When the Machine Does Not Know

Measuring Uncertainty in Deep Learning Models of Medical Images

Biraja Prasad Ghoshal

Department of Computer Science
Brunel University, London, UK

This dissertation is submitted for the degree of
Doctor of Philosophy

September, 2022

I would like to dedicate this thesis to my
loving family who are passionate about
research and still look at new discoveries
with the eyes of a child.

All the images and results presented in this work were produced by the author unless specified differently. They can be reproduced using the code in the thesis repository and the corresponding variational inference framework located at github.com/birajaghoshal.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr Allan Tucker, for his guidance over the years and extend my deepest gratitude. As a great advisor, mentor, and friend, Dr Tucker gave me invaluable guidance and inspiration in researching, writing a paper, making a poster, doing a presentation, and eventually becoming a professional researcher. In particular, I appreciate the freedom he has given me to shape my research direction and his support to structure this endeavour.

I want to thank Dr Stephen Swift and all Intelligent Data Analysis (IDA) Group members, Brunel University, London, for their valuable feedback on my thesis. I would also like to thank the members of the The Human Protein Atlas, Department of Immunology, Genetics and Pathology, Clinical and experimental pathology Unit at Uppsala University, Sweden, for their direct contributions to my work and interesting discussions. In particular, I express my gratitude to Dr Cecilia Lindskog for providing support, despite being very busy. She has shown me how to guide a great research group with dedication and kindness effectively.

I thank my family, particularly my brothers Guru Prasad and Sarada Prasad, for their support and encouragement of my educational pursuits throughout my life. I cannot finish saying how grateful I am to my wife Barnali for her empathetic understanding and my son Bhargab's constant support. They have always supported and encouraged me to do my best in all matters of life.

Abstract

Recently, Deep learning (DL), which involves powerful black box predictors, has outperformed human experts in several medical diagnostic problems. However, these methods focus exclusively on improving the accuracy of point predictions without assessing their outputs' quality and ignore the asymmetric cost involved in different types of misclassification errors. Neural networks also do not deliver confidence in predictions and suffer from over and under confidence, i.e. are not well calibrated. Knowing how much confidence there is in a prediction is essential for gaining clinicians' trust in the technology.

Calibrated uncertainty quantification is a challenging problem as no ground truth is available. To address this, we make two observations: (i) cost-sensitive deep neural networks with Dropweights models better quantify calibrated predictive uncertainty, and (ii) estimated uncertainty with point predictions in Deep Ensembles Bayesian Neural Networks with DropWeights can lead to a more informed decision and improve prediction quality.

This dissertation focuses on quantifying uncertainty using concepts from cost-sensitive neural networks, calibration of confidence, and Dropweights ensemble method. First, we show how to improve predictive uncertainty by deep ensembles of neural networks with Dropweights learning an approximate distribution over its weights in medical image segmentation and its application in active learning. Second, we use the Jackknife resampling technique to correct bias in quantified uncertainty in image classification and propose metrics to measure uncertainty performance. The third part of the thesis is motivated by the discrepancy between the model predictive error and the objective in quantified uncertainty when costs for misclassification errors or unbalanced datasets are asymmetric. We develop cost-sensitive modifications of the neural networks in disease detection and propose metrics to measure the quality of quantified uncertainty. Finally, we leverage an adaptive binning strategy to measure uncertainty calibration error that directly corresponds to estimated uncertainty performance and address problematic evaluation methods.

We evaluate the effectiveness of the tools on nuclei images segmentation, multi-class Brain MRI image classification, multi-level cell type-specific protein expression prediction in ImmunoHistoChemistry (IHC) images and cost-sensitive classification for Covid-19 detection from X-Rays and CT image dataset. Our approach is thoroughly validated by measuring the quality of uncertainty. It produces an equally good or better result and paves the way for the future that addresses the practical problems at the intersection of deep learning and Bayesian decision theory.

In conclusion, our study highlights the opportunities and challenges of the application of estimated uncertainty in deep learning models of medical images, representing the confidence

of the model's prediction, and the uncertainty quality metrics show a significant improvement when using Deep Ensembles Bayesian Neural Networks with DropWeights.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

> “Wisdom is Knowing What We Don’t Know." Socrates

Recent advances in deep learning have achieved a remarkable performance in medical image analysis by improving the diagnostic performance in medical imaging, enhancing the early detection of various diseases, improving a deeper understanding of physiology and pathology, and advancing the field of Computational Radiology. Specifically advanced modalities of digital medical images include ultrasound (US), X-ray, computed tomography (CT) and magnetic resonance imaging (MRI) scans, positron emission tomography (PET) scans, mammography, retinal photography, histology slides, and dermoscopy images. Some of these modalities are organ-specific (retinal photography, dermoscopy). In contrast, others examine multiple organs (such as X-ray, CT, MRI) of the human body or a part of the human body in a non-invasive manner for various tasks such as image detection, classification and segmentation, and registration. As a result, medical images have several traits such as significant variations in pathology, imbalanced long-tail multi-modal distribution of disease patterns, sparse and noisy labels, varied amount of generated data and a high pixel resolution (Shen et al., 2017; Zhou et al., 2017). Figure 1. shows some examples of medical imaging ’ologies’ that have benefited from deep learning (Shen et al., 2017).

