

Leveraging Uncertainty in Deep Learning for Pancreatic Adenocarcinoma Grading

Biraja Ghoshal¹ Bhargab Ghoshal² and Allan Tucker¹

¹ Department of Computer Science, Brunel University, London, UK

² Faculty of Medical Sciences, University College London, UK
Biraja.Ghoshal@brunel.ac.uk

Abstract. Pancreatic cancers have one of the worst prognoses compared to other cancers, as they are diagnosed when cancer has progressed towards its latter stages. The current manual histological grading for diagnosing pancreatic adenocarcinomas is time-consuming and often results in misdiagnosis. In digital pathology, AI-based cancer grading must be extremely accurate in prediction and uncertainty quantification to improve reliability and explainability and are essential for gaining clinicians’ trust in the technology. We present Bayesian Convolutional Neural Networks for automated pancreatic cancer grading from MGG and H&E stained images to estimate uncertainty in model prediction. We show that the estimated uncertainty correlates with prediction error. Specifically, it is useful in setting the acceptance threshold using a metric that weighs classification accuracy-reject trade-off and misclassification cost controlled by hyperparameters and can be employed in clinical settings.

Keywords: Bayesian Deep Learning · Uncertainty Estimation · Cancer grading · Pancreatic Adenocarcinoma.

1 Introduction

Pancreatic adenocarcinoma has one of the worst prognoses compared to other cancers, with a 10% 5 year survival rate, and it is projected to become the second-leading cause of cancer-related mortality by 2030 [31]. Early diagnosis is important to improving the likelihood of survival, which is reflected in the survival rates of 39.4% in patients whose tumours have not metastasised [22]. Intraepithelial neoplasms develop into adenocarcinomas and these are stratified into three progressive categories: Grade I, II and III. Grade I neoplasia comprises of well-differentiated tissue, which do not divide rapidly and so do not respond very well to chemotherapy. Grade II neoplasia represents moderately differentiated tissue, whilst Grade III represents poorly differentiated tissue, which is rapidly proliferating. The grading of neoplasia can be performed using histological samples stained using Haemotxylin and Eosin (H&E) and May-Grunwald-Giemsa stain (MGG). H&E stains nuclear components blue-purple and cytoplasmic components pink, whilst MGG stains help to ascertain the morphology of cells.

Pancreatic cancers are usually detected towards their latter stages in locally advanced (30%-35%) or metastatic (50%-55%) stage as most patients are asymptomatic in the early stages [19]. As a result, pancreatic tumour diagnosis requires urgent action and a definite surgical plan. Pancreatic cancer detection and classification has either focused only on distinguishing benign tumor from malignant ones or tumor presence on radiology images [29]. The diagnostic performance by means of manual examination of pancreatic cancer grading is very tedious, time consuming, depends on clinicians' experience, and often results in misdiagnosis and thus incorrect treatment. There is an urgent need for novel methods to supplement radiologist interpretations in improving the sensitivity of pancreatic cancer detection from histopathology images.

Several methods use deep learning in cancer detection and diagnosis such as the Gleason grading of prostate cancer, colon cancer grading, and breast cancer detection [1,32,7,27]. However, it is critical to estimate uncertainty in medical image analysis as an additional insight to point predictions to improve the reliability in making decisions. Our objective is not to achieve state-of-the-art performance, but rather to evaluate the usefulness of estimating uncertainty approximating Bayesian Convolutional Neural Networks (BCNN) with Dropweights to improve the diagnosis.

In digital pathology, multi-gigapixel Whole Slide digitised histopathology Images (WSI) are divided into small patches because (i) different chemical preparations typically render different slides for the same piece of tissue, (ii) they are generated by different digitisation devices, (iii) the device settings may produce different images from the same slide and (iv) size limitations for CNN image inputs [6].

It is crucial to accurately grade pancreatic cancer in patients where the cost of an error is very high. In order to avoid misdiagnoses, it is necessary to estimate uncertainty in a model's predictions for automatically handling clear-cut diagnoses, whilst elevating difficult decisions to medical professionals, who can request further scans, recognizing the uncertainty and seek assistance.

There are many methods proposed for quantifying uncertainty or confidence estimates in deep learning such as Hamiltonian Monte Carlo, Stochastic Gradient Langevin Dynamic (SGLD), Laplace approximation, Bayes by backprop, Deep ensembles, Monte Carlo (MC) dropout, MC-dropweight and MC-batch normalization [24,10,30,4,28,17,21,14]. There are also many measures to estimate uncertainty such as softmax variance, predictive entropy, mutual information, BALD [20], Uncertainty measure proposed by Leibig [26] and Feinman [8] and averaging predictions over multiple models, which are mostly focused on rejection accuracy and log-likelihood without assessing the quality of predictive uncertainty with calibrated expectations.

