Bayesian Deep Learning Applied to Diabetic Retinopathy with Uncertainty Quantification

Masoud Muhammed Hassan[1*], Halbast Rashid Ismail[2]

[1] Department of Computer Science, Faculty of Science, University of Zakho, Duhok, Kurdistan Regain, Iraq.

[2] Technical College of Informatics-Akre, Duhok Polytechnic University, Duhok, Kurdistan Regain, Iraq.

[*] Corresponding author: email: masoud.hassan@uoz.edu.krd, Phone: 009647504578082.

Abstract

Deep Learning (DL) has significantly contributed to the field of medical imaging in recent years, leading to advancements in disease diagnosis and treatment. In the case of Diabetic Retinopathy (DR), DL models have shown high efficacy in tasks such as classification, segmentation, detection, and prediction. However, DL model's opacity and complexity lead to errors decision-making, particularly in complex cases, making it necessary to estimate the models' uncertainty in predictions. Therefore, there is a need to estimate uncertainty in the model's predictions, which cannot be estimated by classical DL models alone. To address this issue, Bayesian DL methods have been proposed, and their use is increasing in the field. In this paper, we develop a Convolutional Neural Network (CNN) model for DR classification using the simplest architecture. We then apply the Bayesian CNN by Variational Inference (VI) and Monte Carlo dropout (MC-Dropout) methods to the same CNN architecture to obtain the posterior predictive distribution. The Aptos 2019 dataset was used to evaluate the performance of the models. Our experimental results demonstrate that the proposed models outperform other state-of-the-art models in terms of test accuracy, achieving 94.4% for CNN, 94% for BCNN-VI, and 93.3% for MC-Dropout. Finally, we compute the entropy and standard deviation on the obtained predictive distribution to quantify the model uncertainty. This research highlights the potential benefits of using Bayesian DL methods in medical image analysis to enhance the accuracy and reliability of disease diagnosis and treatment.

Keywords:

Bayesian Deep Learning, Convolutional Neural Network, Uncertainty Quantification, Diabetic Retinopathy.

1. Introduction

The emergence of Deep Learning (DL) has made a significant contributions to various fields, including medical diagnoses, financial risk assessment, and autonomous driving [1]. However, despite their success, DL algorithms struggle to communicate the uncertainty in their predictions, rendering them unable to indicate when they are uncertain. This limitation of DL models make them unreliable for critical decision-making, particularly in the medical domain, where the consequences of erroneous decisions can be severe. To address this issue, there is a growing need to develop reliable techniques for quantifying uncertainty in deep learning models. Bayesian modeling has emerged as a promising approach for reducing the risks associated with uncertainty. Bayesian modeling provides a formal framework for developing learning algorithms and training uncertainty-aware neural networks. In the medical field, where the accurate detection of diseases such as cancer can mean the difference between life and death, the ability to quantify uncertainty is crucial for building trust in automated diagnostic systems [2]. Thus, several techniques have been proposed to develop reliable uncertainty-aware deep learning models, with Bayesian modeling being one of the most promising approaches [3].

Bayesian Deep Learning (BDL) is an extension of Deep Learning (DL) that combines the Bayesian probability theory with DL techniques. By imposing prior distributions on the model parameters (weights), Bayesian Neural Networks (BNNs) provide a distribution of posterior for these parameters, which allows for the calculation of prediction and uncertainty estimates [4]. In contrast to conventional DL, which only produces a deterministic output, BDL quantifies uncertainty through probability density over outcomes. This approach is particularly relevant in medical applications, where the accurate quantification of uncertainty is crucial for automated screening and referral of uncertain cases to medical professionals. BDL methods have been increasingly used in medical imaging [5],

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

including the classification and segmentation of diabetic retinopathy , breast cancer detection [6], and brain tumor segmentation [7]. The ability to quantify uncertainty provides valuable information to medical professionals, enabling them to make informed decisions based on the likelihood of a particular diagnosis.

BDL is a probabilistic approach to deep learning that leverages Bayes' theorem. The input components of BDL are the prior distribution and the likelihood of the data, while the output is the posterior distribution [5]. The posterior distribution can be used to quantify various types of uncertainties associated with the used architecture and data [8]. Different type of statistical distributions can be used to represent the prior and posterior distributions in BDL; however, the most popular distributions for BDL with image data are Normal and Bernoulli distribution [9]. To generate output, the Bayesian approach samples from the posterior distribution. One of the most popular exact sampling methods for posterior distribution is Markov Chain Monte Carlo (MCMC). However, MCMC can be computationally expensive thus makes and not feasible to scale up [10] [11]. Fortunately, there are other approximation methods that can be used for larger scale datasets such as images. These methods are faster with fewer parameters compared to MCMC [5]. The two most popular approximation methods are Variational inference (VI) [12] [13] and Monte-Carlo Dropout (MC-Dropout) [14] [15]. VI assumes that the data are normally distributed, resulting in a normal distribution for the model's posterior. On the other hand, MC-Dropout uses Bernoulli distribution to generate a posterior distribution.

Diabetic Retinopathy (DR) is a diabetes-related complication that affects the blood vessels in the retina, leading to vision impairment or blindness if left untreated [16]. It is caused by high blood sugar levels damaging the blood vessels in the retina, leading to leakage or blockage of the vessels. The condition can progress through different stages, from mild non-proliferative DR to severe proliferative DR, which involves the growth of abnormal blood vessels in the retina. Early detection and treatment of DR are crucial to prevent vision loss and blindness, and regular eye exams are recommended for individuals with diabetes. Recently, the automatic classifying of DR by using DL has been of increasing interest [17] [18] [19] [20]. Moreover, the development of reliable DL techniques for classification tasks has received a lot of attention recently, and the most popular method is approximate BDL, which approximates the posterior distribution of BNN in a computationally scalable way. In addition to diagnosis and prognosis, Bayesian deep learning can also be used to develop personalized treatment plans for patients with DR [21]. By incorporating patient-specific information and uncertainty estimates into the model, doctors can make more accurate predictions of how a patient will respond to different treatments. Overall, Bayesian DL has the potential to improve the accuracy and robustness of models for diagnosing, predicting, and treating DR.

The objective of this research is to leverage Bayesian Deep Learning (BDL) techniques to classify Diabetic Retinopathy (DR) images while primarily focusing on quantifying model uncertainty. Additionally, the study aims to compare two common approximation methods for Bayesian Convolutional Neural Networks (BCNNs) - Variational Inference (VI) and Monte Carlo Dropout (MC-Dropout) - for their uncertainty quantification capabilities. The key contributions of this study are as follows:

- Develop a simple CNN model architecture for classifying DR images.
- Apply Bayesian Convolutional Neural Networks (BCNN) to the same architecture of the developed CNN by using both VI and MC-Dropout approximation methods.
- Improving the classification performance of both CNN and BCNN models by identifying optimal hyperparameters and training procedures.
- Development of BCNN architecture that can not only classify DR images but also estimate the uncertainty of model predictions by modeling the posterior predictive distribution.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of the literature on deep learning and Bayesian deep learning in the context of diabetic retinopathy. In Section 3, we describe the proposed methods and the dataset used in this study. Section 4 provides the experimental results and associated

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

discussion, while Section 5 concludes the study by summarizing the key findings and discussing their implications for future research in the field.

2. Related work

In recent years, several studies have explored the use of deep learning, Bayesian deep learning, and uncertainty-aware models for medical diagnosis on public datasets. In this section, we provide a review of some of the most recent studies in this area.

2.1 Deep learning applied on Diabetic Retinopathy

Alahmadi, [22] suggested a deep learning model that combines a content recalibration method and a mechanism that adaptively measures informative areas for image classification of diabetic retinopathy. The authors used a system for separating style from content representation, as well as a spatial normalization module and an attention module. Their recent experiment on the APTOS 2019 dataset showed that their model outperformed previous research.

In another study by Alyoubi et al. [23], two deep learning models were proposed for DR diagnosis and lesion localization. The first model, CNN512, was used to classify images into one of the five DR levels using the entire image as an input to the CNN. The second model utilized an adapted YOLOv3 architecture to identify and localized DR lesions, attaining a 0.216 mAP in lesion localization on the DDR dataset. The authors combined both models on the APTOS-2019 and DDR public datasets, the suggested CNN512 and YOLOv3 structures were combined to accurately classify DR images and pinpoint DR lesions, yielding an accuracy of 89% with 89% sensitivity and 97.3 specificity. They conclude that their proposed models offer high accuracy and can assist medical professionals in the diagnosis and localization of DR.

Following the previous research, Yi et al. [24] proposed a new diagnostic model for DR-grade named RA-EfficientNet, depending on the combination of residual attention blocks and EfficientNet for feature extraction applied on the APTOS 2019 dataset. They used their proposed model for 2- and 5-grade diagnosis and classification of DR. According to their results, the proposed model effectively improves the efficiency of DR detection, achieving an accuracy of 93.55% in 5-grade classifiers and an accuracy of 98.36% in 2-grade classifiers. The study suggests that the use of advanced DL architectures and attention mechanisms can enhance the performance of DR detection models.

Gangwar & Ravi [25] proposed a novel hybrid DL model to solve the issue of automatic DR detection. The researcher employed transfer learning by utilizing a pretrained Inception-ResNet-v2 as the base, and included a customized block of CNN layers on top of it to build a hybrid model. They obtained a test accuracy of 82.18% and 72.33% on APTOS and Messidor-1 dataset, respectively.

Majumder & Kehtarnavaz [26] presented a multitask DL model to diagnose the five phases of DR. The multitask model that built contained a classification model and a regression model. A modified, densely connected deep neural network has also been developed for excitation as part of this multitasking method. The two substantial Kaggle datasets for APTOS and EyePACS were applied to the multitask model. According to their results, the multitask model created for the EyePACS and APTOS datasets has performed a weighted Kappa value of 0.88 and 0.90 respectively.

