction.

The convergence of the standard deviation of a weight parameter drawn from the variational posterior $q(w|\theta)$ is shown in Fig. 11. We can see that the standard deviation decreases as the network is trained.

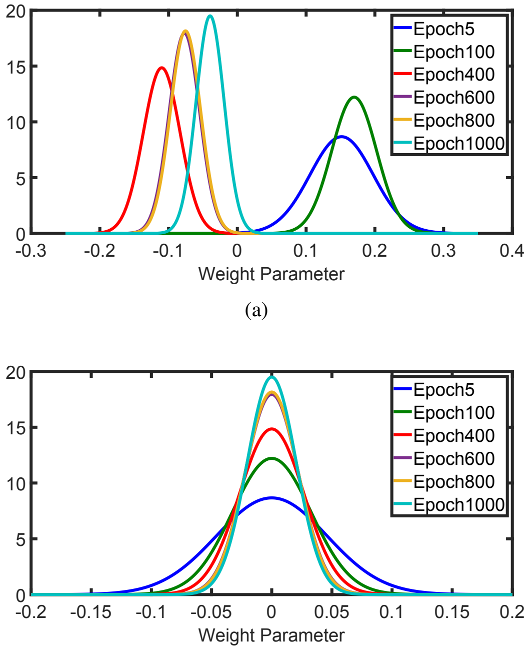


Figure 10: a) Probability density functions b) Shifted (zero mean) Probability density of a parameter for different epochs of the modified Bayesian-CNN.

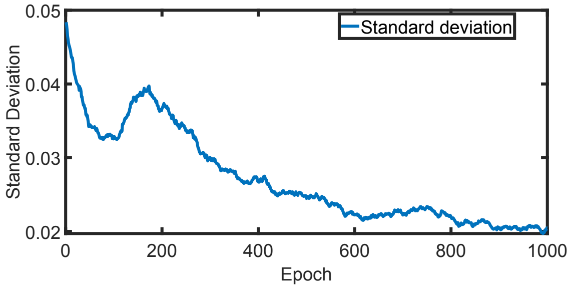


Figure 11: Figure to show the standard deviation of a weight parameter over different epochs.

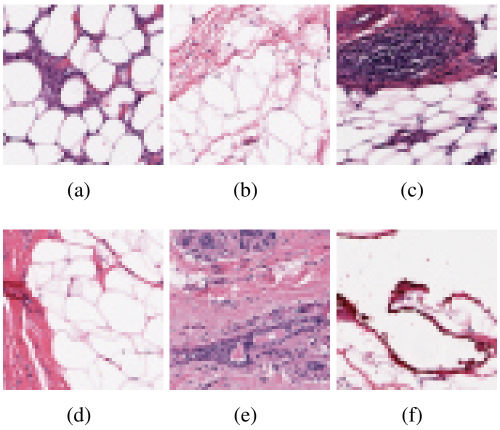


Figure 12: Sample images showing the predictions of the model. Aleatoric and epistemic uncertainties are given by A and E respectively. a) False positive [$A = 4.05 \times 10^{-1}$, $E = 2.48 \times 10^{-2}$], b) True Negative [$A = 2.32 \times 10^{-2}$, $E = 3.97 \times 10^{-4}$], c) True Negative [$A = 4.25 \times 10^{-1}$, $E = 3.61 \times 10^{-2}$], d) True Negative [$A = 2.92 \times 10^{-2}$, $E = 6.74 \times 10^{-5}$], e) False Positive [$A = 4.67 \times 10^{-1}$, $E = 3.19 \times 10^{-2}$], f) True Negative [$A = 2.53 \times 10^{-3}$, $E = 9.17 \times 10^{-6}$]