In this paper, we propose Bayesian Convolutional Neural Networks (BCNN) approach to predict pancreatic cancer grading from two different May-Grunwald-Giemsa (MGG) and Haematoxylin and Eosin (H&E) stained images and show that the estimated uncertainty in prediction has a strong correlation with classification accuracy, thus enabling the identification of false predictions or unknown cases.

We present the first approach (to the best of our knowledge) in leveraging uncertainty in automated grading of pancreatic cancer based on histopathology images. The proposed Bayesian deep learning model can be very useful to clinicians in diagnosing cancer grading system, which can address the problems in manual grading. We believe that the availability of uncertainty-aware deep learning solution will enable a wider adoption of Artificial Intelligence (AI) in a clinical setting.

2 Proposed Method

2.1 Dataset

Our approach was based on the patch-level prediction to include all tumor histological subtypes, ensuring that the selected patches were widely representative for practical diagnosis in order to be adaptable to meet the input size of most neural networks training and derive the likelihood of neoplasia at patch-level. The patches of approximately 200 X 200 pixels of non-overlapping regions in a WSI were firstly extracted from 138 high-resolution tissue-samples stained with MGG and H&E [33] with varying dimensions annotated with Normal, Grade-I, Grade-II and Grade-III. Overall, 49.5% (3201) of the patches were selected after discarding patches with non-tissue information. The flow chart is shown in Fig. 1. An imbalanced number of patch images in each class i.e. bias towards the majority proportions of cancer cells is reduced by class weight to regularise the loss function.

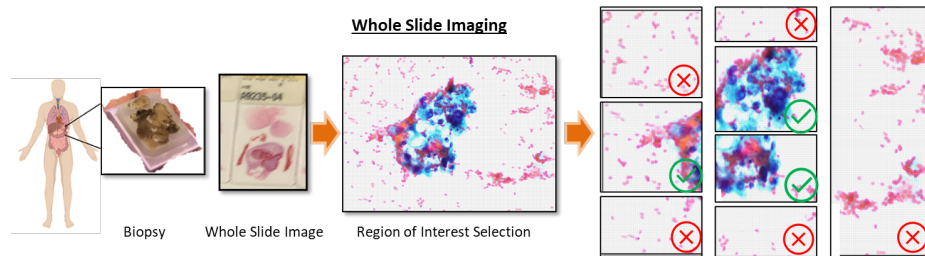


Fig. 1: Whole Slide Image (WSI) processing and selection of tissue patches. Whole slide image was considered by dividing the slide into patches of 200x200 pixels. Each patch was then selected based on tissue presence.

2.2 Approximate Bayesian Neural Networks (BCNN) and Model Uncertainty

During the training phase, histological pancreatic cancer image dataset $X = \{x_1, x_2 \dots x_N\}$ and the corresponding grade $Y = \{y_1, y_2 \dots y_N\}$ where $X \in R^d$ is a d-dimensional input vector and $Y \in \{1 \dots C\}$ with $y_i \in \{1 \dots C\}$, given C class labels (here four: Normal, Grade I, Grade II and Grade III), a set of independent

and identically distributed (i.i.d.) training samples size $N\{x_i, y_i\}$ for $i = 1$ to N , are used to learn the weights of the neural networks. Uncertainty of the model prediction was captured by placing a prior distribution over weights W . The principled predictive distribution of an unknown grading of cancer prediction label \hat{y} of a test input data \hat{x} by marginalizing the parameters:

$$p(\hat{y}|\hat{x}, X, Y) = \int P(\hat{y}|\hat{x}, w)P(w|X, Y, \hat{x})dw \quad (1)$$

Unfortunately, finding the posterior distribution $p(w|X, Y)$ is often computationally intractable. Following Gal, Ghoshal et al. showed that neural networks with dropweights with the cross-entropy loss of the network, is equivalent to a variational approximation by minimising the Kullback-Leibler (KL) divergence on a Bayesian neural network. At test time, the unseen images are passed through the network, the posterior $P(W|X, Y, \hat{x})$ was approximated by averaging stochastic feed forward Monte Carlo (MC) sampling to estimate uncertainty.