In another study by Al-Antary & Arafa [27] a novel DL model (Multi-Scale Attention network (MSA-Net)) was developed for classification of the damage resulted in DR. The MSA-Net technique was used on top of the high-level representation, to enhance the discriminatory power of features. In a standard way, the model was trained by the loss cross-entropy to classify the DR levels. The proposed approach has achieved good accuracy on two public datasets with 84% for APTOS and 79 % for EyePACS.

In another study by Padmanayana & B.K [28], a DL model for binary classification of DR detection was developed using an online interface for easy of interaction. The provided input images were improved, preprocessed, and classified as 0 or 1. The website showed the anticipated outcome as DR or with No-DR. The model's performance was compared with various optimizers, including Adam, RMSPROP with momentum, and Adagrad. Kaggle's APTOS-2019 dataset was utilized.

Islam et al. [29] presented the supervised contrastive learning (SCL) method and the transfer learning Xception CNN pre-trained model for DR classification. The developed approach achieved high accuracy of 98.36% for binary classification and 84.364% for multi-classification for APTOS 2019 dataset. The proposed model's performance was

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

examined using the Messidor-2 dataset. It was found that the suggested approach outperformed the normal CNN without SCL in terms of performance for DR detection [5].

Menaouer et al. [30] suggested a hybrid DL method for DR detection and classification according to the visual risk associated with the severity of retinal disease, utilizing the CNN method and two models of VGG: VGG19 and VGG16. The primary goal of this study was to create a reliable system for automatically detecting and classifying DR. Their hybrid method performed well on multiclass tasks with a 90.60% accuracy rate.

2.2 Bayesian Deep Learning Applied on Diabetic Retinopathy

BDL has emerged as a promising approach to quantify model uncertainty, which is particularly relevant in medical applications where decisions based on the model outputs can have serious consequences. Recent research by Jaskari et al. [31] has demonstrated the benefits of applying BDL to clinical datasets, such as diabetic retinopathy. They investigated the performance of various Bayesian approximation methods on a clinical dataset from a hospital in Finland, as well as publicly available datasets including APTOS, EyePACS, KSHHP, and Messidor-2. The authors also developed a novel uncertainty metric based on risk-based classier, which improved density uncertainty performance on both the two datasets used. The advantage of this study lies in its comprehensive evaluation of various BDL methods on multiple datasets, which allows for a more robust comparison of performance. The findings demonstrate the potential of BDL in improving the accuracy and reliability of deep learning models in medical applications, ultimately leading to better patient outcomes.

In a related study, Band et al. [32] proposed benchmarking tasks for the detection of DR using Bayesian deep learning methods. They evaluated both well-established and state-of-the-art Bayesian and non-Bayesian approaches on a set of task-specific performance and reliability measures. The authors specifically focused on two large retinal datasets, EyePACS and APTOS, and applied several Bayesian inference methods, Mean-Field Variational Inference (MFVI), Structured Mean-Field Variational Inference (SFVI), (SFVI), Monte Carlo Dropout (MC-Dropout), Maximum a Posteriori (MAP) and deep ensembles. The strength of this study lies in its comprehensive evaluation of various BDL methods on multiple datasets, along with providing implementations of different benchmark approaches, and findings computed over 20-GPU days, 100-TPU days, 400 hyper-parameters configurations with evaluations conducted on at least six random seeds for each method [21].

Ahsan et al. [33] developed a hybrid model for DR classification that addresses the challenge of uncertainty and also leverages unlabelled data. The proposed framework consists of two main parts: a Bayesian CNN model with MC-Dropout, which was utilized as a feature descriptor, and an Active Learning (AL) part. The Bayesian CNN was designed to reduce the uncertainty of predictions, while the AL component enables the model to learn from unlabeled data. The authors evaluated their proposed model against state-of-the-art approaches using various metrics and deduced that their hybrid method outperformed these methods. The strength of this study lies in its novel approach to addressing uncertainty and leveraging unlabeled data, which can be challenging in medical applications.

Kwon et al. [34] proposed a novel method for quantifying uncertainty in DR classification based on a Bayesian NN model. Their proposed approach was based on utilizing two medical datasets, ISLES and DRIVE, and applied VI as a Bayesian approximation technique for quantifying model uncertainty. Compared to existing methods, the proposed method has some advantages, including numerical stability and the ability to express inherent variability in terms of the underlying distribution of the outcome. These features make the proposed method more reliable and informative for uncertainty quantification in medical classification tasks.

In another paper by Toledo-Cortés et al. [35], a hybrid DL Gaussian Process (DLGP) model was introduced for DR binary classification and quantifying model uncertainty. The developed approach utilized Radial Basis Function as an approximate for model uncertainty quantification. The advantage of this research comes from utilizing the Gaussian Process with the DL in two various datasets, Messidor-2 and EyePACS, and comparing its performance with other state-of-the-art approaches. They concluded that their proposed approach outperformed others, demonstrating the effectiveness of the DLGP-DR method for uncertainty quantification in medical classification tasks.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

Filos et al. [36] also proposed a new benchmark for Bayesian DL model for diagnosing and classification of DR. They used various MC-Dropout and MFVI as Bayesian approximation methods on two datasets, APTOS and EyePACS. Their experimental results showed that the MC-Dropout provided better results compared to other methods investigated in terms of the classification accuracy and uncertainty quantification.

Farquhar et al. [37] suggested using the Radial Bayesian Neural Network (RBNN) as a new approach to Bayesian neural network modeling that avoids the "soap-bubble" problem encountered by MFVI, which arises when using multivariate Gaussians as posterior distributions. The proposed method successfully identified a simple approximate posterior distribution in a hyperspherical space. The study's main contribution is the demonstration that the RBNN outperforms other state-of-the-art Bayesian DL models, including MFVI, deep ensembles, and MC-Dropout, on various tasks, such as image classification and regression. The authors also show that the proposed method can provide more accurate uncertainty estimates than other Bayesian methods.

Toledo-Cortés et al. [38] proposed a novel Probabilistic DL Ordinal Regression (PDLOR) model for diagnosing medical images. They evaluated their proposed method on two different medical imaging tasks: diagnosing prostate cancer and the detecting diabetic retinopathy levels. They utilized the RBF to approximate the posterior distribution and quantify uncertainty, and compared its performance to classical DL architectures. The authors reported that their proposed method outperformed these architectures in both tasks and provided better interpretability of outcomes. The strength of this paper lies in its contribution to improving the accuracy and interpretability of medical image analysis outcomes.

3. Proposed Method

In this paper, we have developed a new multi-layer architecture of Convolutional Neural Networks (CNN) and Bayesian CNN (BCNN) to classify DR based on retinal images. The proposed method in this paper consists of three main phases: pre-processing retinal images, training a CNN and Bayesian CNN model to detect and classify DR, and quantifying the model's uncertainty. The developed multi-layer architecture of CNN and Bayesian CNN has the potential to improve the accuracy and reliability of DR diagnosis based on retinal images. The following steps provide a detailed description of the proposed method.

1- Pre-processing:
The pre-processing step is performed on retinal images to perform contrast enhancement. This step aims to improve the quality of the retinal images and enhance the features of the retina that are essential for DR diagnosis.

2- CNN and Bayesian CNN models:
After the pre-processing step, the suggested CNN and Bayesian CNN models are used to detect and classify the various statuses of DR. The CNN model is trained on a large dataset of retinal images to learn the patterns and features that are indicative of DR. The Bayesian CNN model, on the other hand, takes into consideration the uncertainty in the model's predictions by estimating the model's posterior distribution over its parameters.

3- Model uncertainty quantification:
The Bayesian CNN method is then used to quantify the model's uncertainty. This step is crucial in medical diagnosis, as it provides a measure of how confident the model is in its predictions. The model's uncertainty is quantified by estimating the predictive distribution of the model, which takes into account the uncertainty in the model's parameters. Figure 1 shows a diagram of the proposed method.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

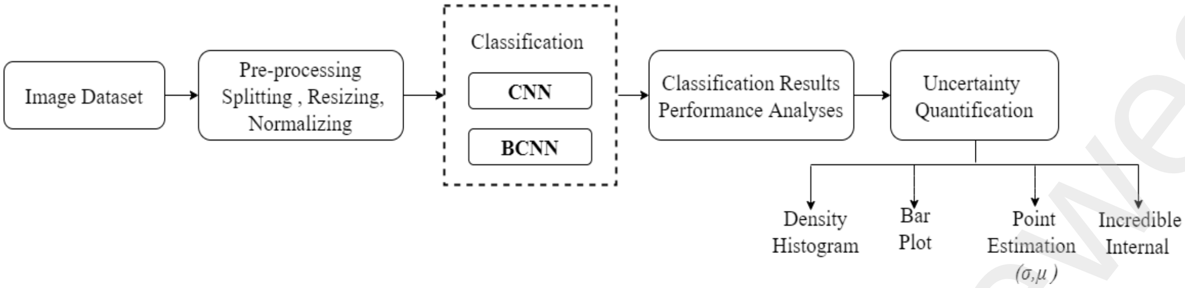


Figure 1: Block diagram of the proposed method.

3.1 Dataset Used

In this study, the APTOS 2019 dataset provided by the Asia Pacific Tele-Ophthalmology Society (APTOS) was used [39]. This dataset was used as part of the Detection Competition for Blindness [40] in 2019, and is widely used for diabetic retinopathy, which is publicly available in Kaggle [41]. The dataset consists of 3662 high-resolution color retinal images of different sizes taken using fundus imaging under different imaging conditions. The smallest and largest original image sizes in this dataset are 640 × 480 and 2848 × 4288, respectively. The retinal images in the APTOS 2019 dataset are categorized into 5 classes representing the severity of diabetic retinopathy: 0 (No-DR), 1 (Mild-DR), 2 (Moderate-DR), 3 (Severe-DR), and 4 (Proliferative-DR). The dataset includes 1805 images labeled as No-DR, 370 images labeled as Mild-DR, 999 images labeled as Moderate-DR, 193 images labeled as Severe-DR, and 293 images labeled as Proliferative-DR. Sample images from the APTOS dataset are shown in Figure 2.