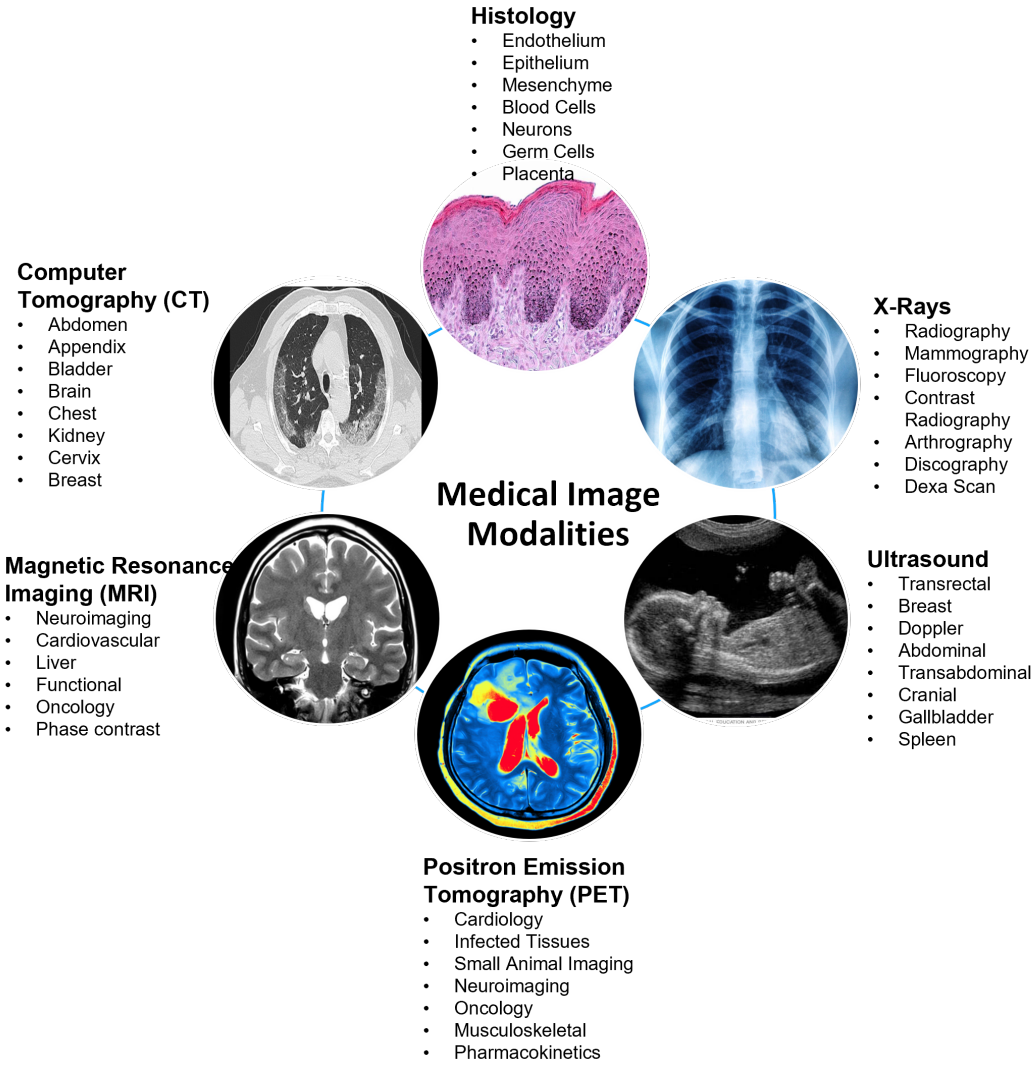


Fig. 1.1 Clinical Applications for Deep Learning in Medical Imaging.

As a routine, clinicians make crucial decisions to determine the diagnosis of patients. They use their personal experience as "prior information" along with data gathered through different approaches, such as the medical interview, physical examination or diagnostic test as the likelihood of a diagnosis. Clinicians express their assumptions about various possible diagnoses when making a decision. Ambiguity in data, the asymmetric cost of misdiagnoses, biases and the black-box nature of deep learning models lead to the lack of model interpretability, reliability and explainability, which contributes to the uncertainty in the final diagnosis. Estimating uncertainty in deep learning models' predictions improves predictive performance and model interpretability for clinical applications in computer-based medical image analysis.

1.1 Motivation

Deep Learning methods, which involves powerful black box predictors, focus exclusively on improving the accuracy of point predictions without assessing the quality of their outputs and tend to produce overconfident predictions. In deep learning, two distinct types of predictive uncertainties exist: aleatoric uncertainty and epistemic uncertainty (Der Kiureghian and Ditlevsen, 2009). Aleatoric uncertainty arises due to inherent randomness in the data. Consequently, in regions that are well represented by the training data, a model's aleatoric uncertainty should accurately estimate capturing the stochastic pattern in the data. On the other hand, epistemic uncertainty arises due to a lack of knowledge about the data. Hence in regions unexplored by the training data, the model's epistemic uncertainty should increase to capture the model's lack of confidence in the predictions.

However, quantifying uncertainty in model prediction is crucial for interpretability and explainability of the model in computer-based clinical applications (Ghahramani, 2015; Krzywinski and Altman, 2013). In fact, a mismatch between a model's confidence and its accuracy due to the model architecture, normalisation, and regularisation techniques is a key reason why a standard neural network training is miscalibrated. The consequences of an overconfident incorrect prediction can be fatal. Hence, it is essential to consider the uncertainty in deep learning where serious decisions are being made upon the model's predictions. In particular, the following two intuitive desiderata:

- A prediction with a low uncertainty which is likely to be accurate
- A prediction with high uncertainty is likely to be incorrect

Quantifying reliable uncertainty in deep neural networks is a challenging and unsolved problem. Bayesian Neural Networks (BNN) provide a natural and principled way of modelling uncertainty, robust to over-fitting (i.e. regularisation). However, exact Bayesian inference on the weights of a deep neural network is computationally intractable.

There are variety of approximations that have been developed including Deep Ensembles, Markov Chain Monte Carlo (MCMC) techniques, variational inference-based Monte Carlo Dropout (MCDO), and Monte Carlo Batch Normalisation (MCBN) uncertainty estimation methods (Blundell et al., 2015; Gal, 2016; Lakshminarayanan et al., 2017; MacKay, 1992a; Neal, 1993; Teye et al., 2018). However, all these methods capture variance of network parameters and the amount of noise in the input data is constant. On the other hand in Monte Carlo Batch Normalisation (MCBN), by increasing the mini-batch size, BN layers become deterministic and so unable to capture model uncertainty.

Current approaches to approximate Bayesian Deep learning assumes an equal cost for classification errors. Therefore, deep learning models are poorly calibrated at quantifying

predictive uncertainty: i.e. the mismatch between a model's uncertainty and its error, for the applications sensitive to misclassification costs. In practical applications such as medical imaging, safety is critical and prediction problems are asymmetric as different types of misclassification errors incur different costs or significant losses. Because of this, overconfident incorrect predictions may result in the loss of life in some circumstances. Although higher accuracy is a well recognised benefit for deep learning approaches, the enormous potential for improvements in calibrated uncertainty for cost-sensitive applications is a primarily overlooked advantage. Knowing how much confidence there is in a prediction is essential for gaining clinicians' trust in the technology.