Practically, the expectation of \hat{y} is called the predictive mean of the model. The estimate of the vector of Softmax probabilities i.e., the predictive mean μ_{pred} over the MC iterations is then used as the final prediction on the test sample:

$$\mu_{pred} \approx \frac{1}{T} \sum_{t=1}^T p(\hat{y} = c|\hat{x}, \hat{\omega}_t); \quad c \in \{1, \dots, C\} \quad (2)$$

For each test sample \hat{x} , the class with the largest predictive mean μ_{pred} is selected as the output prediction. We present a novel approximation of predictive model uncertainty as below:

$$\sigma_{uncertainty} = \frac{1}{C} \sum_{i=1}^C \sqrt{\frac{1}{T} \sum_{t=1}^T [p(\hat{y}_t = c|\hat{x}, \hat{\omega}_t) - \hat{\mu}_{pred}]^2} \quad (3)$$

$$\text{where } \hat{y}_t = y(\hat{\omega}_t) = \text{Softmax}\{f^{\hat{\omega}_t}(\hat{x})\} \quad (4)$$

In our approximated uncertainty measure in model prediction (i.e. equation 3), we take into account the uncertainty associated with every class in the predictive mean μ_{pred} . Furthermore in the approximation, we take the mean of the standard deviations of the class probabilities, instead of the variance. It assigns the highest average uncertainty to the most frequently mislabelled class.

2.3 Uncertainty Metrics in Deep Learning

We can compute an uncertainty metric from the multiple predictions per input image captured during test time. In this work we compared four well established metrics: predictive entropy; mutual information [9,20], Feinman uncertainty, which measures of uncertainty for each observation by averaging the mean squared prediction error of each class [8]; Leibig uncertainty, which returns the

empirical standard deviation as a proxy for predictive uncertainty [26]; and moment based aleatoric and epistemic uncertainty metrics by Kwon [23]. In addition, we introduce a novel approximation of predictive model uncertainty which averages the standard deviations of the predictions in the predictive mean vector of class probabilities [11,16].

2.4 Experiment

Instead of training a very deep model from scratch on a small dataset, we decided to run this experiment in a transfer learning setting, where we used a pretrained DenseNet-201, VGG-19 and ResNet-152V2 model [18] and acquired data only to fine-tune the original model. We split the whole dataset into 60% - 20% - 20% between training, validation and test sets respectively. This is suitable when the data is abundant for an auxiliary domain, but very limited labelled data is available for the domain of experiment. We introduced fully connected layers on top of the pre-trained convolutional base. Dropweights followed by a softmax activated layer is applied to the network as an approximation to the Gaussian Process (GP) and cast as an approximate Bayesian inference in the fully connected layer to estimate meaningful model uncertainty [15]. The softmax layer outputs the probability distribution over each possible class label. We resized all images to 224 x 224 pixels (using a bicubic interpolation over 4 x 4 pixel neighbourhood). The images [33] were then standardised using the mean and standard deviation values of the MGG and H&E dataset.

Real-time data augmentation was also applied, leveraging Keras ImageDataGenerator during training, to prevent overfitting and enhance the learning capability of the model. Training images were rotated 90 degrees, randomly flipped horizontally and vertically, scaled outward and inward, shifted, and sheared. The Adam optimiser was used with default initial learning rate of $\alpha = 0.01$ and moment decay rate of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. All our experiments were run for 100 epochs and batch size was set to 64. Dropweights with rates of 0.5 were added to the fully-connected layer. We monitored the validation accuracy after every epoch and saved the model with the best accuracy on the validation dataset. During test time, Dropweights were active and Monte Carlo sampling was performed by feeding the input image with MC-samples 50 through the Bayesian Deep Convolutional Neural Networks.

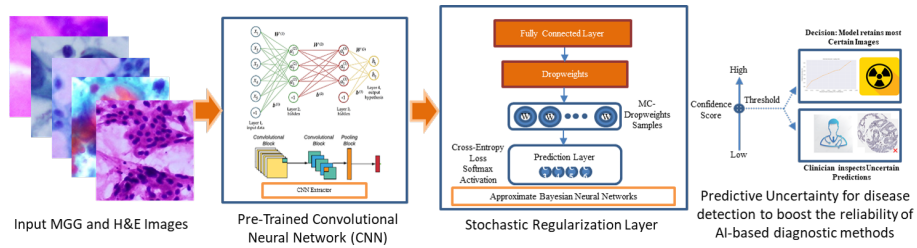


Fig. 2: Overview of the proposed Bayesian Convolutional Neural Network (BCNN) framework.

3 Model Performance

On average, Bayesian DenseNet-201 model based inference improves the detection accuracy of the ResNet-152V2 and VGG-19 model in our sample dataset. Table 1 summarizes the Precision, Recall, F1-Score and prediction accuracy of our implemented models.