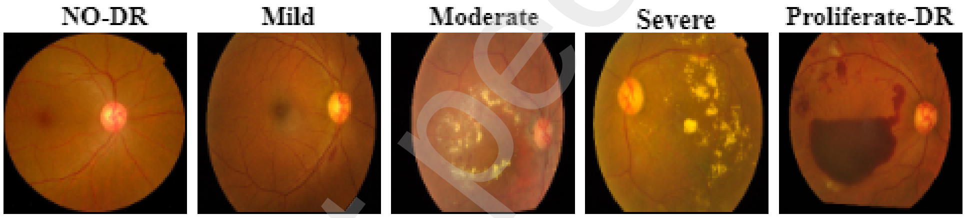


Figure 2: Different labels of DR in the APTOS-2019 dataset.

3.2 CNN Model

The proposed method utilizes Convolutional neural networks (CNN) to detect and classify diabetic retinopathy based on retinal images. CNNs are a type of ANN that can automatically learning low and high-level features from medical images. These features can be used to identify, categorize, and stage diseases [42] [43]. The developed CNN architecture consists of a convolutional layer, a pooling layer, and a fully connected layers (as shown in Figure 3). The convolutional layers apply a set of linear filters to extract various low- and high-level features, such as edges, curves, blood vessels, from the input image or the activation map in the previous convolutional layer. Figure 3 illustrates the proposed CNN architecture, which is trained on the APTOS 2019 dataset to learn the patterns and features indicative of diabetic retinopathy. The model uses cross-entropy loss as the objective function to optimize the model's parameters during training. The trained CNN model is then used to classify the different statuses of DR based on retinal images.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

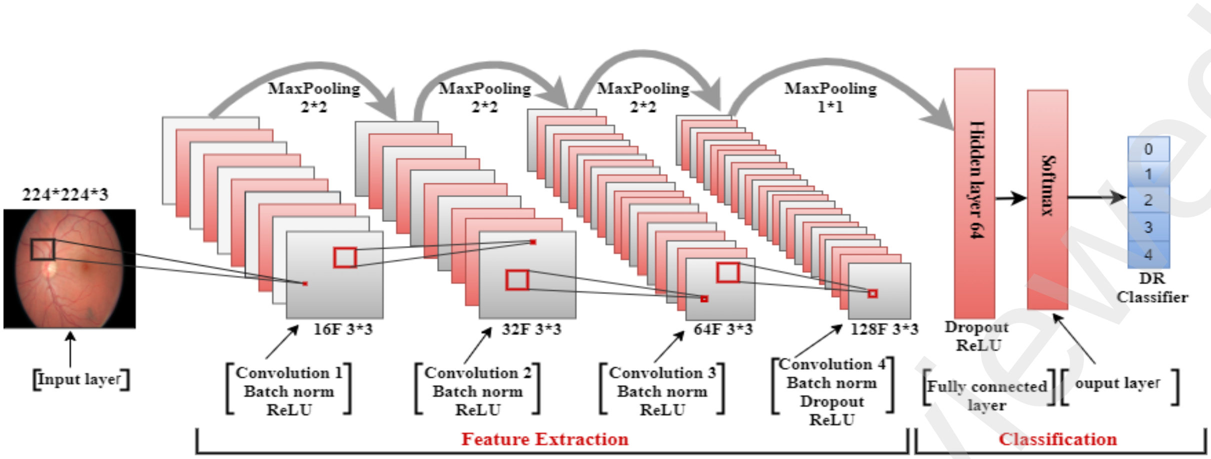


Figure 3: Architecture of the proposed Convolutional Neural Network model.

The proposed CNN model in this research consists of five convolutional layers, five Maxpool layers, five batch-normalization layers, two dropout layers, and two fully connected layers. The architecture is designed to be simple yet effective, in contrast to the complex models or pre-trained algorithms used in recent studies. For each convolution layer, the proposed CNN uses a ReLU activation function. The last dense layer uses Softmax activation to generate class probabilities. The Maxpool layers are used to reduce the dimensions of the output image, and batch-normalization layers normalize the output of the previous layer to speed up training and improve performance. To reduce overfitting during training, two dropout layers are used to randomly drop some neurons. The fully connected layers are used to combine the features and make a classification decision. Table 1 represents each layer and the parameters used in the proposed CNN model. The architecture of the proposed CNN model is shown in Figure 3. The goal of this architecture is to achieve high accuracy with a simple model, which can be useful in real-world applications where computational resources may be limited.

Table 1: Details of the proposed CNN architecture.

Layer	Parameters
Conv-1	8 filter 3*3, padding = “valid”
Max-pooling	2*2
Batch-normalization	
Conv-2	16 filter 3*3, padding = “valid”
Max-pooling	2*2
Batch-normalization	
Conv-3	32 filter 3*3, padding = “valid”
Max-pooling	2*2
Batch-normalization	
Conv-4	64 filter 3*3, padding= “valid”
Max-pooling	2*2
Batch-normalization	
Conv-5	128 filter 3*3, padding = “valid”
Max-pooling	1*1
Batch-normalization	
Dropout	0.1
FCL	64
Dropout	0.2

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

3.3 Bayesian CNN model

In this research, the proposed CNN architecture is converted into a Bayesian CNN to capture the uncertainty in the model's predictions. In a Bayesian CNN, the weights and biases of the model are treated as random variables rather than deterministic variables utilized in a traditional CNN. This allows the Bayesian CNN to capture the variability in the dataset and estimate uncertainty in its predictions [44] [45]. To make the proposed CNN architecture probabilistic, a prior distribution is defined for the weights and biases, and a posterior distribution is obtained by applying Bayes' rule. The posterior distribution is then used to generate multiple sets of weights, which are used to train an ensemble of networks. The outputs of the ensemble networks are averaged to obtain a probability distribution for each weight.

Given a training dataset and a supervised learning environment, $\mathcal{D} = \{\boldsymbol{x_n}, y_n\}_{n=1}^{N}$, where N represents the size of the dataset, $\boldsymbol{x_n}$ represents a vector of input features, where $\boldsymbol{x}_n \in \mathcal{R}^m = [x_{1,n}, x_{2,n}, \ldots, x_{m,n}]$, and y_n denotes the corresponding label, where $y_n \in \{1, 2, \ldots \mathrm{C}\}$ C is the number of categories. The ultimate goal of the NN model is to accurately estimate $\dot{y}_n = f(\boldsymbol{x}_n)$. This is achieved by minimizing the prediction error on the training dataset and by generalizing the learned relationship to unseen data.

In an ANN model with L layers, such as CNN, the set of weights used to determine the model is denoted by $(\boldsymbol{w} = \{w_i\}_{i=1}^{L})$. The goal of Bayesian methods is to estimate the posterior uncertainty on the network parameters $P(w|\mathcal{D})$ given a $(\mathcal{D})$ dataset, by assuming a distribution of prior on neural network parameters $P(w)$. This prior distribution provides an assumption for which neural network parameters' functions are most likely to produce the data. During inference, the prediction probability of the model $(\hat{y})$ over a test data input $(\boldsymbol{x}^*)$ can be computed by integrating all feasible values in w:

$$p(\dot{y}|\boldsymbol{x}^*,\mathcal{D}) = \mathbb{E}_{p(\mathbf{w}|\mathcal{D})}[p(\dot{y}|\boldsymbol{x}^*,\mathbf{w})] = \int_{\mathbf{w}} p(\dot{y}|\boldsymbol{x}^*,\mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w} \tag{1}$$

However, in practice, computing this integral exactly in Eq. 1 is computationally intractable. This is because calculating the $P(w|\mathcal{D})$ probability distribution is a challenging task. Therefore, several approximation methods have been proposed to achieve an analytically solvable inference. These methods include expectation propagation [46], Markov Chain Monte Carlo probabilistic sampling-based inference (MCMC) [11] [47] [48], Monte Carlo Dropout (MC-Dropout) approximation inference [49] [50] [51], and Variational Inference (VI) [13] [52]. These approximation methods help to compute the probability distribution and estimate the posterior uncertainty of the network parameters by approximating the intractable integral in Eq. 1. In our proposed, we used both VI and MC-Dropout methods, as follows.

3.3.1 Variational Inference

Variational Inference (VI) is a technique used to approximate the posterior distribution of a Bayesian model. To approximate the posterior distribution $p(\mathbf{w}|\mathcal{D})$ via a fit of an approximation $q_\theta(\mathbf{w}) \approx p(\mathbf{w}|\mathcal{D})$, we evaluate the VI methods [45] [53] [13]. The goal is to find an approximate distribution $q_\theta(\mathbf{w})$ that is as close as possible to the true posterior distribution p(w|D) given the observed data D. To measure the proximity between $q_\theta(\mathbf{w})$ and $p(\mathbf{w}|\mathcal{D})$, the Kullback-Leibler (KL) divergence is commonly used in information theory [54]. The KL-divergence measures the amount of information lost when approximating one distribution with another. In this case, we want to minimize the KL-divergence between $q_\theta(\mathbf{w})$ and $p(\mathbf{w}|\mathcal{D})$, which is defined as:

$$\mathrm{KL}\left(q_\theta(\mathbf{w}) \parallel p(\mathbf{w}|\mathcal{D})\right) \tag{2}$$

Minimizing KL divergence in Eq. 2 is equivalent to reducing the negative evidence lower bound function "ELBO" [49] [53] [55] for θ:

$$\mathcal{L}(\theta) = -\ \mathbb{E}_{q_\theta(\mathbf{w})}[\log p(\mathcal{D}|\mathbf{w})]\ + \mathrm{KL}\left(q_\theta(\mathbf{w}) \parallel p(\mathbf{w})\right) \tag{3}$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

Where $\mathbb{E}_{q_\theta(\mathbf{w})}$ is the expectation with respect to the distribution denotes "a description of how the variational distribution of neural parameters explains the data observed" in the first term, [56] and KL- divergence is in the second term that calculates proximity between true and approximate posterior densities. The cost function defined by Eq.3 is minimized by the small batch random gradient descent method while training the neural networks to get the ideal value of θ, which establishes the distribution over weights' parameters [57].