The main goal of this thesis is to quantify uncertainty to make medical imaging with deep learning more robust and more accurate, which leads us to our research questions:

1. How to measure model uncertainty in deep learning?
2. Is approximate Bayesian neural networks a good principle for deep learning?
3. Are there any alternatives of quantifying uncertainty that align better with our goal?
4. Can we develop practical inference algorithms to measure model uncertainty in cost-calibrated situations?

Therefore, the objective of this thesis is to use Deep Ensembles of Bayesian deep learning with variational inference that provides calibrated uncertainty quantification in deep neural networks and addresses overconfident predictions due to asymmetric costs involved in misclassification errors. Furthermore, most of the baseline performance measures of the deep learning models are uncertainty - independent metrics, which are not reliable for real-world applications. Therefore, this thesis aims to demonstrate performance metrics accounting quality of estimated uncertainty in applications.

1.2 Contributions

This thesis explores ideas using concepts from the weight uncertainty in neural networks, calibration of confidence, cost-sensitive applications and Bayesian decision theory. To summarise, the contributions of this thesis are as follows:

- We have developed a technique called "DropWeights", which randomly drops connections; incoming or outgoing weights are set to zeros, including drop neurones. DropWeights can be considered as the combination of generalised version of Dropout

and DropConnects, and this comprises of the method used for regularising deep neural networks. We proposed "Bayesian Deep Ensembles of DropWeights" to quantify uncertainty and metrics to evaluate the estimated uncertainty performance as well as the quality of uncertainty of the Bayesian Deep Learning models for the classification and semantic segmentation.

- Various methods used in the literature to estimate the uncertainty of neural network predictions: prediction variance, Leibig's uncertainty, Feinman's predictive uncertainty, and stochastic sampling-based measure: variance of MC samples, predictive entropy, and mutual information. Estimation of entropy as a form of quantified uncertainty from the finite set of data suffers from a severe downward bias when the data is under-sampled. Even small biases can result in significant inaccuracies when estimating epistemic entropy. We have leveraged the Jackknife resampling technique to quantify bias-corrected uncertainty.

- The goal of uncertainty estimation is to characterise predictive uncertainty properly. We have decomposed predictive uncertainty into aleatoric uncertainty and epistemic uncertainty, which can provide additional information. We have also shown that there is a strong correlation between classification accuracy and estimated uncertainty in predictions.

- We investigated neural networks with Dropweights to calibrate model uncertainty, revising backpropagation learning classification procedures that attempt to minimise the cost of misclassified samples, rather than the number of misclassified samples to represent the model error. We encoded asymmetries due to the different types of misclassification errors or probability of occurrence of different classes in the form of a utility function. We obtained calibrated predictive uncertainty for applications with an asymmetric cost, by maximising the utility function (i.e. minimising asymmetric costs) in backpropagation learning procedure.

- Calibrated uncertainties provide an additional confidence to identify false predictions whilst minimising the asymmetric misclassification costs, yielding more reliable results and further improves overall model accuracy. We leveraged the adaptive binning strategy to measure uncertainty calibration error which directly corresponds to estimated uncertainty performance to address non-uniformity issues with fixed binning calibration metrics.

- We further evaluated uncertainty quality from Bayesian neural networks with Dropweights using two metrics: Predictive Log-Likelihood (PLL) and Brier Score (BS),

which all produce an equally good or better result than standard Bayesian neural networks.

This thesis makes three observations, which can be used to address the weaknesses of deep learning, which in turn will improve model interpretability and explainability in medical imaging ranging from nuclei image segmentation, Brain MR image classification, cell type-specific protein expression prediction in immunohistochemistry (IHC) images and Covid-19 detection from X-Rays and CT images:

1. The prediction uncertainty is correlated with prediction accuracy. Cost-sensitive approximate variational inference better quantifies calibrated predictive uncertainty

2. Deep learning models tend to be overconfident about their predictions. Estimated uncertainty with point predictions can lead to a more informed decision and improve the quality of prediction

3. The standard conventional evaluation method such as AUROC and AUPR is either misleading or meaningless when the predictive models are different (Lobo et al., 2008). Performance metrics accounting uncertainty information avoids pathologies of existing metrics to provide reliable and confident results.

This research question is becoming increasingly important as Bayesian Neural Networks (BNN) are becoming more prevalent in medical image diagnosis (Araújo et al., 2020; Awate et al., 2019; Ayhan et al., 2020; Baumgartner et al., 2019; Esteva et al., 2017; Ghesu et al., 2019; Irvin et al., 2019; Jungo and Reyes, 2019; Leibig et al., 2017; Litjens et al., 2017; Nair et al., 2020; Wickstrøm et al., 2020).

1.3 Co-Authored Papers

The following publications have resulted from the research presented in this thesis:

1. **Estimating uncertainty in deep learning for reporting confidence to clinicians when segmenting nuclei image data**; B Ghoshal, A Tucker, B Sanghera, WL Wong; IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) 2019 Jun 5 (pp. 318-324) https://doi.org/10.1109/CBMS.2019.00072]

2. **Hyperspherical Weight Uncertainty in Neural Networks**; B Ghoshal, A Tucker; International Symposium on Intelligent Data Analysis (IDA) 2021 April 26–28, 2021, LNCS 12695 (Pages 3-11). Springer.

3. **Estimating Uncertainty in Deep Learning for Reporting Confidence to Clinicians in Medical Image Segmentation and Diseases Detection**; B Ghoshal, A Tucker, B Sanghera, W Lup Wong; Computational Intelligence; Wiley Online Library; 22 October 2020;

4. **Bayesian Deep Active Learning for Medical Image Analysis**; B Ghoshal, S Swift, A Tucker; 19th International Conference on Artificial Intelligence in Medicine in Europe (AIME 2021), 2021 Jun 15, LNAI, volume 19 (Pages 36-42), Springer.

5. **Uncertainty Estimation in SARS-CoV-2 B-cell Epitope Prediction for Vaccine Development**; B Ghoshal, B Ghoshal, S Swift, A Tucker; 19th International Conference on Artificial Intelligence in Medicine in Europe (AIME 2021), 2021 Jun 15, LNAI, volume 12721 (Pages 361-366), Springer.