Dataset	VGG-19				ResNet-152V2				DenseNet-201			
May Grunwald-Giemsa (MGG)	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy
Normal	79.60%	78.80%	79.00%		80.20%	81.40%	80.60%		85.20%	83.40%	84.00%	
Grade I	53.00%	22.00%	31.50%		65.40%	40.80%	50.80%		72.20%	45.60%	55.40%	
Grade II	66.00%	82.40%	73.20%		75.80%	80.60%	78.20%		78.20%	84.20%	81.20%	
Grade III	50.20%	30.00%	38.00%		65.60%	58.60%	61.20%		72.20%	66.00%	68.80%	
Average Score	62.89%	66.69%	63.66%	66.69%	74.14%	74.12%	73.77%	74.23%	78.19%	78.19%	77.86%	85.53%
Haematoxylin and Eosin (H&E)												
Normal	87.60%	83.40%	85.20%		90.00%	87.80%	88.80%		90.00%	90.00%	90.00%	
Grade I	84.40%	79.60%	81.80%		85.40%	86.80%	86.20%		88.20%	88.40%	88.20%	
Grade II	62.40%	77.40%	68.80%		79.60%	79.60%	79.80%		81.00%	81.40%	81.20%	
Grade III	76.80%	65.80%	70.80%		79.80%	78.60%	79.80%		83.20%	82.60%	82.80%	
Average Score	78.02%	76.51%	76.79%	76.52%	83.34%	83.48%	83.62%	83.60%	85.65%	85.71%	85.61%	85.60%
Mixed (H&E and MGG)												
Normal	66.40%	57.00%	59.00%		79.00%	82.60%	80.60%		82.20%	86.20%	84.20%	
Grade I	79.60%	66.80%	72.60%		84.00%	80.00%	81.80%		85.80%	78.60%	82.00%	
Grade II	57.60%	83.00%	63.80%		75.00%	80.40%	77.80%		76.80%	81.80%	79.20%	
Grade III	40.80%	37.00%	54.50%		72.60%	63.20%	67.60%		75.00%	69.60%	71.80%	
Average Score	60.55%	59.81%	55.12%	59.75%	77.19%	77.16%	77.08%	77.16%	79.35%	79.33%	79.15%	79.29%

Table 1: summarizes the Precision, Recall, F1-Score and prediction accuracy of our implemented models

All models trained with the H&E stained images achieved the highest accuracy and F1-score compared to MGG and mixed. VGG-19 model trained with the mixed dataset performed the lowest compared to H&E and MGG which indicates H&E stained images are easier to learn and achieve better prediction than MGG. All models achieved an exceptional performance in identifying normal tissue from H&E stained images, which could be due to the distinct differences between neoplastic and non-neoplastic tissue. We also observed that all models performed the weakest in Grade-III class prediction compared other grades.

4 Prediction Error vs Uncertainty

4.1 Bayesian Model Uncertainty

We measured the uncertainty associated with the predictive probabilities of the deep learning model by keeping dropweights on during test time. Figure 3 shows that the pancreatic adenocarcinoma grade-specific approximations of uncertainty is consistent with the confusion matrix. Higher uncertainty seems to be associated with slides that tend to be misclassified.

Figure 4 shows the boxplots of estimated uncertainty in deep learning model for pancreatic adenocarcinoma grading.