In short, variational inference is a method for approximating the posterior distribution of a Bayesian model by finding an approximate distribution that minimizes the KL-divergence to the true posterior distribution. The optimization is done by maximizing the evidence lower bound (ELBO) with respect to the parameters of the approximate distribution.

3.3.2 MC Dropout

MC Dropout is another technique used in DL to estimate the uncertainty of the model predictions. It is an extension of the standard dropout [58] regularization technique commonly used in neural networks. In a standard dropout neural network, during training, some nodes are randomly dropped out of the network with a probability p. This forces the network to learn more robust features and helps prevent overfitting. However, during inference, all nodes are kept, and the network makes deterministic predictions.

Bayesian MC Dropout extends this approach by treating the dropout rate p as a probability distribution, which allows us to perform Bayesian inference over the model weights. During inference, instead of making deterministic predictions, the model is sampled multiple times with different dropout masks, and the predictions are averaged. This gives us an estimate of the predictive distribution, which can be used to estimate the model uncertainty. By doing this, it ensures that every neuron participates in the prediction without overfitting.

More recently, the use of dropout regularization in DL models has been demonstrated to be equal to variational approximation Bayesian inference [59]. The main idea for obtaining an uncertainty model in dropout regularization is used at training time, followed by dropout (Bernoulli) sampling at testing time. It was displayed [49],[60] that the group of mean-weight matrix (L layered NN) and probabilities of dropout (parameters of variational) for a distribution dropout satisfies $\theta = \{\mathbf{M}_l, p_l\}_{l=1}^{L}$, in which $q_{\mathbf{M}_l}(\mathbf{w}_l) = \mathbf{M}_l \cdot [\text{ Bernoulli } (1 - p_l)^{K_l}]$ and $q_\theta(\mathbf{w}) = \prod_l q_{\mathbf{M}_l}(\mathbf{w}_l)$ for a single-random weight matrix, $\mathbf{w}_l$, with K_{l+1} by K_l dimension.

3.4 Model Uncertainties

Uncertainty quantification is a crucial aspect of predictive modeling, which measures how certain is the model. In Bayesian modeling, two fundamental types of uncertainties, namely Aleatoric and Epistemic are often considered [61]. Although DL methods have proven to be powerful for generating accurate predictions, they may not be able to express the model's uncertainty. Therefore, Bayesian deep learning offers a mechanism to overcome this limitation by allowing the network to express uncertainty in its predictions [62]. By incorporating Bayesian principles into the network, it can learn to say, "I don't know" when it is uncertain about a prediction. To this end, we propose a Bayesian CNN model for DR classification. We apply VI approximation and MC-Dropout in the CNN model to estimate its uncertainty. The posterior distribution of the neural network's weights is analyzed to quantify uncertainty in Bayesian modeling. The mean probability predictive is computed using Eq.1 by marginalizing the approximate posterior distribution of the wights $q(w)$ with MC integration over T samples [49].

$$
\begin{aligned}
p(\grave{y} = c|\boldsymbol{x}^*, \mathcal{D}) \quad & \approx \int p(\grave{y} = c|\boldsymbol{x}^*, \mathbf{w}) q_\theta(\mathbf{w}) d\mathbf{w} \\
& \approx \frac{1}{T}\sum_{t=1}^{T} p(\grave{y} = c|\boldsymbol{x}^*, \grave{\mathbf{w}}_t) \\
& \approx \frac{1}{T}\sum_{t=1}^{T} \grave{p}_{c_t} = \overleftharpoon{p}_c
\end{aligned}
\tag{4}
$$

Where $\grave{\mathbf{w}}_t$ is a vector of estimated weights, and c indicates the true class. Moreover, the final classifier is set using Eq. 4 to assign the category based on the greatest mean prediction probability. Therefore, it is essential to state that in

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

inference for each input to the trained NN, the prediction was repeated T times for both MC Dropout and Flipout. Total variance and predictive entropy are used to estimate the uncertainty of Bayesian CNN. The average-amount of knowledge present in the predictive distribution is measured by predictive-entropy, a well-known uncertainty measure, which is provided by:

$$H_p(\hat{y}|\boldsymbol{x}^*) = -\sum_c \bar{p}_c \log \bar{p}_c \tag{5}$$

Where Hp can be normalized to lie between (0 &1) by dividing by $\log 2^C$ [63] as shown in Eq. 6.

$$H_p^*(\hat{y}|\boldsymbol{x}^*) = -\sum_c \bar{p}_c \frac{\log \bar{p}_c}{\log 2^C} \tag{6}$$

Our proposed approach offers a principled framework to incorporate uncertainty into DL models, allowing the network to indicate its uncertainty when making predictions. This is particularly relevant in medical applications such as DR classification, where the ability to quantify uncertainty can aid clinical decision-making. The effectiveness of the proposed approach is demonstrated through experimental results in Section 4.

3.5 Evaluating Metrics

Once the developed models have been implemented and trained, the subsequent step involves applying evaluation metrics to assess their performance. In this work, several evaluation methods are employed. First, comparison evaluation metrics such as accuracy, recall, precision, and F1-score are utilized to compare the predicted and actual values. Secondly, a normalized confusion matrix is presented, which provides a comprehensive of the model's classification performance. Thirdly, the performance of the model's classification task is analyzed using the Receiver Operating Characteristic (ROC) curve and the area under the ROC (AUC). The ROC curve graphically represents the trade-off between true positive and false positive rates, while the AUC provides a quantitative measure of the model's classification performance. Finally, we used entropy as a measure of uncertainty in data prediction for Bayesian CNN model. Entropy is a measure of the disorder or randomness in the data, and it is used for quantifying uncertainty associated with the prediction values. Overall, these evaluation metrics provide a robust framework for assessing the performance of developed models in this study. The metrics used are given as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 - Score = \frac{2 * \ Precision \ * \ Recall}{Precision \ + \ Recall} \tag{10}$$

Where TP, TN, FP, and FN are the four possible outcomes of classification of predicting the presence or absence of DR in medical images. These outcomes represent the following:

- True Positive (TP): The model predicts a positive outcome (DR is present) and it is correct. This means that the patient has DR and the model correctly identifies it.
- True Negative (TN): The model predicts a negative outcome (DR is not present) and it is correct. This means that the patient does not have DR and the model correctly identifies it.
- False Positive (FP): The model predicts a positive outcome (DR is present) but it is incorrect. This means that the patient does not have DR, but the model erroneously identifies it as present.
- False Negative (FN): The model predicts a negative outcome (DR is not present) but it is incorrect. This means that the patient has DR, but the model fails to identify it.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

These outcomes are important because they are used to calculate evaluation metrics such as sensitivity, specificity, accuracy, and F1 score, which provide insights into how well the model is performing on the task at hand.

For the uncertainty quantification, we used Entropy metric, which is a measure of uncertainty or randomness in data prediction for the Bayesian model (as described in Sec 3.4). The entropy value ranges from 0 to 1, and represents the level of uncertainty and disorder in the dataset based on the number of categories present. The primary goal of this model was to reduce the uncertainty while maintaining as low entropy as possible [63].

4. Experimental Results and Discussion

In this section, we present the results of our experimental evaluation of the proposed method and discuss its effectiveness. The goal of this analysis was to assess the potential of Bayesian CNN model for classification and quantifying uncertainty associated with the classification model to achieve the stated objectives, and to identify areas for improvement or further investigation. Through a series of carefully designed experiments, we explore various aspects of the method's performance and provide an in-depth analysis of the findings. The results are discussed in the context of existing literature, and implications for future research are highlighted.

4.1 Data Prepossessing

In this study, we used data processing using Python, and various libraries such as TensorFlow, TensorFlow probability on Jupyter Notebook. First, we imported the images dataset using OpenCV transform functions, and subsequently, the dataset was divided into three parts, namely training (60%), testing (20%), and validation (20%). To facilitate direct handling of these datasets, a new directory was created using “os.makedirs To ensure uniformity in the dataset, which comprised images of varying sizes, we resized all images to have the same width and height (224 x 224 pixels). Finally, to eliminate feature bias and achieve a uniform distribution across the dataset, we normalized all input image intensity values to range between 0 and 1.

4.2 Training Process for CNN Model

Once the new dataset was generated through preprocessing, the CNN model was trained on 100 epochs using the Adam optimization algorithm with different values of learning rate (Lr) ranging from 0.001 to 1e-5. Categorical cross entropy was applied as the loss function to evaluate the model's performance during training. For multi-class classification of the input images, the output layer of the CNN model consisted of a convolution layer with the Softmax activation function to classify the five levels of diabetic retinopathy. Following numerous experiments, we determined the optimal optimization algorithm and learning rate values that yielded the highest accuracy of our CNN model. To provide further details, we summarize the hyperparameters configuration and input images size used in the model in Table 2. The training process involved iterative experimentation with various hyperparameters to achieve the best performance in classifying diabetic retinopathy.