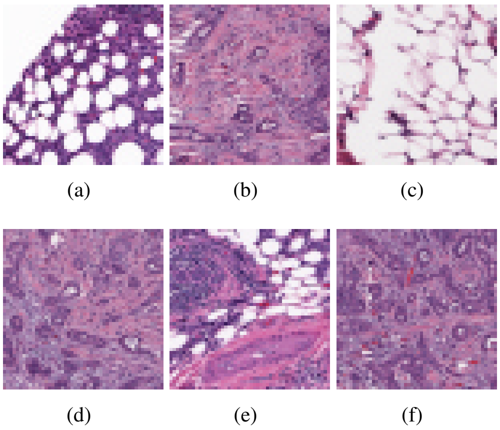
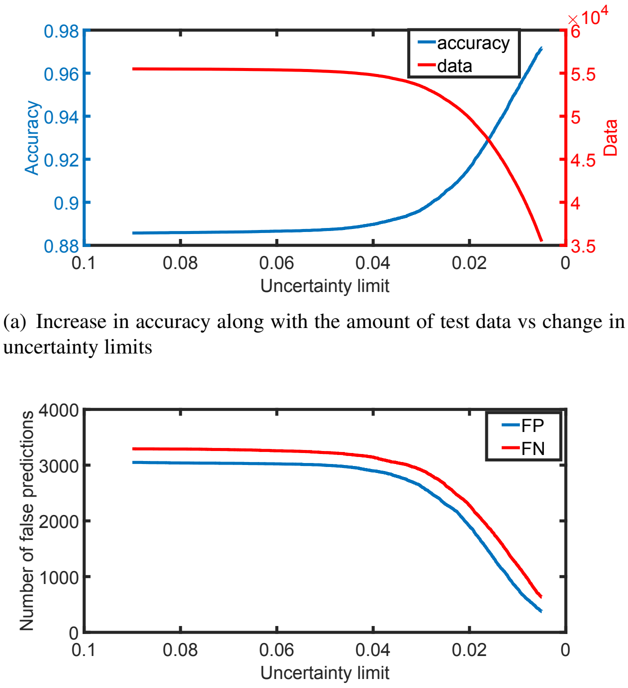


Figure 13: Sample images showing the predictions of the model. Aleatoric and epistemic uncertainties are given by A and E respectively. a) True Positive [$A = 4.17 \times 10^{-1}$, $E = 8.32 \times 10^{-2}$], b) True Positive [$A = 5.11 \times 10^{-2}$, $E = 1.80 \times 10^{-4}$], c) True positive [$A = 3.39 \times 10^{-1}$, $E = 1.71 \times 10^{-2}$], d) True positive [$A = 4.06 \times 10^{-2}$, $E = 2.29 \times 10^{-4}$], e) False negative [$A = 4.56 \times 10^{-1}$, $E = 4.35 \times 10^{-2}$] f) True Positive [$A = 1.37 \times 10^{-1}$, $E = 1.29 \times 10^{-4}$]

The uncertainty values associated with an individual image denotes the confidence with which the network predicts the class label of that image. The aleatoric uncertainty corresponds to the noise in the data for which we don't have any control when we design our model. The epistemic uncertainty corresponds to the variability coming from the model predictions which can be improved. 12 images from the test data set with their epistemic uncertainty estimates are shown in Fig. 12 and 13. The images with negative ground truth labels are shown in Fig. 12 and the ones with positive ground truth labels are shown in Fig. 13. The important features that we can see from the images of the respective classes are texture density and color. It appears that IDC positive class has darker and denser images. The images with high epistemic uncertainty, Fig. 12(a),12(c),12(e), and 13(a),13(c),13(e), has features from both the classes. Therefore, no matter its prediction, the model gives high uncertainty to those images. That is, the network is less confident about these predictions because of their mixed features. Thus, uncertainty estimates can be a measure that can be used to filter out images that the machine is less confident in predicting so that these images can be put to further scrutiny either by a pathologist or a more rigorous machine learning algorithm.

The improvement in results in terms of both accuracy and the reduction in false predictions as we impose stricter constraints on the uncertainty limits is depicted in Fig. 14. Thus, we demonstrate that modified Bayesian–CNN has the potential to improve the performance of classification, especially when the data is imbalanced towards a particular class which is the case for a majority of medical imaging data.



(a) Increase in accuracy along with the amount of test data vs change in uncertainty limits

(b) Reduction in false predictions with the change in uncertainty limits

Figure 14: Improvements in the results with the stricter uncertainty limits

4. Conclusions

In this study, we proposed an extension of Bayesian–CNN model. We utilised it for classification of breast histopathological images and its uncertainty quantification. To conclude this study, we recapitulate upon its four main findings.

Firstly, we have proposed a novel adaptive activation enabled Bayesian–CNN and call it modified Bayesian–CNN. In the proposed model we introduce a learnable activation function that adapts to the training data in order to improve predictive capabilities. In addition, the proposed model has faster convergence and better accuracy over Bayesian–CNN.

Secondly, we have shown that the proposed modified Bayesian-CNN can reduce the number of false negative predictions remarkably (about 38%) over CNN, as compared to 28% reduction through Bayesian–CNN.

Thirdly, we have demonstrated that using the proposed modified Bayesian–CNN and the Bayesian–CNN, the problem of overfitting can be nearly eliminated without any recourse to regularization. That is, the Bayesian approach work with almost equal accuracy for unknown and known data sets. The Bayesian approach automatically regularizes, since through stochastic parameters it represents an ensemble average over a set of CNNs instead of a single deterministic CNN.