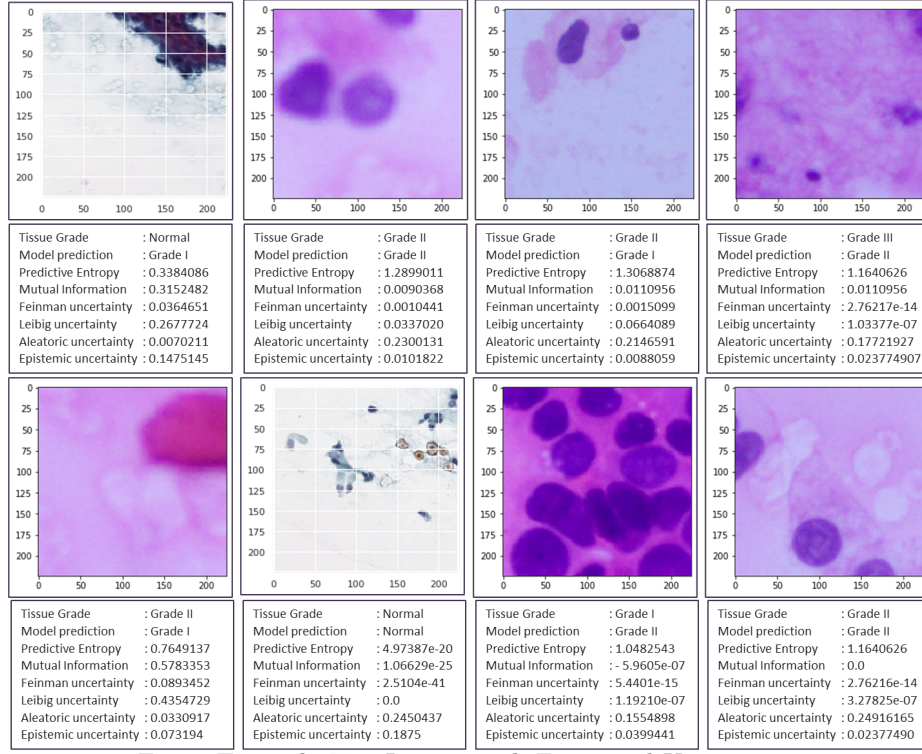


Fig. 3: Example Test Images with Estimated Uncertainty

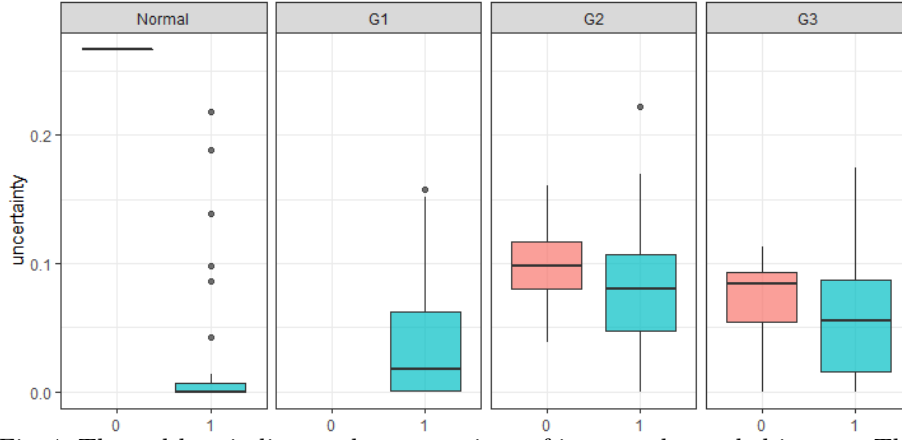


Fig. 4: The red box indicates the uncertainty of incorrectly graded images. The green box corresponds to correctly classified images. Our model assigns the highest average uncertainty to the mislabelled grades and less contribution to high uncertainty values by correctly classified images. Therefore, uncertainty information provides as an additional insight to point prediction to refer the uncertain images to radiologists for further investigation, which improves the overall prediction performance.

4.2 Relationship between the Accuracy and Uncertainty

Trustworthy clinical Artificial Intelligence (AI) systems should not only return accurate predictions, but also a credible representation of their uncertainty. The model's expected accuracy increases as the model's uncertainty in prediction decreases [12].

The true error is the difference between estimated values and actual values. In order to assess the quality of predictive uncertainty, we leveraged Spearman's correlation coefficient. We quantified the predictive accuracy by 1-Wasserstein distance (WD) to measure how much the estimated uncertainty correlates with the true errors [2,25]. The Wasserstein distance for the real data distribution P_r and the generated data distribution P_g is mathematically defined as the greatest lower bound (infimum) for any transport plan (i.e. the cost for the cheapest plan):

$$W(P_r, P_g) = \inf_{\gamma \sim \Pi(P_r, P_g)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|] \quad (5)$$

, $\Pi(P_r, P_g)$ is the set of all possible joint probability distributions $\gamma(x, y)$ whose marginals are respectively P_r and P_g . However, the equation (5) for the Wasserstein distance is intractable. Using the Kantorovich-Rubinstein duality it can be simplified to

$$W(P_r, P_g) = \frac{1}{K} \sup_{\|f\|_L \leq K} \mathbb{E}_{x \sim P_r} [f(x)] - \mathbb{E}_{x \sim P_g} [f(x)] \quad (6)$$

, where sup (supremum) is the least upper bound and f is a 1-Lipschitz continuous functions $\{f_w\}_{w \in W}$, parameterized by w and the K-Lipschitz constraint $|f(x_1) - f(x_2)| \leq K|x_1 - x_2|$. The error function can be configured as measuring the 1-Wasserstein distance between P_r and P_g .

$$E(P_r, P_g) = W(P_r, P_g) = \max_{w \in W} \mathbb{E}_{x \sim P_r} [f_w(x)] - \mathbb{E}_{z \sim P_r(z)} [f_w(g_\theta(z))] \quad (7)$$

The advantage of Wasserstein distance (WD) is that it can reflect the distance of two non-overlapping or little overlapping distributions.

The figure 5 below shows the correlation between estimated uncertainty and the error of prediction and Spearman correlation. The results show strong correlation with $\rho > 0.95$ between entropy of the probabilities as a measure of the epistemic uncertainty and prediction errors.

Our experiments show that the prediction uncertainty correlates with accuracy, thus enabling the identification of false predictions or unknown cases.