4.3 Training Process for Bayesian CNN Model

To train the BCNN model, we utilized the same base CNN architecture that we developed earlier. For the MC-Dropout model, we used a 0.2 dropout probability to all layers of MC-Dropout, and trained the model for 550 epochs. On the other hand, for the VI model, we used TensorFlow Probability and implemented Convolutional 2DFlipout while initializing the convolutional kernel's prior and posteriors with default values. As the classifications were one-hot encoded, we used the OneHotCategorial layer as the output layer in the VI model, and the model was trained for 200 epochs. Adam was also used as an optimizer, with a learning rate of (1e-3) for VI and (1e-5) for MC-Dropout. The batch-size was set to 32. The loss function used in the VI model was negative log-likelihood (nll) because the output is a distribution. Table 2 summarizes the hyperparameters configuration and input image size used in our models. The training process for the Bayesian CNN models involved implementing dropout and uncertainty estimation techniques to improve the model's accuracy and performance in classifying diabetic retinopathy.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

Table 2: Hyperparameters configuration of the proposed CNN and BCNN models.

Configuration	**CNN**	**BCNN-MC**	**BCNN-VI**
Epochs	100	550	200
Optimization method	Adam	Adam	Adam
Learning-rate	(1e-5)	(1e-5)	(1e-3)
Input size	(224*224*3)	(224*224*3)	(224*224*3)
Loss-function	Categorical-cross-entropy	Categorical-cross-entropy	negative log-likelihood
Activation-function	Softmax and Relu	Softmax and Relu	Relu

4.4 Models Performances

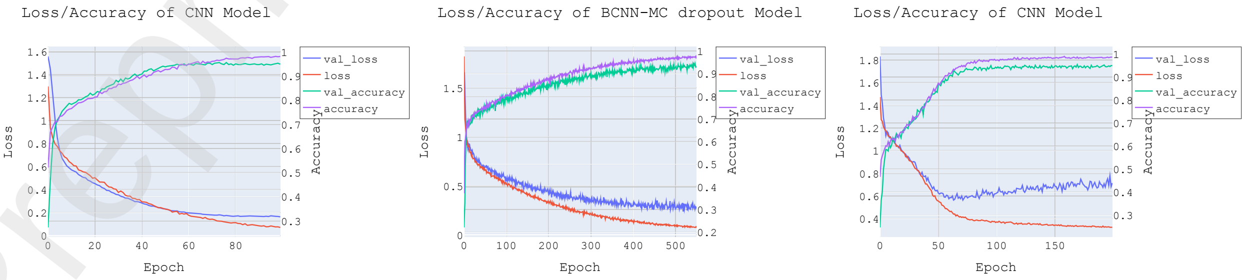
In this section, we present the performance evaluation of our proposed models for DR classification. We employed a dataset of retinal images and utilized the proposed models to evaluate their prediction accuracy and provide an analysis of model uncertainty. Then, we compared the performance of our proposed models with the state-of-the-art studies in DR classification.

4.4.1 Classification Performance of the Proposed Models

The classification performance for diagnosing diabetic retinopathy using simple CNN and Bayesian CNN (BCNN) models are shown in Table 3. We assess the classification performance of these models in terms of accuracy, recall, precision, and F1-score on both training and testing data. Additionally, the learning curves of CNN and BCNN models in terms of accuracy and loss are depicted in Figure 4(a-c). Our results show that the CNN model outperform the BCNN models slightly in terms of accuracy, achieving rates of 94.7%, 93.3%, and 94% for CNN, BCNN MC-Dropout, and BCNN-VI, respectively. This is attributed to the CNN's architecture having less information blocking. However, BCNN-VI outperforms CNN and BCNN MC-Dropout in terms of recall and F1-score, achieving precision, recall, and F1-score of 90.4%, 91.2%, and 90.8%, respectively. In contrast, CNN achieved precision, recall, and F1-score of 91.6%, 89.4%, and 90.4%, respectively, and BCNN MC-Dropout achieved precision, recall, and F1-score of 89.2%, 85.2%, and 87.2%, respectively. These results demonstrate that BCNN-VI model outperforms CNN and BCNN MC-Dropout in term recall and F1-score, despite having slightly lower accuracy. The learning curves depicted in Figure 4(a-c) show that the validation accuracy of BCNN-VI gradually increases and approaches that of CNN over the training epochs, further supporting our performance evaluation. Overall, these results highlight the efficacy of the proposed BCNN-VI model in the classification of DR and its potential for improving disease diagnosis and treatment outcomes.

Table 3: Performance of CNN and BCNN models for classification of DR.

Model	Train Acc (%)	Train Loss	Test Acc (%)	Test Loss	Precision (%)	Recall (%)	F1-score (%)
CNN	98	0.08	**94.7**	0.23	**91.6**	89.4	90.4
BCNN-MC Dropout	97	0.07	93.3	0.3	89.2	85.5	87.2
BCNN-VI	98	0.3	94	0.7	90.4	**91.2**	**90.8**



(a)

(b)

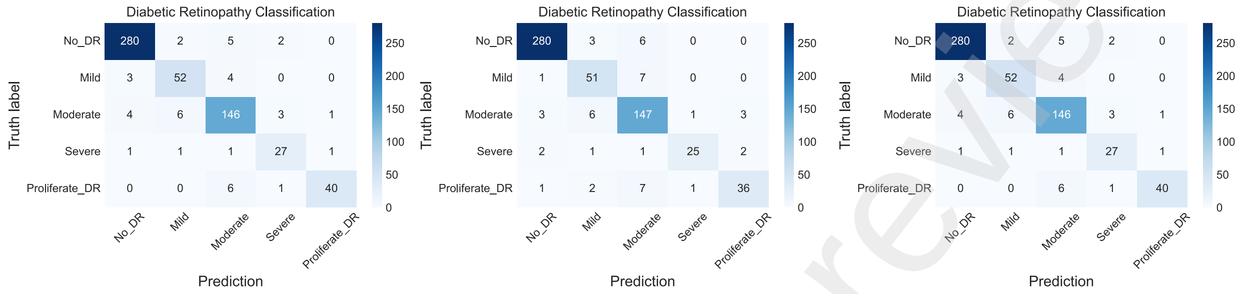
(c)

Figure 4: Accuracy and loss graphs during validation and training for the proposed models.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

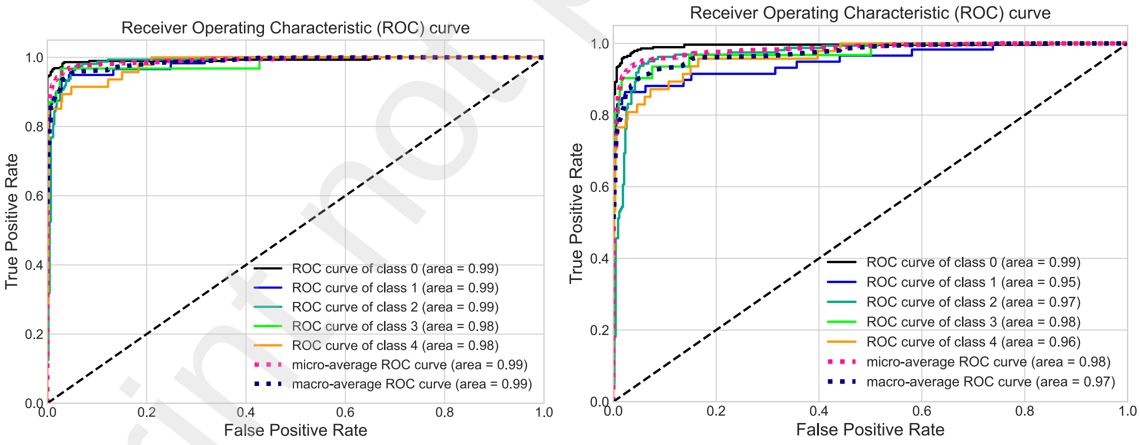
The performances evaluation of the three different models are further illustrated via confusion matrices in Figure 5 (a–c) and the ROC curves in Figure 6 (a-c). The confusion matrices in Figure 5(a) demonstrates that the CNN model produced lower values of false-negative and false-positive for multiclassification applied to testing data. Besides this, Figures 5(b) and 5(c) illustrate the confusion matrices for BCNN MC-Dropout and BCNN-VI, respectively, but the values of false-negative and false-positive in these models are slightly increased. Moreover, we observed from Figures 5(a), 5(b), and 5(c) that the NO-DR class has the highest value of the correct classification rate, and the severe class has the lowest value of the correct classification rate compared to the other severity levels.



(a) CNN (b) BCNN-MC Dropout (c) BCNN-VI

Figure 5: Confusion matrix of the proposed models for test data.

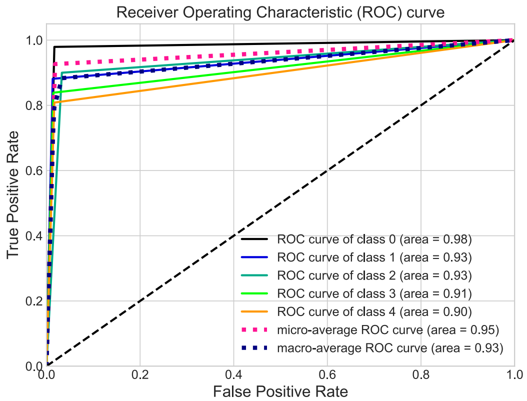
The ROC curves in Figure 6(a-c) illustrate how close the prediction of each class resembles perfect classification, with the AUC value representing the area under the curve. The AUC values for the five classes (No-DR, Mild, Moderate, Severe, and Proliferate-DR) are highest for No-DR (100%) and Moderate (99%), followed by Mild (98%) and Proliferate-DR (98%), and lowest for Severe (97%). A higher AUC value indicates better model performance, and the results suggest that the models perform well in predicting the absence of diabetic retinopathy but struggle with the severe class.



(a) CNN (b) BCNN-MC Dropout

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872



(c) BCNN-VI

Figure 6: ROC curve of the proposed models for testing data.