6. **Estimating Uncertainty in Deep Learning for Reporting Confidence: An Application on Cell Type Prediction in Testes Based on Proteomics**; B Ghoshal, C Lindskog, A Tucker; International Symposium on Intelligent Data Analysis 2020 Apr 27 (pp. 223-234). Springer, Cham.

7. **DeepHistoClass: A novel strategy for confident classification of immunohistochemistry images using Deep Learning**; Mr Biraja Ghoshal, Allan Tucker, Dr Charles Pineau, Mr Feria Hikmet Norradin, Dr. Cecilia Lindskog; Journal of Molecular and Cellular Proteomic published by American Society for Biochemistry and Molecular Biology; Page: 100140; ISSN: 1535-9476; 2021;

8. **Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection**; B Ghoshal, A Tucker; arXiv preprint arXiv:2003.10769 (372+ citations)

9. **On Calibrated Model Uncertainty in Deep Learning**; B Ghoshal, A Tucker; The European Conference on Machine Learning (ECML PKDD 2020).

10. **On Cost-Sensitive Calibrated Uncertainty in Deep Learning: An application on COVID-19 detection**; B Ghoshal, A Tucker; IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS) 2021 Jun 7-9, ISBN: 978-1-6654-4121-6 (Pages 509-515)

11. **Leveraging Uncertainty in Deep Learning for Pancreatic Adenocarcinoma Grading**; B Ghoshal, B Ghoshal, A Tucker; Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022. Lecture Notes in Computer Science, vol 13413. Springer, Cham. https://doi.org/10.1007/978-3-031-12053-4-42

Publication 1 and 2 results from foundation work conducted for this thesis, presented in chapter 3. Publication 3, 4, and 5 results from work presented in chapter 4 and publication 6 is an early output from work further developed in publication 7 presented in chapter 5. Publication 8, 9 and 10 results from work presented in chapter 6.

1.4 Thesis Structure

This thesis presents our work around measuring uncertainty in deep learning models, practical issues in quality of estimated uncertainty, and applications in medical images. An overview of this dissertation is shown in Figure 1.2.

The main focus of Chapter 2 is to introduce the Bayesian neural networks, highlighting methods for approximate inference, focused on the literature associated with Uncertainty Quantification in Medical Image Analysis on both the background and recent approaches. Chapter 3 is focused on developing tools which measure predictive uncertainty by Dropweights based Bayesian neural networks learning an approximate distribution over its weights and assess empirically.

The chapters thereafter follow different research questions on the theme of the application of estimated uncertainty in deep learning in the context of medical image analysis to improve accuracy of automated predictions and identification of manual errors, while minimising the total misclassification cost.

Chapter 4 provides a Bayesian perspective for Neural Networks applications in image segmentation and its application in active learning.

In Chapter 5, we use the Jackknife resampling technique to correct bias in quantified uncertainty in image classification and propose metrics to quantify uncertainty estimates in both multi-class classification and multi-level classification medical image analysis. Our experimental results show that the MC-Dropweights visibly improve performance to estimate uncertainty compared to current approaches in image classification.

Chapter 6 provides a cost-sensitive classification in Bayesian neural networks, which means cost-sensitive calibrated predictive uncertainty in medical imaging can be estimated whilst minimising the asymmetric cost in misclassification with improved accuracy and proposed adaptive binning strategy, based revised metrics to mitigate them.

In Chapter 7, we make overall conclusions, discuss the application of this technology and suggest directions for future research.