4.3 Performance improvement via Uncertainty-Aware Cancer grading Classification

We performed predictions for all test images and sorted the predictions by their associated predictive uncertainty. We then referred predictions based on the various levels of uncertainty for further diagnosis and measured the accuracy of the predictions for the remaining cases. We observed in figure 6, uncertainty

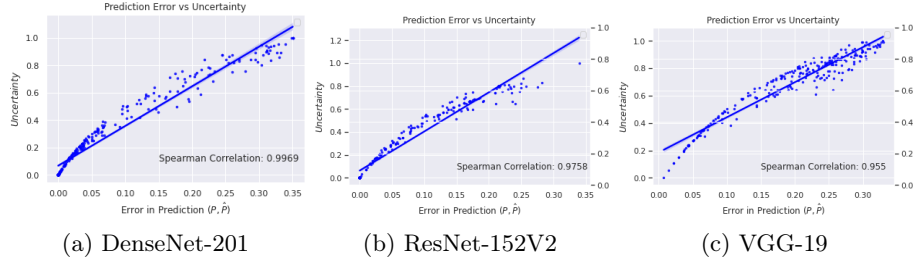


Fig. 5: Correlation between estimated predictive entropy as a measure of Uncertainty and Accuracy in prediction

estimation can usually be used in every image classifier to improve prediction accuracy of man-machine combination via uncertainty-aware referral with the additional computational load cost of performing multiple forward passes.

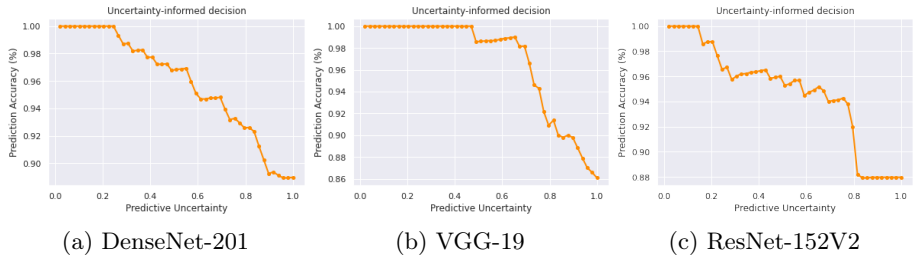


Fig. 6: The classification accuracy as a function of the tolerated normalized model uncertainty

5 Leveraging uncertainty in classification error and reject tradeoff

In safety critical decision making scenarios, the goal of a deep learning model is that the algorithm abstains from making a prediction on the most uncertain images rather than making an incorrect prediction, where the risk of making a incorrect prediction is too large i.e. allowing the model to say "I don't know".

A low model uncertainty ascertained when the probability of predicting into the most possible category has much greater margin over the 2nd most likely category [13]. In rejection, the prediction was ambiguous among all categories because the model failed to reach a definitive conclusion between all the grades.

For example, if the model was presented with a out-of-distribution or non-domain image, it would still classify the image into one of the 4 grades albeit with a high estimated uncertainty measure, which is not an expected outcome. Instead, the model should be able to reject predicting the grade for the test images where the uncertainty is too high. Such abstention is known as classification with a

reject option. In the context of the clinical diagnostic process, accuracy is of paramount importance. The overall decision is the following: the model would accept if the acceptance criteria defined as $\Lambda(x)$ defined as [5,3]:

$$\Lambda(x) := |P(\mathbf{x}^{(i)})_1 - P(\mathbf{x}^{(i)})_2| \geq \epsilon * [\sigma_{uncertainty}(\mathbf{x}^{(i)})_1] \quad (8)$$

where $P(\mathbf{x}^{(i)})$ is the predictive probability for input image $\mathbf{x}^{(i)}$ with $P(\mathbf{x}^{(i)})_j$ being defined as the probability of the j^{th} most likely class, and $U(\mathbf{x}^{(i)})_1$ is the estimated uncertainty of the most likely class returned by the model for input $\mathbf{x}^{(i)}$. Note that the ϵ probability threshold parameter defines a trade-off between the number of classified examples at the number of examples that would have been incorrectly classified but were accepted instead.

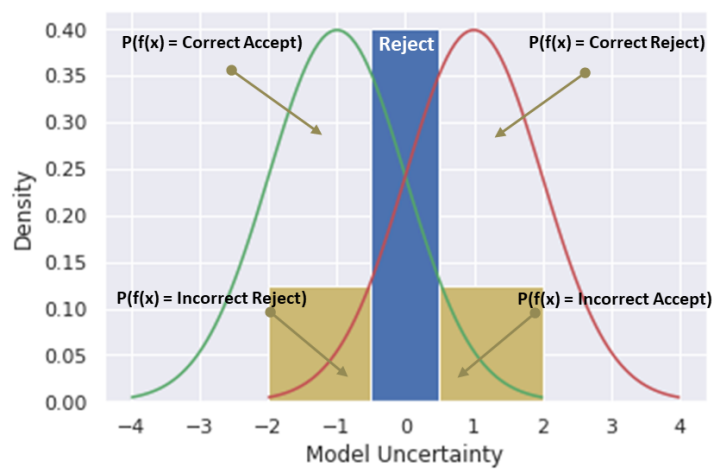


Fig. 7: Distribution of predictive uncertainty values grouped by correct and erroneous predictions

Figure 7 illustrates two class-conditional distribution of predictive uncertainty for the discriminant. The conditional accept or reject with estimated uncertainty depends on uncertainty thresholds.