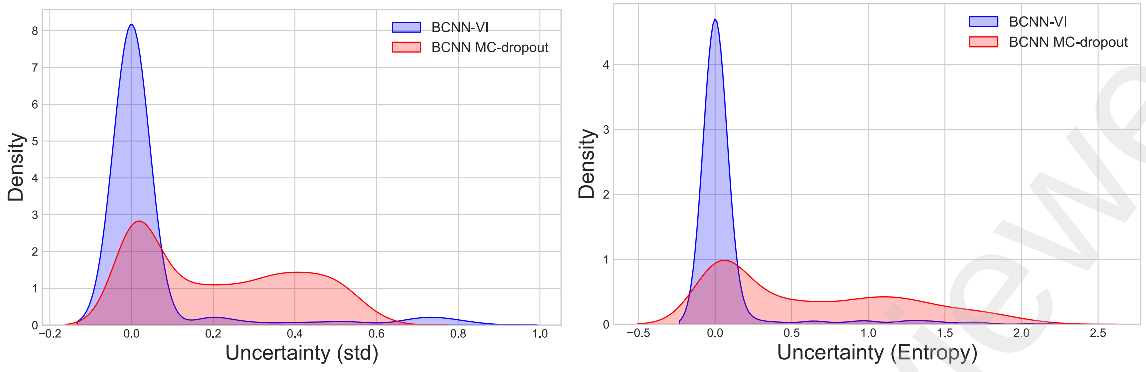
4.4.2 Uncertainty Quantification of BCCN Models

In this section, we delve into the crucial topic of quantifying the uncertainty that arises from the employment of two classification models, namely BCNN MC-Dropout and BCNN-VI. For the BCNN MC-Dropout model, the first step in quantifying model uncertainty was to select an appropriate MC dropout rate, which requires striking a delicate balance between overfitting and underfitting the training data. The MC dropout rate is a hyperparameter that controls the probability of dropping out a neuron in the network during training. This technique was used to prevent overfitting and improve the model's generalization performance. However, if the dropout rate is too small, the model may overfit to the training data, while if it is too large, the model may underfit and fail to capture important patterns in the data. Based on our experiments, we have determined that a dropout rate of 0.20 was optimal in this method. To measure the uncertainty of the models, we have employed the calculation of both entropy and standard deviation (SD) of the predictive distribution for both BCNN models. The entropy provides a gauge of the amount of uncertainty in the model's predictions, while the SD serves as an estimator of the variance of the predictive distribution. These quantifications can facilitate the assessment of the model's reliability and discern cases where the model may require further refinement.

The uncertainty quantifications for these two models are shown in Figures 7 (a and b). The results reveal that the BCNN-VI model exhibits a lower level of uncertainty compared to the BCNN MC-Dropout model, which may indicate that the VI method used to train the BCNN-VI model is more effective than the MC-Dropout for capturing the underlying uncertainty in the data. Additionally, Figure 8 demonstrates the uncertainty per class, indicating that class 3 with the red color representation has the lowest level of uncertainty among other classes. Overall, we observed that the classification models have high uncertainty, and our proposed methods have successfully quantified such uncertainty associated with the modes. The high uncertainty observed in these models, coupled with the low confidence in the classification models, highlights the necessity for further improvements in the models to reduce the uncertainty. This is particularly crucial in medical applications, where the integration of BNN models into diagnostic systems can facilitate the human clinical workflow. Future work could explore alternative uncertainty quantification techniques, such as ensemble models or advanced Bayesian inference approaches, to evaluate their performance in this context.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

Preprint not peer reviewed



(a) (b)

Figure 7: Uncertainty distribution for model predictions on test data, based on (a) Entropy and (b) SD.

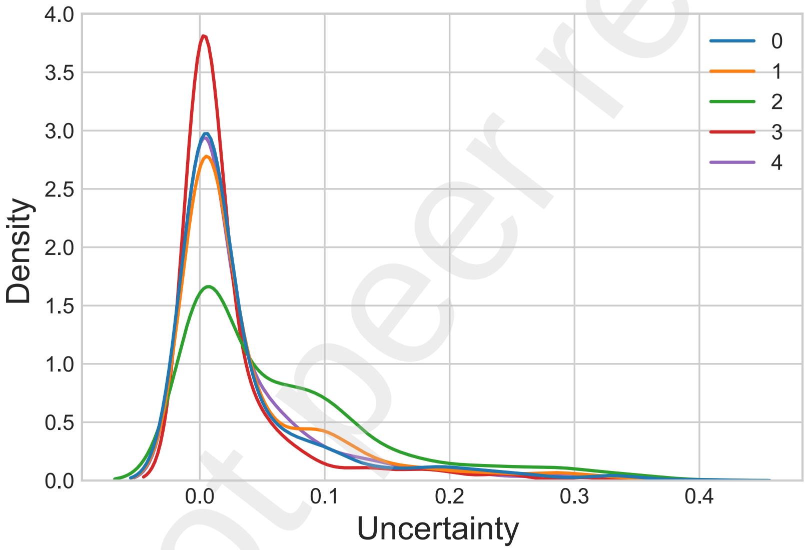


Figure 8: Predictive posterior distribution for each class during test data.

To further investigate the effectiveness of our proposed method, we analyzed the outputs and calculated probabilistic model predictions by examining some samples from the test dataset and evaluating the probability of the model's predictions for each class. The results are presented in Figure 9 (a - f). We represented the certainty of the model's predictions using a bar graph with a green color indicating the true class. The shorter bars indicate a lower level of certainty, while taller bars represent higher levels of certainty.

In our analysis, we found that the model probabilities assigned to some samples were not certain enough, even though they were assigned to a specific class. For instance, Figure 9 (a) shows that the model assigned the highest probability to the true Proliferate-DR class, but with a high uncertainty rate. In contrast, the model assigned a low probability to the Moderate class in Figure 9 (b), indicating a high level of uncertainty in the prediction. Additionally, the estimated probabilities for Figure 9 (d) are accurate, because the highest probability was assigned to the true class (Moderate) with high certainty, and the probabilities of surrounding classes are lower.

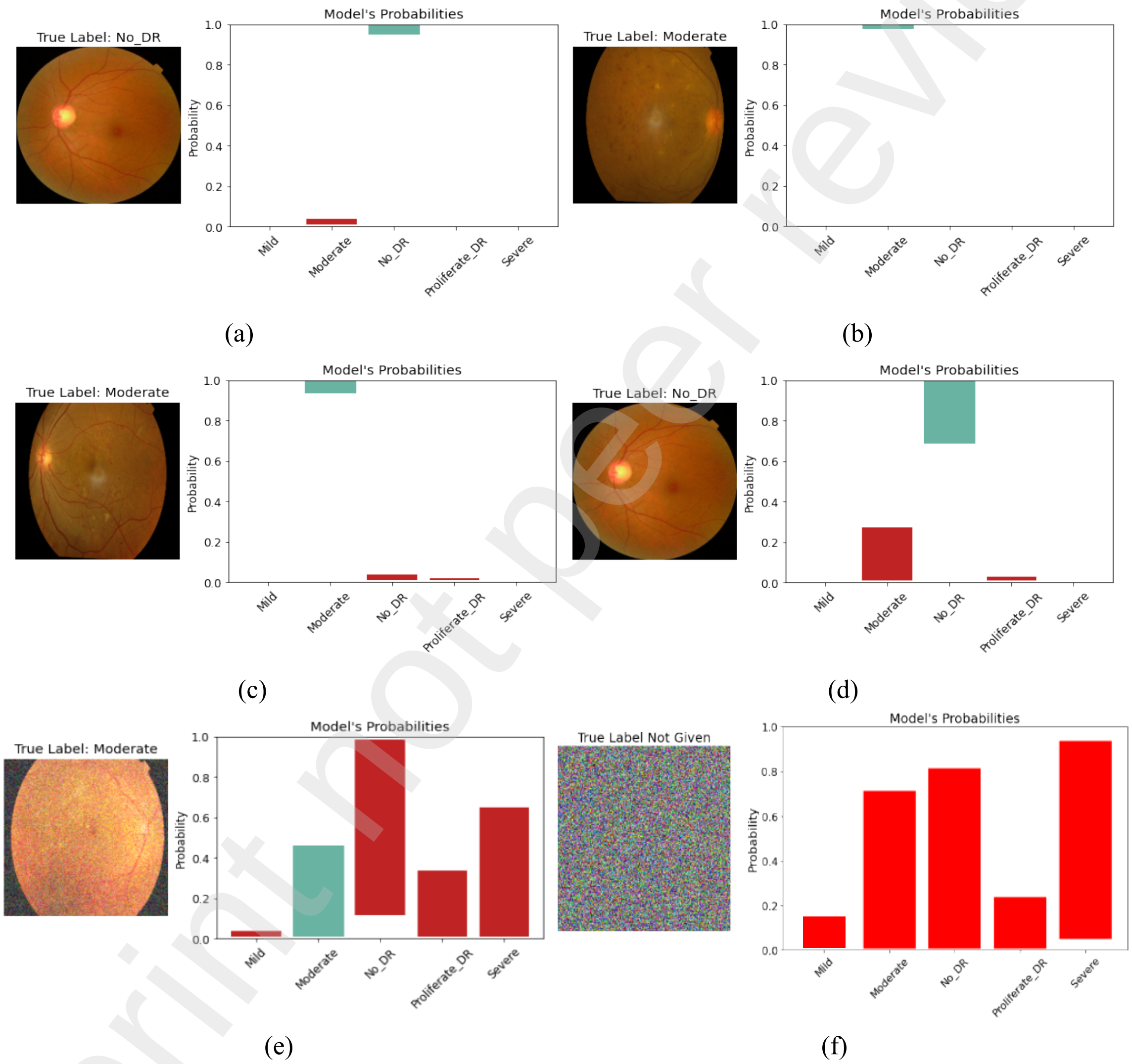
To demonstrate that the Bayesian model can identify its uncertainty, we introduced random noise vectors to the test images to observe the effect on the model's predictions. Figure 9 (e) shows that the model predicted the sample belonged to the Moderate class, but with a much greater level of uncertainty. Moreover, in Figure 9 (f), we increased the noise level, and a normal CNN model would still generate high output probabilities for certain classes. In contrast,

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

the Bayesian model could not make a reliable prediction, and there was no assigned class for this image. Therefore, the Bayesian model admits that it cannot properly classify this image, as indicated by the absence of any prediction. In other words, when a Bayesian model encounters a situation where the evidence is ambiguous or inconclusive, it can express its uncertainty by assigning a range of probabilities to different outcomes or by explicitly stating that it does not know the answer.