Fourthly, through analyzing the uncertainty-quantification, we found that the images which have higher uncertainties have features of both the classes. We have also demonstrated that the accuracy of predictions can be improved by imposing stricter constraints on the uncertainty limits. Thus, uncertainty measures can guide the need for human expert intervention.

To summarize, the present work demonstrates that a classifier based on the modified Bayesian–CNN can be used as an accurate and automated classifier to detect breast cancer. Further, it can help in diagnosis by quantifying the uncertainty in the prediction. These findings require further investigation on larger data sets, however, they show potential routes to improve upon the state–of–the–art automated classifier for a broad range of biomedical imaging applications.

Acknowledgements

We acknowledge the high performance computing facilities from the SUPERIOR at MTU and XSEDE (supported by the NSF grant# ACI-1548562). (Request#: MSS200004).

References

[1] C. P. Wild, B. W. Stewart, and C. Wild, *World cancer report 2014*. World Health Organization Geneva, Switzerland, 2014.

[2] C. Fitzmaurice *et al.*, "Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study," *JAMA oncology*, vol. 4, no. 11, pp. 1553–1568, 2018.

[3] F. Bray, P. McCarron, and D. M. Parkin, "The changing global patterns of female breast cancer incidence and mortality," *Breast cancer research*, vol. 6, no. 6, p. 229, 2004.

[4] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast cancer multi-classification from histopathological images with structured deep learning model," *Scientific reports*, vol. 7, no. 1, pp. 1–10, 2017.

[5] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–171, 2009.

[6] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, "Digital imaging in pathology: whole-slide imaging and beyond," *Annual Review of Pathology: Mechanisms of Disease*, vol. 8, pp. 331–359, 2013.

[7] S. Al-Janabi, A. Huisman, and P. J. Van Diest, "Digital pathology: current status and future perspectives," *Histopathology*, vol. 61, no. 1, pp. 1–9, 2012.

[8] T. J. Fuchs and J. M. Buhmann, "Computational pathology: Challenges and promises for tissue analysis," *Computerized Medical Imaging and Graphics*, vol. 35, no. 7-8, pp. 515–530, 2011.

[9] M. L. Giger, H.-P. Chan, and J. Boone, "Anniversary paper: history and status of cad and quantitative image analysis: the role of medical physics and aapm," *Medical physics*, vol. 35, no. 12, pp. 5799–5820, 2008.

[10] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Computerized medical imaging and graphics*, vol. 31, no. 4-5, pp. 198–211, 2007.

[11] M. Aswathy and M. Jagannath, "Detection of breast cancer on digital histopathology images: Present status and future possibilities," *Informatics in Medicine Unlocked*, vol. 8, pp. 74–79, 2017.

[12] P. Filipczuk, T. Fevens, A. Krzyżak, and R. Monczak, "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies," *IEEE transactions on medical imaging*, vol. 32, no. 12, pp. 2169–2178, 2013.

[13] Y. M. George, H. H. Zayed, M. I. Roushdy, and B. M. Elbagoury, "Remote computer-aided breast cancer detection and diagnosis system based on cytological images," *IEEE Systems Journal*, vol. 8, no. 3, pp. 949–964, 2013.

[14] M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, and R. Monczak, "Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images," *Computers in biology and medicine*, vol. 43, no. 10, pp. 1563–1572, 2013.

[15] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.

[16] Y. Zhang, B. Zhang, F. Coenen, and W. Lu, “Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles,” *Machine vision and applications*, vol. 24, no. 7, pp. 1405–1420, 2013.

[17] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “A dataset for breast cancer histopathological image classification,” *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[18] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, “Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images,” *IEEE transactions on medical imaging*, vol. 37, no. 1, pp. 316–325, 2017.

[19] N. Bayramoglu, J. Kannala, and J. Heikkilä, “Deep learning for magnification independent breast cancer histopathology image classification,” in *2016 23rd International conference on pattern recognition (ICPR)*, pp. 2440–2445, IEEE, 2016.

[20] A. Cruz-Roa *et al.*, “Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks,” in *Medical Imaging 2014: Digital Pathology*, vol. 9041, p. 904103, International Society for Optics and Photonics, 2014.

[21] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “Breast cancer histopathological image classification using convolutional neural networks,” in *2016 international joint conference on neural networks (IJCNN)*, pp. 2560–2567, IEEE, 2016.

[22] T. Araújo *et al.*, “Classification of breast cancer histology images using convolutional neural networks,” *PloS one*, vol. 12, no. 6, 2017.