5.1 Evaluation Metric - Accuracy-Rejection Quotient (ARQ)

The standard metric for evaluating a model f on a classification task is the accuracy:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(P(\mathbf{x}^{(i)}) = y^{(i)}) \quad (9)$$

where N is the number of test data points and $(\mathbf{x}^{(i)}, y^{(i)})$ is the i^{th} test point. Note that the accuracy does not take into account the uncertainty with which a model makes the prediction and only rewards correctly classified examples.

The metric Accuracy-Rejection Quotient (ARQ) for classification with a accuracy-reject trade-off by assigning a cost α to misclassification, and a cost β to the acceptance option $\Lambda(x)$ is defined as [5,3]:

$$\text{ARQ}_{\alpha,\beta} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(P(\mathbf{x}^{(i)}) = y^{(i)}) - \alpha \mathbb{1}(P(\mathbf{x}^{(i)}) \neq y^{(i)}) - \beta \mathbb{1}(P(\mathbf{x}^{(i)}) \neq \Lambda(\mathbf{x}^{(i)})) \quad (10)$$

Here, α and β are hyperparameters. A large value for α corresponds to cases where cost of misclassification is extremely high such as in medical diagnosis. The β intuitively is a trade-off indicator - higher β will decrease the cost of the accept option.

The figure 8 shows the trade-off between the number of misclassified but accepted images and the number of correctly classified but rejected images.

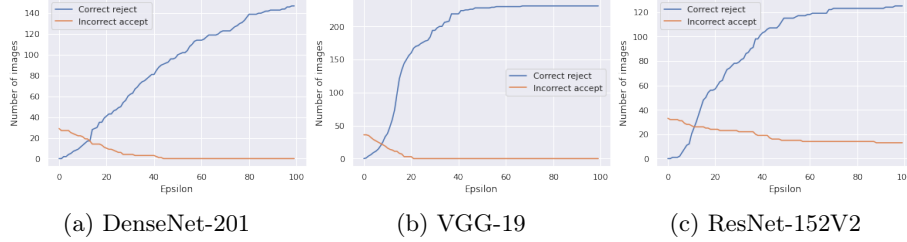


Fig. 8: Accuracy-Rejection Quotient (ARQ) obtained using varying ϵ and α but setting $\beta = 0$. For $\alpha > 0$ there seems to be an optimal value of ϵ around 17 in case of DenseNet-201.

6 Conclusion and Future work

Clinical AI systems are now regularly being used in medical settings although regulatory oversight is inconsistent and undeveloped. Computer based medical systems in clinical setting requires informed users, who are generally responsible for identifying and reporting emerging problems [32]. Understanding the confidence of predictions made by a deep learning model can aide clinicians identify when AI systems fail is essential for gaining clinicians' trust in the technology. In this work, Bayesian Deep Learning classifier has been trained using transfer learning method on pancreatic cancer grade from histopathology images to estimate model uncertainty. The evaluation shows that the Bayesian DenseNet-201 model trained by H&E stained image dataset achieved the best performance compared to the other models and dataset. Our experiment has shown a strong correlation between model uncertainty and error in prediction. We have also shown how to leverage estimated uncertainty in prediction as rejection threshold in classifying images by user-defined hyperparameters for a given cost of misclassification and rejection cost to control the accuracy-acceptance rate of the model. The estimated

uncertainty in deep learning yields more reliable predictions; protects against model limitations from possible scenarios in training dataset such as minority class, dismissal, automation bias or out-of-distribution test dataset, which can alert radiologists on false predictions, increasing the acceptance of deep learning into clinical practice in disease detection. With this Bayesian Deep Learning based classification, studies correlating with multi "omics" datasets (pathology with radiological, genomic, and proteomic) and treatment responses could further reveal insights about imaging markers and findings towards improved diagnosis and treatment for pancreatic adenocarcinomas.