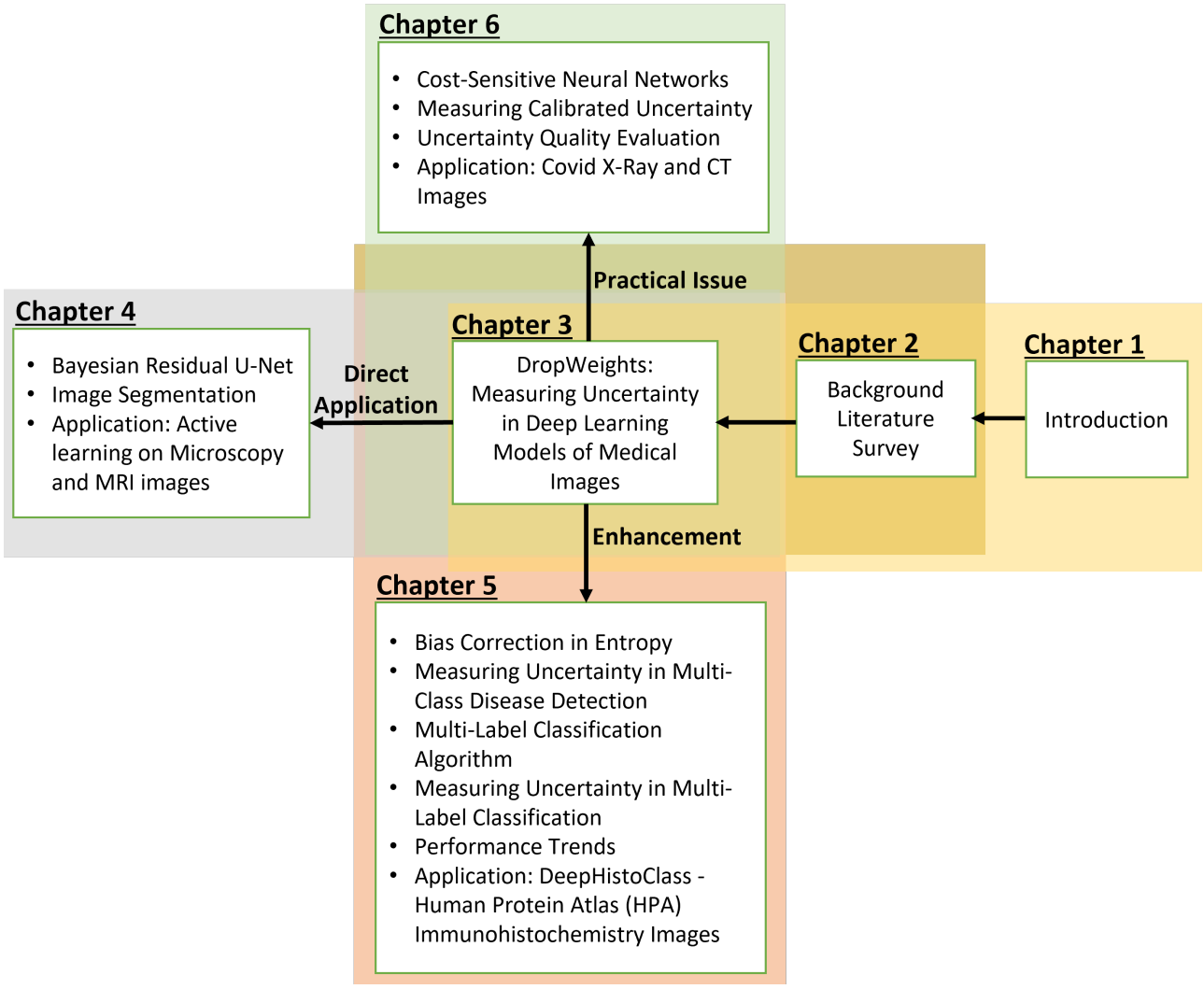


Fig. 1.2 Overview of this dissertation

1.5 Datasets

Deep Learning (DL) is one of the most exciting subfields within the Artificial Intelligence (AI) and Machine Learning (ML) domain with different Computer Vision (CV) tasks, including semantic segmentation, image classification, object detection, and cost-sensitive prediction. In addition, the nature of the images affects deep learning models learn from them. The below table shows various deep learning tasks and image datasets used in this thesis.

Deep Learning Task	Method	Dataset used	Number of Images
Measured epistemic and aleatoric uncertainty in Segmentation (chapter 4)	Semantic segmentation is one of the essential tasks for medical image understanding. Semantic segmentation is classifying each pixel belonging to a particular label.	1. Microscopy images of segmented nuclei images 2. IHC image of Cytokeratin-Supervised Epithelial Cells in Breast Cancers 3. Diabetic retinopathy on digital fundus images	1. Training set with 670 images and the test set with 4000 Segmented nuclei images accordingly. 2. Total 13344 images included from 152 patient samples stained with fluorochromogenic cytokeratin-Ki-67 double staining or sequential hematoxylin-IHC. Active learning experiment: Initial Training Size: 100; Pool Size: 5000; Batch Size: 10, AL Iterations: 50 3. Digital Retinal Images for Vessel Extraction dataset (DRIVE) consists of 40 images, 20 for training purpose and 20 for testing.

continued . . .

... continued

Deep Learning Task	Method	Dataset used	Number of Images
Measured epistemic uncertainty in Multi-Class classification (chapter 5)	A classification problem where the task is to categorise an image into three or more classes.	1. Magnetic Resonance Imaging (MRI) of brain tumours	This dataset contains 96115 MRI images, containing 21307 Astrocytoma, 17983 Glioblastoma, 12460 Oligodendroglioma, 13677 Unidentified and 31244 Healthy brain.
Measured epistemic uncertainty in Multi-Label classification (chapter 5)	Multi-label image classification is the task of predicting a set of mutually non-exclusive classes or "labels" corresponding to objects, attributes or other entities present in an image.	1. IHC (Immunohistochemical) staining images of cell type-specific protein expression of 8 different cell types in human testis.	A total of 7848 IHC stained images of human testis, corresponding to 3046 different antibody stainings and 2794 unique proteins, were divided into three different sets: a training set (5411 images), a validation set (1063 images), and a test set (1374 images).

continued ...

. . . continued

Deep Learning Task	Method	Dataset used	Number of Images
Measured epistemic uncertainty in Cost-Sensitive classification (chapter 6)	Specialised learning that considers the cost of misclassification errors. The goal is to minimise the total cost.	1. Chest X-Ray Images radiography images across four classes (Normal, Bacterial Pneumonia, non-COVID-19 Viral Pneumonia, and COVID-19)	Total of 5941 Posterior-Anterior (PA) chest radiography images across 4 classes (Normal: 1583, Bacterial Pneumonia: 2786, non-COVID-19 Viral Pneumonia: 1504, and COVID-19: 68).

Chapter 2

Uncertainty Quantification in Medical Image Analysis

"Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge." Ronald Fisher (1890 - 1962).

Deep learning is ubiquitous in the field of computer vision. This chapter reviews the current work and background related to measuring uncertainty and its applications in medical image analysis. We will also review cost-sensitive neural networks and the existing calibration methods related to Bayesian deep neural networks. For further information on uncertainty estimation in Bayesian settings, we refer the reader to (Gal, 2016).

2.1 Background

Our ability to learn from observations is the primary source of knowledge about the world. Deep learning provides a robust framework, have become popular in recent years due to their outstanding performances in complex prediction tasks (Krizhevsky et al., 2012; LeCun et al., 2015). Deep learning (Goodfellow et al., 2016) parameterised models with compositions of functions trained using backpropagation and stochastic gradients decent, usually learned by maximising the log-likelihood. Deep learning is based on the philosophy of connectionism: deep neural networks aim to optimise a set of algebraic functions in order to perform the prediction with the highest degree of accuracy using multiple layers of neurons and the basic unit of the model (Schmidhuber, 2015). As a machine learning tool, deep neural networks effectively understand high dimensional data, such as images.

However, despite these successes, the main drawback of neural networks lies in their lack of interpretability. They are often deemed as "black boxes" (Benítez et al., 1997; Duch, 2003; Lundberg and Lee, 2017; Shrikumar et al., 2017). Despite their ability to outperform simpler models for a variety of tasks and domain applications, point prediction score (i.e. the accuracy of the prediction) is not sufficient (Ghahramani, 2015). Deep learning has been highly successful in a range of applications. However, for validation and interpretability, we need the predictions made by the model and how confident it is while making those predictions. This is very important in medical imaging for clinicians to accept it.

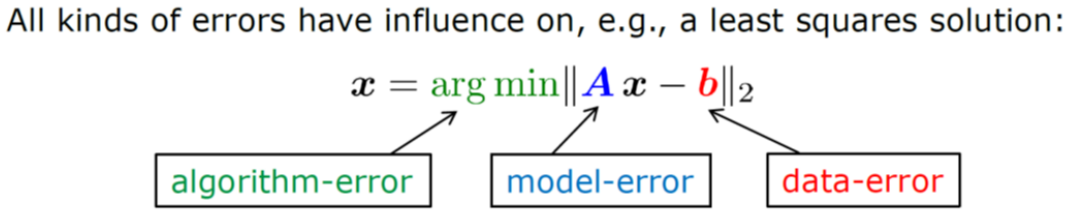


Fig. 2.1 Errors in deep learning.