Overall, these results highlight the advantages of using Bayesian models, as they can identify their uncertainty and indicate when their predictions may be unreliable. By reducing uncertainty, we can enhance the accuracy of the model, which is critical in medical applications, where the models are used to aid in diagnosis and clinical decision-making.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 9: Probabilistic model predictions with uncertainty for single test samples.

4.4.3 Comparison with the state-of-the-art Studies

To show the efficiency of the proposed method, we compare its results with recent state-of-the-art studies on the diabetic retinopathy (DR) dataset. We present the comparison in Table 4, where all works used the same images from the DR dataset. Our results show that the CNN model slightly outperforms the BCNN MC-Dropout and BCNN-VI models, achieving an accuracy of 94.7%, while other studies, [26], [30] and [24], achieved an accuracy of 86.5%, 90%, and 93.5%, respectively. The results of our proposed CNN model outperformed all other models developed in the literature. The non-Bayesian models presented in Table 4, such as SEDenseNet in [26], a Hybrid method based on

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

CNN and two VGGNets (VGG16 and VGG19) in [30], and the RA-EfficientNet in [24], also performed well, but not as well as our proposed Bayesian models.

As for the Bayesian models, the findings indicate that our Bayesian approach outperforms recent studies in terms of accuracy values. We noted from the results in Table 4 that the performance of our BCNN MC-Dropout method demonstrated superior performance compared to other works [36], [33], [32], achieving an accuracy of 93.3% compared to those studies that used MC-Dropout. They achieved an accuracy of 87.1%, 85%, and 88.2% in [36], [33], [32], respectively, which were lower than our achieved accuracy. Moreover, when comparing the performance of the Variational Inference method, our BCNN-VI method with an accuracy of 94% demonstrated superior performance compared to the Mean-Field VI method with an accuracy of 81.1% [36]. Notably, and there were not significant difference in the results for our VI method, Mean-Field VI, and Mean-Field VI Ensemble methods [32].

Our results demonstrate that the methods we used, particularly Bayesian-CNN, are effective for classifying diabetic retinopathy. While the accuracy of traditional CNN is slightly higher, the use of Bayesian methods allows for the quantification of model uncertainty, and BCNNs are more computationally efficient, making them better suited for large-scale applications, particularly in medical contexts. Moreover, BCNNs can be used to identify potential sources of bias in the model, enabling more accurate predictions. Additionally, BCNNs can help identify areas of data that are not well-represented in the training data, enabling more robust models to be developed, thus reducing the risk of overfitting and enhancing the overall performance of the model. By presenting our results in terms of accuracy, precision, recall, F1-score, and ROC curve and comparing them to state-of-the-art works, we can confidently rely on the methods we used for diabetic retinopathy classification.

Table 4: Comparison of the proposed models with existing studies for DR classification on the Aptos2019 dataset.

Reference	Year	Method	Accuracy
Islam et al. **[29]**	2022	SCL (Supervised Contrastive Learning)	84%
Alyoubi et al. **[23]**	2021	CNN512 + Dropout	84%
Gangwar & Ravi **[25]**	2021	Hybrid Inception-ResNet-v2	82.18%
Alahmadi **[22]**	2022	Use the mechanism of recalibration of style and content by DL	85%
Majumder & Kehtarnavaz **[26]**	2021	SEDenseNet	86%
Menaouer et al. **[30]**	2022	Hybrid method based on CNN and two VGGNet	90%
Yi et al. **[24]**	2021	RA-EfficientNet	93.5%
Filos et al. **[36]**	2019	Ensemble MC Dropout Mean-Field VI	87.1% 81.1%
Ahsan et al. **[33]**	2020	MC-Dropout	85%
Band et al. **[32]**	2020	MC-Dropout Mean-Field VI MC- Dropout Ensemble Mean-Field VI Ensemble Deep Ensemble	88.2% 92.7% 87.7% 94% 90.1%
Our Proposed Models		Simple CNN BCNN (VI) BCNN (MC-Dropout)	94.7 % 94 % 93.3%

5. Conclusion

In this work, we aimed to develop a CNN model for the classification of diabetic retinopathy. We combined the developed CNN model with two Bayesian approximation methods, MC-Dropout and VI, to create Bayesian CNN models that can quantify model uncertainty by analyzing the posterior predictive distribution. We discussed how uncertainty can be modeled in order to detect and classify DR in retinal images. Our experimental results on the Aptos 2019 dataset demonstrate the effectiveness of our proposed models in terms of accuracy and the ability to provide

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

uncertainty in predictions. Furthermore, the developed models not only produced high accuracy in classifying DR but also provided uncertainty quantification in predictions. The proposed models achieved high accuracy, with CNN achieving an accuracy of 94.7%, BCNN-VI achieving an accuracy of 94%, and MC-Dropout achieving an accuracy of 93%. Moreover, our models outperformed the state-of-the-art works in classifying diabetic retinopathy.

In future works, we aim to use different datasets for diabetic retinopathy to further validate our proposed models. We will explore ways to reduce and improve model uncertainty and investigate the potential to incorporate uncertainty estimations into the network as feedback information to enhance its predictive capability. Overall, our study contributes to the growing body of research on Bayesian models in medical image analysis and provides insights into the use of uncertainty quantification for improved disease diagnosis and treatment.

Reference

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.

[2] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Sci. Rep.*, vol. 7, no. 1, Art. no. 1, Dec. 2017, doi: 10.1038/s41598-017-17876-z.

[3] H. D. Kabir, A. Khosravi, M. A. Hosen, and S. Nahavandi, "Neural network-based uncertainty quantification: A survey of methodologies and applications," *IEEE Access*, vol. 6, pp. 36218–36234, 2018.

[4] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015.

[5] A. A. Abdullah, M. M. Hassan, and Y. T. Mustafa, "A Review on Bayesian Deep Learning in Healthcare: Applications and Challenges," *IEEE Access*, vol. 10, pp. 36538–36562, 2022, doi: 10.1109/ACCESS.2022.3163384.

[6] P. Kaur, A. Singh, and I. Chana, "BSense: A parallel Bayesian hyperparameter optimized Stacked ensemble model for breast cancer survival prediction," *J. Comput. Sci.*, vol. 60, p. 101570, Apr. 2022, doi: 10.1016/j.jocs.2022.101570.

[7] R. Jena and S. P. Awate, "A Bayesian Neural Net to Segment Images with Uncertainty Estimates and Good Calibration," in *Information Processing in Medical Imaging*, Cham, 2019, pp. 3–15. doi: 10.1007/978-3-030-20351-1_1.

[8] E. Hüllermeier and W. Waegeman, "Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, Mar. 2021, doi: 10.1007/s10994-021-05946-3.

[9] Y. Gal and Z. Ghahramani, "Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference." arXiv, Jan. 18, 2016. Accessed: Feb. 28, 2023. [Online]. Available: http://arxiv.org/abs/1506.02158

[10] M. A. Kupinski, J. W. Hoppin, E. Clarkson, and H. H. Barrett, "Ideal-observer computation in medical imaging with use of Markov-chain Monte Carlo techniques," *JOSA A*, vol. 20, no. 3, pp. 430–438, 2003.

[11] T. Salimans, D. P. Kingma, and M. Welling, "Markov Chain Monte Carlo and Variational Inference:Bridging the Gap," *Bridg. Gap*.

[12] T. Huix, S. Majewski, A. Durmus, E. Moulines, and A. Korba, "Variational Inference of overparameterized Bayesian Neural Networks: a theoretical and empirical study." arXiv, Jul. 08, 2022. Accessed: Mar. 01, 2023. [Online]. Available: http://arxiv.org/abs/2207.03859

[13] A. Graves, "Practical variational inference for neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 24, 2011.

[14] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[15] A. Shamsi *et al.*, "Improving MC-Dropout Uncertainty Estimates with Calibration Error-based Optimization." arXiv, Oct. 07, 2021. Accessed: Nov. 22, 2022. [Online]. Available: http://arxiv.org/abs/2110.03260

[16] R. Muc, A. Saracen, and I. Grabska-Liberek, "Associations of diabetic retinopathy with retinal neurodegeneration on the background of diabetes mellitus. Overview of recent medical studies with an assessment of the impact on healthcare systems," *Open Med.*, vol. 13, no. 1, pp. 130–136, 2018.

[17] M. D. Abràmoff *et al.*, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Invest. Ophthalmol. Vis. Sci.*, vol. 57, no. 13, pp. 5200–5206, 2016.

[18] "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs | Diabetic Retinopathy | JAMA | JAMA Network." https://jamanetwork.com/journals/jama/fullarticle/2588763/ (accessed Feb. 24, 2023).

[19] D. S. W. Ting *et al.*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *Jama*, vol. 318, no. 22, pp. 2211–2223, 2017.

[20] J. Sahlsten *et al.*, "Deep learning fundus image analysis for diabetic retinopathy and macular edema grading," *Sci. Rep.*, vol. 9, no. 1, p. 10750, 2019.

[21] H. R. Ismail and M. M. Hassan, "Bayesian deep learning methods applied to diabetic retinopathy disease: a review," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 30, no. 2, Art. no. 2, May 2023, doi: 10.11591/ijeecs.v30.i2.pp1167-1177.