References

1. Anaya-Isaza, A., Mera-Jiménez, L., Zequera-Diaz, M.: An overview of deep learning in medical imaging. *Informatics in Medicine Unlocked* **26**, 100723 (2021)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: *International conference on machine learning*. pp. 214–223. PMLR (2017)
3. Borgeaud dit Avocat, S.: Gaussian process classifiers for cnn uncertainty. https://github.com/seb5666/cnn_gaussian_process_uncertainty (2018)
4. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424* (2015)
5. Chow, C.: On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory* **16**(1), 41–46 (1970)
6. Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., Socher, R.: Deep learning-enabled medical computer vision. *NPJ digital medicine* **4**(1), 1–9 (2021)
7. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J.: A guide to deep learning in healthcare. *Nature medicine* **25**(1), 24–29 (2019)
8. Feinman, R., Curtin, R.R., Shintre, S., Gardner, A.B.: Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410* (2017)
9. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *33rd International Conference on Machine Learning, ICML 2016*, 3 pp. 1651–1660 (2016)
10. Gal, Y.: Uncertainty in deep learning. PhD thesis, University of Cambridge (2016)
11. Ghoshal, B., Tucker, A.: On calibrated model uncertainty in deep learning. In: *The European Conference on Machine Learning (ECML PKDD 2020) Workshop on Uncertainty in Machine Learning* (2020)
12. Ghoshal, B., Ghoshal, B., Swift, S., Tucker, A.: Uncertainty estimation in sars-cov-2 b-cell epitope prediction for vaccine development. In: *International Conference on Artificial Intelligence in Medicine*. pp. 361–366. Springer (2021)
13. Ghoshal, B., Hikmet, F., Pineau, C., Tucker, A., Lindskog, C.: Deephistoclass: A novel strategy for confident classification of immunohistochemistry images using deep learning. *Molecular & Cellular Proteomics* **20** (2021)
14. Ghoshal, B., Swift, S., Tucker, A.: Bayesian deep active learning for medical image analysis. In: *International Conference on Artificial Intelligence in Medicine*. pp. 36–42. Springer (2021)
15. Ghoshal, B., Tucker, A.: Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. *arXiv preprint arXiv:2003.10769* (2020)

16. Ghoshal, B., Tucker, A.: On cost-sensitive calibrated uncertainty in deep learning: An application on covid-19 detection. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). pp. 503–509. IEEE (2021)
17. Ghoshal, B., Tucker, A., Sanghera, B., Lup Wong, W.: Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. *Computational Intelligence* **37**(2), 701–734 (2021)
18. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press Cambridge **1** (2016)
19. Hidalgo, M.: Pancreatic cancer. *New England Journal of Medicine* **362**(17), 1605–1617 (2010)
20. Houlisby, N., Huszár, F., Ghahramani, Z., Lengyel, M.: Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745 (2011)
21. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems* pp. 5580–5590 (2017)
22. Kenner, B., Chari, S.T., Kelsen, D., Klimstra, D.S., Pandol, S.J., Rosenthal, M., Rustgi, A.K., Taylor, J.A., Yala, A., Abul-Husn, N., et al.: Artificial intelligence and early detection of pancreatic cancer: 2020 summative review. *Pancreas* **50**(3), 251 (2021)
23. Kwon, Y., Won, J.H., Kim, B.J., Paik, M.C.: Uncertainty quantification using bayesian neural networks in classification: Application to ischemic stroke lesion segmentation. *Medical Imaging with Deep Learning Conference* (2018)
24. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* **30** (2017)
25. Laves, M.H., Ihler, S., Ortmaier, T., Kahrs, L.A.: Quantifying the uncertainty of deep learning-based computer-aided diagnosis for patient safety. *Current Directions in Biomedical Engineering* **5**(1), 223–226 (2019)
26. Lebig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports* **7**(1), 1–14 (2017)
27. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. *Medical image analysis* **42**, 60–88 (2017)
28. MacKay, D.J.: A practical bayesian framework for backpropagation networks. *Neural computation* **4**(3):448–472 (1992)
29. Naito, Y., Tsuneki, M., Fukushima, N., Koga, Y., Higashi, M., Notohara, K., Aishima, S., Ohike, N., Tajiri, T., Yamaguchi, H., et al.: A deep learning model to detect pancreatic ductal adenocarcinoma on endoscopic ultrasound-guided fine-needle biopsy. *Scientific reports* **11**(1), 1–8 (2021)
30. Neal, R.M.: Bayesian learning via stochastic dynamics. In *Advances in neural information processing systems* pp. 475–482 (1993)
31. Park, W., Chawla, A., O'Reilly, E.M.: Pancreatic cancer: a review. *Jama* **326**(9), 851–862 (2021)
32. Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J.: Ai in health and medicine. *Nature Medicine* pp. 1–8 (2022)
33. Sehmi, M.N.M., Fauzi, M.F.A., Ahmad, W.S.H.M.W., Chan, E.W.L.: Pancreatic cancer grading in pathological images using deep learning convolutional neural networks. *F1000Research* **10**(1057), 1057 (2021)