Uncertainty is the most common and unavoidable feature of deep learning tasks. Understanding what a model does not know is a critical part of many machine learning systems (Ghahramani, 2015) . Unfortunately, today's deep learning algorithms are usually unable to explain "why is the model made this prediction" and "why is the model uncertain about a prediction?". It is also equally important to understand "what deep learning models do not know". Therefore, it is not sufficient to depend on deep learning models' regression or classification score alone. Estimating uncertainty in deep neural networks is a challenging and yet unsolved problem. The Bayesian framework provides a natural and principled way of modelling uncertainty via probability density over outcomes, resistant to overfitting. In order to address this problem, the deep neural networks need to provide uncertainty estimation as an additional insight to point prediction to improve the reliability, trustworthiness and safety of these systems in the decision-making process.

2.2 Medical Imaging Modalities

Medical imaging is a valuable tool for clinicians and radiologists aim to assist physicians in clinical examination in the diagnosis, pathology of the disease state, estimation of treatment response and appropriate treatment decisions. There are several different types of medical image modalities. Each medical image has its organ and properties. Medical imaging incorporates various disciplines, including radiology, nuclear medicine, radiation physics and

tomography. The medical imaging modalities are based on the method in which images are generated, including light, electrons, lasers, X-rays, radionuclides, ultrasound and nuclear magnetic resonance. Recent advances in medical image modalities produce 2-D and 3-D digital images both anatomical (X-ray radiography, Computed Tomography (CT), Magnetic Resonance Imaging (MRI)) and physiological (Ultrasonography (U/S), Immunohistochemistry (IHC)), or functional images (Nuclear Medicine - PET and SPECT), ranging from molecules and cells to organ systems with acceptable degrees of contrast and resolution. In histopathology, Whole-slide images (WSI) are very large in image size, and usually, every WSI has high spatial resolution used in digital pathology. A broad scope of abnormalities has been extensively explored in different research areas, such as detecting cancerous cells in different medical structures, identifying dead tissues in different organs, and detecting brain abnormalities in subjects with brain disorders. Unlike general images, medical images have various aspects such as shape, posture complexities, texture, colour and visual features. Since all the work presented in this dissertation mainly deals with medical images, figure 2.2 shows the different types of medical imaging modalities.

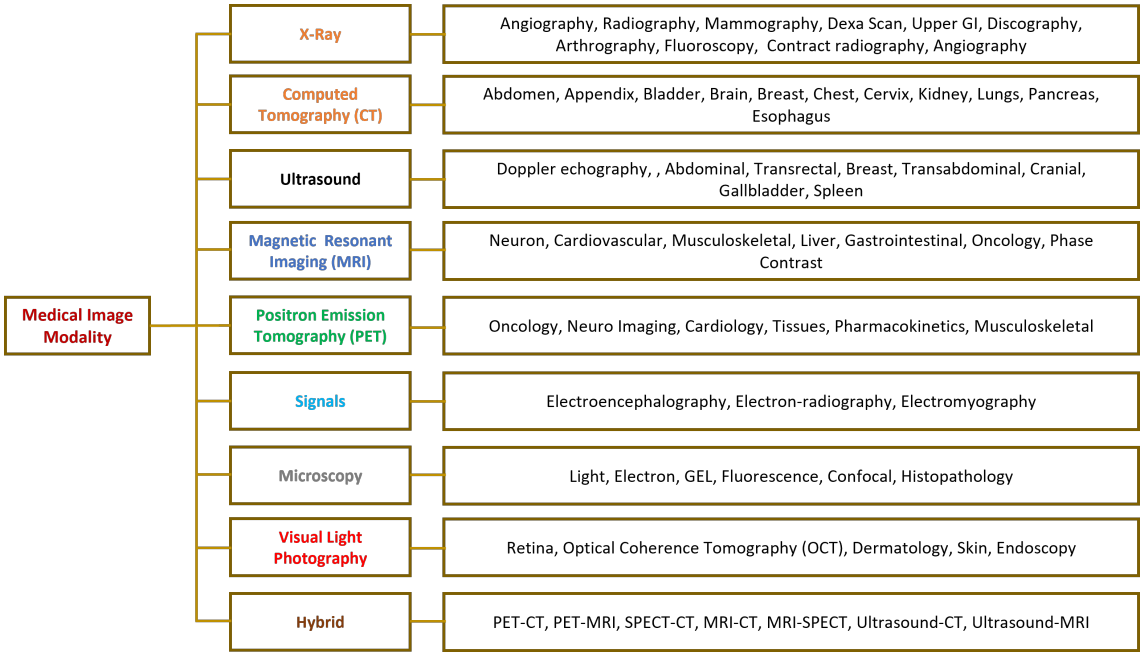


Fig. 2.2 Classification of medical imaging modalities

2.3 Artificial Neural Networks

Inspired by the early research in neuroscience (Hebb, 1949; Rosenblatt, 1958), the Artificial Neural Network (ANN) was developed to process information similar to how the brain

process information. A perceptron is conceived as a mathematical model of how the neurons function in our brain. Learning in biological nervous systems involves adjustments to the synaptic connections, which is similar to weight updates in a Neural Network. ANN models have been extensively applied to a wide range of machine learning tasks such as computer vision (Krizhevsky et al., 2012; Rowley et al., 1998), reinforcement learning (Mnih et al., 2013), speech synthesis (Oord et al., 2016), speech and text recognition (Bengio et al., 2003), chemical and molecular modelling (Wei et al., 2016), and many others.

We review a single hidden layer fully connected network, as it is a prerequisite for future discussions. Our input x to the network is a vector with Q elements, and we transform it with a linear map to a K elements vector. The weight matrix W_1 (i.e. a linear map) and the bias vector b1 (i.e. a translation) operate the affine linear transformation. A non-linear differentiable activation function $\sigma(\dot{})$ is then applied to $xW_1 + b$. Accordingly, the network output y^* is obtained by means of a second linear transformation W_2 that connects the hidden layer to the model output,

$$y^* = \sigma(xW_1 + b)W_2 \tag{2.1}$$

where y^* is a vector of C elements. Therefore, W_1 is a QxK matrix, W_2 is a KxC matrix, and b is a K dimensional vector. W_1, W_2, and b are the learnable parameters in our network.