[22] M. D. Alahmadi, "Texture Attention Network for Diabetic Retinopathy Classification," *IEEE Access*, vol. 10, pp. 55522–55532, 2022, doi: 10.1109/ACCESS.2022.3177651.

[23] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning," *Sensors*, vol. 21, no. 11, p. 3704, May 2021, doi: 10.3390/s21113704.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

[24] S.-L. Yi, X.-L. Yang, T.-W. Wang, F.-R. She, X. Xiong, and J.-F. He, “Diabetic Retinopathy Diagnosis Based on RA-EfficientNet,” *Appl. Sci.*, vol. 11, no. 22, p. 11035, Nov. 2021, doi: 10.3390/app112211035.

[25] A. K. Gangwar and V. Ravi, “Diabetic Retinopathy Detection Using Transfer Learning and Deep Learning,” in *Evolution in Computational Intelligence*, vol. 1176, V. Bhateja, S.-L. Peng, S. C. Satapathy, and Y.-D. Zhang, Eds. Singapore: Springer Singapore, 2021, pp. 679–689. doi: 10.1007/978-981-15-5788-0_64.

[26] S. Majumder and N. Kehtarnavaz, “Multitasking Deep Learning Model for Detection of Five Stages of Diabetic Retinopathy,” *IEEE Access*, vol. 9, pp. 123220–123230, 2021, doi: 10.1109/ACCESS.2021.3109240.

[27] M. T. Al-Antary and Y. Arafa, “Multi-Scale Attention Network for Diabetic Retinopathy Classification,” *IEEE Access*, vol. 9, pp. 54190–54200, 2021, doi: 10.1109/ACCESS.2021.3070685.

[28] Padmanayana and Dr. A. B.K, “Binary Classification of DR-Diabetic Retinopathy using CNN with Fundus Colour Images,” *Mater. Today Proc.*, vol. 58, pp. 212–216, 2022, doi: 10.1016/j.matpr.2022.01.466.

[29] M. R. Islam *et al.*, “Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images,” *Comput. Biol. Med.*, vol. 146, p. 105602, Jul. 2022, doi: 10.1016/j.compbiomed.2022.105602.

[30] B. Menaouer, Z. Dermane, N. El Houda Kebir, and N. Matta, “Diabetic Retinopathy Classification Using Hybrid Deep Learning Approach,” *SN Comput. Sci.*, vol. 3, no. 5, p. 357, Jul. 2022, doi: 10.1007/s42979-022-01240-8.

[31] J. Jaskari *et al.*, “Uncertainty-aware deep learning methods for robust diabetic retinopathy classification.” arXiv, Feb. 02, 2022. Accessed: May 26, 2022. [Online]. Available: http://arxiv.org/abs/2201.09042

[32] N. Band *et al.*, “Benchmarking Bayesian Deep Learning on Diabetic Retinopathy Detection Tasks,” p. 40.

[33] M. A. Ahsan, A. Qayyum, J. Qadir, and A. Razi, “An Active Learning Method for Diabetic Retinopathy Classification with Uncertainty Quantification.” arXiv, Dec. 26, 2020. Accessed: May 26, 2022. [Online]. Available: http://arxiv.org/abs/2012.13325

[34] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, “Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation,” *Comput. Stat. Data Anal.*, vol. 142, p. 106816, Feb. 2020, doi: 10.1016/j.csda.2019.106816.

[35] S. Toledo-Cortés, M. de la Pava, O. Perdomo, and F. A. González, “Hybrid Deep Learning Gaussian Process for Diabetic Retinopathy Diagnosis and Uncertainty Quantification,” in *Ophthalmic Medical Image Analysis*, Cham, 2020, pp. 206–215. doi: 10.1007/978-3-030-63419-3_21.

[36] A. Filos *et al.*, “A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks.” arXiv, Dec. 22, 2019. Accessed: May 26, 2022. [Online]. Available: http://arxiv.org/abs/1912.10481

[37] S. Farquhar, M. A. Osborne, and Y. Gal, “Radial Bayesian Neural Networks: Beyond Discrete Support In Large-Scale Bayesian Deep Learning,” p. 10.

[38] S. Toledo Cortés, D. Useche Reyes, H. Müller, and F. González, “Grading diabetic retinopathy and prostate cancer diagnostic images with deep quantum ordinal regression,” *Comput. Biol. Med.*, vol. 145, p. 105472, Apr. 2022, doi: 10.1016/j.compbiomed.2022.105472.

[39] N. E. M. Khalifa, M. Loey, M. H. N. Taha, and H. N. E. T. Mohamed, “Deep transfer learning models for medical diabetic retinopathy detection,” *Acta Inform. Medica*, vol. 27, no. 5, p. 327, 2019.

[40] A. Pak, A. Ziyaden, K. Tukeshev, A. Jaxylykova, and D. Abdullina, “Comparative analysis of deep learning methods of detection of diabetic retinopathy,” *Cogent Eng.*, vol. 7, no. 1, p. 1805144, 2020.

[41] “APTOS 2019 Blindness Detection.” https://kaggle.com/competitions/aptos2019-blindness-detection (accessed Feb. 24, 2023).

[42] A. Sungheetha and R. Sharma, “Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural network,” *J. Trends Comput. Sci. Smart Technol. TCSST*, vol. 3, no. 02, pp. 81–94, 2021.

[43] H. R. Ismael, A. M. Abdulazeez, and D. A. Hasan, “Detection of Diabetic Retinopathy Based on Convolutional Neural Networks: A Review,” *Asian J. Res. Comput. Sci.*, vol. 8, no. 3, Art. no. 3, May 2021, doi: 10.9734/ajrcos/2021/v8i330200.

[44] H. Wang and D.-Y. Yeung, “A Survey on Bayesian Deep Learning,” *ACM Comput. Surv.*, vol. 53, no. 5, p. 108:1-108:37, Sep. 2020, doi: 10.1145/3409383.

[45] K. Shridhar, F. Laumann, and M. Liwicki, “A Comprehensive guide to Bayesian Convolutional Neural Network with Variational Inference.” arXiv, Jan. 08, 2019. doi: 10.48550/arXiv.1901.02731.

[46] T. P. Minka, “Expectation propagation for approximate Bayesian inference,” in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, San Francisco, CA, USA, Aug. 2001, pp. 362–369.

[47] S. Chib, “Chapter 57 - Markov Chain Monte Carlo Methods: Computation and Inference,” in *Handbook of Econometrics*, vol. 5, J. J. Heckman and E. Leamer, Eds. Elsevier, 2001, pp. 3569–3649. doi: 10.1016/S1573-4412(01)05010-3.

[48] T. Papamarkou, J. Hinkle, M. T. Young, and D. Womble, “Challenges in Markov Chain Monte Carlo for Bayesian Neural Networks,” *Stat. Sci.*, vol. 37, no. 3, pp. 425–442, Aug. 2022, doi: 10.1214/21-STS840.

[49] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *international conference on machine learning*, 2016, pp. 1050–1059.

[50] M. A. M. Sadr, J. Gante, B. Champagne, G. Falcao, and L. Sousa, “Uncertainty Estimation via Monte Carlo Dropout in CNN-Based mmWave MIMO Localization,” *IEEE Signal Process. Lett.*, vol. 29, pp. 269–273, 2022, doi: 10.1109/LSP.2021.3130504.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872

[51] T. Myojin, S. Hashimoto, and N. Ishihama, “Detecting Uncertain BNN Outputs on FPGA Using Monte Carlo Dropout Sampling,” in *Artificial Neural Networks and Machine Learning – ICANN 2020*, Cham, 2020, pp. 27–38. doi: 10.1007/978-3-030-61616-8_3.

[52] J. Paisley, D. Blei, and M. Jordan, “Variational Bayesian inference with stochastic search,” *ArXiv Prepr. ArXiv12066430*, 2012.

[53] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight Uncertainty in Neural Network,” in *Proceedings of the 32nd International Conference on Machine Learning*, Jun. 2015, pp. 1613–1622. Accessed: Feb. 24, 2023. [Online]. Available: https://proceedings.mlr.press/v37/blundell15.html

[54] D. P. Kingma, T. Salimans, and M. Welling, “Variational Dropout and the Local Reparameterization Trick,” in *Advances in Neural Information Processing Systems*, 2015, vol. 28. Accessed: Feb. 24, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/bc7316929fe1545bf0b98d114ee3ecb8-Abstract.html

[55] Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse, “Flipout: Efficient pseudo-independent weight perturbations on mini-batches,” *ArXiv Prepr. ArXiv180304386*, 2018.

[56] R. Feng, N. Balling, D. Grana, J. S. Dramsch, and T. M. Hansen, “Bayesian convolutional neural networks for seismic facies classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8933–8940, 2021.

[57] S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih, “Monte carlo gradient estimation in machine learning,” *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5183–5244, 2020.

[58] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”.

[59] M. Combalia, F. Hueto, S. Puig, J. Malvehy, and V. Vilaplana, “Uncertainty estimation in deep neural networks for dermoscopic image classification,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 744–745.

[60] Y. Gal, J. Hron, and A. Kendall, “Concrete dropout,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[61] E. Hüllermeier and W. Waegeman, “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods,” *Mach. Learn.*, vol. 110, pp. 457–506, 2021.

[62] M. Abdar *et al.*, “A review of uncertainty quantification in deep learning: Techniques, applications and challenges,” *Inf. Fusion*, vol. 76, pp. 243–297, Dec. 2021, doi: 10.1016/j.inffus.2021.05.008.

[63] L. A. Park and S. Simoff, “Using entropy as a measure of acceptance for multi-label classification,” in *Advances in Intelligent Data Analysis XIV: 14th International Symposium, IDA 2015, Saint Etienne. France, October 22-24, 2015. Proceedings 14*, 2015, pp. 217–228.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4402872