[23] H. Lin, H. Chen, S. Graham, Q. Dou, N. Rajpoot, and P.-A. Heng, “Fast scannet: fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection,” *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1948–1958, 2019.

[24] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, “Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images,” *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[26] Z. Jia, X. Huang, I. Eric, C. Chang, and Y. Xu, “Constrained deep weak supervision for histopathology image segmentation,” *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2376–2388, 2017.

[27] Y. Li, J. Wu, and Q. Wu, “Classification of breast cancer histology images using multi-size and discriminative patches based on deep learning,” *IEEE Access*, vol. 7, pp. 21400–21408, 2019.

[28] H. Yang, J.-Y. Kim, H. Kim, and S. P. Adhikari, “Guided soft attention network for classification of breast cancer histopathology images,” *IEEE transactions on medical imaging*, 2019.

[29] B. Xu *et al.*, “Attention by selection: A deep selective attention approach to breast cancer classification,” *IEEE Transactions on Medical Imaging*, 2019.

[30] M. Nawaz, A. A. Sewissy, and T. H. A. Soliman, “Multi-class breast cancer classification using deep learning convolutional neural network,” *Int. J. Adv. Comput. Sci. Appl*, vol. 9, no. 6, pp. 316–332, 2018.

[31] N. H. Motlagh, M. Jannesary, H. Aboulkheyr, P. Khosravi, O. Elemento, M. Totonchi, and I. Hajirasouliha, “Breast cancer histopathological image classification: A deep learning approach,” *bioRxiv*, p. 242818, 2018.

[32] J. Xie, R. Liu, J. Luttrell IV, and C. Zhang, “Deep learning based analysis of histopathological images of breast cancer,” *Frontiers in genetics*, vol. 10, p. 80, 2019.

[33] G. Aresta *et al.*, “Bach: Grand challenge on breast cancer histology images,” *Medical image analysis*, vol. 56, pp. 122–139, 2019.

[34] S. S. Chennamsetty, M. Safwan, and V. Alex, “Classification of breast cancer histology image using ensemble of pre-trained neural networks,” in *International conference image analysis and recognition*, pp. 804–811, Springer, 2018.

[35] S. Kwok, “Multiclass classification of breast cancer in whole-slide images,” in *International conference image analysis and recognition*, pp. 931–940, Springer, 2018.

[36] N. Brancati, M. Frucci, and D. Riccio, “Multi-classification of breast cancer histology images by using a fine-tuning strategy,” in *International conference image analysis and recognition*, pp. 771–778, Springer, 2018.

[37] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[39] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[40] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[42] E. Begoli, T. Bhattacharya, and D. Kusnezov, “The need for uncertainty quantification in machine-assisted medical decision making,” *Nature Machine Intelligence*, vol. 1, no. 1, pp. 20–23, 2019.

[43] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight uncertainty in neural networks,” *arXiv preprint arXiv:1505.05424*, 2015.

[44] K. Shridhar, F. Laumann, and M. Liwicki, “A comprehensive guide to bayesian convolutional neural network with variational inference,” *arXiv preprint arXiv:1901.02731*, 2019.

[45] K. Shridhar, F. Laumann, and M. Liwicki, “Uncertainty estimations by softplus normalization in bayesian convolutional neural networks with variational inference,” *arXiv preprint arXiv:1806.05978*, 2018.

[46] A. Kendall and Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?,” in *Advances in neural information processing systems*, pp. 5574–5584, 2017.

[47] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, “Uncertainty quantification using bayesian neural networks in classification: Application to ischemic stroke lesion segmentation,” 2018.

[48] Paul Mooney, “Breast Histopathology Images.” `https://www.kaggle.com/paultimothymooney/breast-histopathology-images`. 2017.

[49] A. Graves, “Practical variational inference for neural networks,” in *Advances in neural information processing systems*, pp. 2348–2356, 2011.

[50] A. D. Jagtap, K. Kawaguchi, and G. E. Karniadakis, “Adaptive activation functions accelerate convergence in deep and physics-informed neural networks,” *Journal of Computational Physics*, vol. 404, p. 109136, 2020.

[51] M. Dushkoff and R. Ptucha, “Adaptive activation functions for deep networks,” *Electronic Imaging*, vol. 2016, no. 19, pp. 1–5, 2016.

[52] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[53] G. B. Orr and K.-R. Müller, *Neural networks: tricks of the trade*. Springer, 2003.

[54] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.

[55] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *Neural Networks*, vol. 106, pp. 249–259, 2018.

[56] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.