We can easily generalise to L hidden layers network by treating each layer's output $f_i^{W_i}(\dot{})$ as a non-linear function by computing

$$\phi(x) = f(\sum_{i=1}^{d} W_i x_i + b) \tag{2.2}$$

where $f(z) = \frac{1}{1+\exp(-z)}$ is the component-wise logistic function (e.g.: sigmoid, softmax, tanh, or ReLU). The output of the network is:

$$y^* = f_L^{W_L}(\ldots f_1^{W_1}(x)) \tag{2.3}$$

where each network's weight matrix W_i has dimensions of $K_{i-1}xK_i$ and the bias b_i has dimension of K_i for each layer $i = 1,\ldots,L$. In classification, the network learns a categorical distribution over the classes. The model output is:

$$y^* = Cat(y|f^W(x)) \tag{2.4}$$

where y^* is a categorical distribution over C classes.

To perform multiclass classification, an extra softmax layer, parameterized by $\theta_l = \{\omega_l, b_l\}$, is placed after the L-th hidden layer. There are k neurons in the softmax layer, where the j-th

neuron comes with weights $W_j^{(l)}$ and bias $b_j^{(l)}$ and is responsible for estimating the probability of class j given x:

$$P(y=j|x) = \frac{\exp(\phi(x)^T W_j^{(l)} b_j^{(l)})}{\sum_{j=1}^{d} \exp(\phi(x)^T W_j^{(d)} b_j^{(d)})} \tag{2.5}$$

Traditionally, the parameters $(\{w_i\}_{i=1}^{L}, w_l)$ of the network are optimized by the back-propagation algorithm, which is essentially stochastic gradient descent (SGD), with respect to the negative loglikelihood loss function over the training set S:

$$L_{NLL}(S) = \sum_{i=1}^{N} -ln(P(y=y_n|x_n)) \tag{2.6}$$

One of the challenges when using gradient descent is setting the appropriate learning rate. There are variants of SGD, e.g., RMSProp, AdaDelta and Adam, which adaptively change the learning rate for each individual weight. One of the effective normalisation operators is Batch Normalization to keep the output of each layer in a certain range of values and ease the parameter optimisation process (Goodfellow et al., 2016).

The three most basic families are feedforward neural networks (often referred to as multi-layer perceptrons (MLP)), recurrent neural networks (RNN) (Hochreiter and Schmidhuber, 1997; Sundermeyer et al., 2012) for the temporal structure of the data and convolutional neural networks (CNN) (Krizhevsky et al., 2012) represents spatial structure.

2.3.1 Convolutional neural networks

Convolutional neural networks (CNNs) have become standard in many deep learning applications, especially in image processing or vision tasks (Goodfellow et al., 2016). A convolutional neural network is a type of feedforward neural network, typically consisting of convolutional layers, pooling layers and fully connected layers:

- **Convolutional layers** are composed of several convolution kernels, each computing a different feature map. The output feature maps are obtained by convolving the input with the convolution kernel and applying an element-wise nonlinearity. Mathematically, the feature value $z_{i,j,k}^l$ at location (i,j) of the k^{th} feature map in the l^{th} layer is computed as:

$$z_{i,j,k}^l = {w_k^l}^T x_{i,j}^l + b_k^l \tag{2.7}$$

  where w_k^l and b_k^l are the weight and bias vectors for the k^{th} convolution kernel in the l^{th} layer and $x_{i,j}^l$ is the input patch centered around (i,j) in the l^{th} layer. The output

value is computed by apply a nonlinearity $a(\cdot)$ point-wise:

$$x_{i,j,k}^{(l+1)} = a(z_{i,j,k}^{l}) \tag{2.8}$$

- **Pooling layers** aim to achieve shift-invariance and reduce the number of parameters in the network by reducing the resolution of the feature maps. The pooling layer operates on each feature map independently. Mathematically, the output of a pooling layer with pooling operation $\text{pool}(\cdot)$ is given by

$$y_{i,j,k}^{l} = \text{pool}(x_{m,n,k}^{l}), \forall (m,n) \in R_{i,j} \tag{2.9}$$

  where $R_{i,j}$ is a local neighbourhood around (i,j). Typically, the pooling operation computes the average or the maximum.

- **Fully connected layers** connect every neuron in the previous layer to every neuron in the current layer. Mathematically, the output of a fully connected layer is given by:

$$x_{i}^{(l+1)} = a\Big(\big(\sum_{j} w_{i,j}^{l} x_{j}^{l} \big) + b_{i}^{l} \Big) \tag{2.10}$$

  where $a(\cdot)$ is a nonlinearity, $w_{i,j}^{l}$ is the weight connecting neuron j in the l^{th} layer to neuron i in layer $(l+1)^{\text{th}}$, and b_{i}^{l} is the bias weight for neuron i.

The learning process for a convolutional neural network is identical to the learning process for standard neural networks. First, a differentiable loss function is computed for the training examples (often done in batches), and the gradients w.r.t. the network weights are computed. Then, the weights are updated in a gradient descent step using these gradients. Typically, more complex update rules that take into account momentum (e.g. Adam optimisation (Goodfellow et al., 2016)) are used as they converge faster.

2.3.2 Neural networks limitations

In supervised learning, given a data set $D = \{(x_n, y_n)\}_{n=1}^{N}$ formed by feature vectors $x_n \in R$ and targets y_n, most networks are trained to learn the optimal set of network parameters $w_{\hat{MLE}} = arg\max_w \prod_{i=1}^{n} p_w(D|x_i) = arg\max_w \sum_{i=1}^{n} \log p_w(D|x_i)$ maximizing the probability of the observed data according to the model, using maximum likelihood estimation (MLE) criterion:

$$\min_{w} \sum_{n=1}^{N} -\log p(y|x,w) \tag{2.11}$$

Where the negative log-likelihood $-\log p(y|x,w)$ results in a cross-entropy loss in classification (or a squared error in the case of regression), the neural network prediction may be real-valued continuous output in regression or categorical for classification problem.

Regularization is usually added through a prior distribution $p(w)$ on the weights to avoid overfitting, and performing maximum a posteriori (MAP) $p(D|w)p(w)$:

$$\min_{w} \sum_{n=1}^{N} -\underbrace{\log p(D|w)}_{\text{Likelihood}} - \underbrace{\log p(w)}_{\text{Prior}} - constant \tag{2.12}$$

It corresponds to the cross-entropy loss in classification. We assume that data points are drawn independently from a Gaussian distribution with an unknown mean but constant variance σ^2.

In regression, maximising a Gaussian log-likelihood w.r.t. the model parameters is equivalent to minimising the mean squared error (MSE loss) using mini-batches, which corresponds to L2 regularisation as weight decay with a unit Gaussian prior resulting in stochastic gradient estimation.

Overall, the MAP solution is computationally efficient but only provide point estimates of the network weights are obtained and not a distribution over parameters. With the success of deep neural networks, understanding if a model is under-confident or falsely over-confident (i.e. its uncertainty estimates are too small) can help to improve reliability in terms of robustness and confidence in the prediction. Bayesian neural networks (BNN) promises improved predictions and address these issues by directly modelling the uncertainty of the network weights.

2.4 Deep Learning in Medical Imaging

Since Convolutional neural networks (CNN) (LeCun et al., 1989, 1998) were first introduced in 1989, many complex and deep CNN models have been extended in several directions, as represented by VGGNet, Inception Net, and ResNet (He et al., 2016a; Simonyan and Zisserman, 2014). The use of skip connections makes a deep network more trainable, as in DenseNet and U-Net (Ronneberger et al., 2015). The availability of big data, improvement in the hardware technology and several inspiring ideas such as the use of parameter optimisation, regularisation, different activation, loss functions, and architectural innovations into different categories based on spatial exploitation, depth, multi-path, width, feature-map exploitation, channel boosting, and attention has accelerated the research in CNNs (Khan et al., 2020). The following Fig. 2.3 shows a brief timeline of CNN architectural developments, starting from 1989 all the way to 2020:

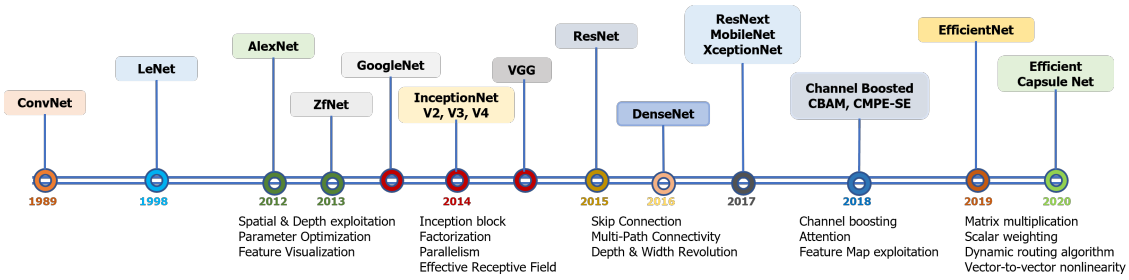


Fig. 2.3 Evolution of CNN architectures

Recently, research into the connectivity of the layers in DNNs has led to new architectures such as RepVGG (Ding et al., 2021). Attempts to adopt these types of models to mobile devices (Howard et al., 2017) as well as the automatic design of neural network architectures in an emerging field called neural architecture search (NAS) (Elsken et al., 2019) have been presented.

Deep learning methods have been widely used in various medical imaging; for example, Rajpurkar et al. (Patel et al., 2019; Rajpurkar et al., 2017) developed (CheXNet) deep learning model to detect fourteen types of chest pneumonia disease using X-ray images. The model was able to detect at the level of radiologists with reduced human efforts. In another study, Gulshan et al. (Gulshan et al., 2016) applied a Deep learning model for diabetic detection using retinal fundus images. Varadarajan et al. (Varadarajan et al., 2020) developed a deep learning model to predict diabetic macular oedema grades from optical coherence tomography images. Similarly, Esteva et al. (Esteva et al., 2017) proposed a CNN image-based model for skin cancer detection and successfully classified the disease. Moeskop et al. (Moeskops et al., 2016), used CNN for segmentation of brain MR images. XU et al. (Xu et al., 2016) used a deep convolutional neural network for segmenting and classifying microscopy cell nuclei in histopathology biopsy images for diagnosis of cancer cells.

In the survey of the literature, Zhou et al., Litjens et al. and Mohammed et al. highlighted both clinical needs and technical challenges in medical imaging and described how emerging trends in deep learning are addressing these issues (Litjens et al., 2017; Mohammed and Al-Ani, 2020; Zhou et al., 2021).

However, deep learning methods have often been described as 'black boxes'. Due to the scarcity of experts clinicians, rough estimate procedures, complex shapes, locations and structures of the medical images makes the analysis difficult even for specialised physicians. So there is a need to express the ambiguity of an image and unreliable predictions in the same way a doctor may express uncertainty and ask for experts' help. Furthermore, where accountability in the decision is important, and there are legal implications, it is often not

enough to have just a good prediction score. This computer-based medical system also has to be able to explain itself in a certain way.

Several strategies such as deconvolution networks (Zeiler and Fergus, 2014), guided backpropagation (Springenberg et al., 2014), deep Taylor composition (Montavon et al., 2017) or prediction to textual representations of the image (i.e. captioning) (Karpathy and Fei-Fei, 2015) have been developed to understand what intermediate layers of convolutional networks.

P-value derived from classical statistics often confuses probability with Bayesian posterior probabilities. Scientists are interested in the conditional probability of parameter values of given data. Bayesian statistics allow probabilistic inferences about the true population mean and other parameters. Recently, researchers have tried to combine Bayesian decision theory with neural networks to quantify uncertainty in deep learning.

Medical imaging datasets are often noisy, incomplete, and prior knowledge may be inconsistent with the measurements. There is a need to quantify uncertainty in deep learning to improve interpretability and make them more reliable and trustworthy for clinicians.

2.5 Uncertainty in Deep Learning

Estimating uncertainty is important in deep learning and explaining what a model does not know is crucial for practitioners in safety-critical application domains such as medical imaging. There are two primary sources of uncertainty in deep learning: epistemic and aleatoric uncertainty (Hüllermeier and Waegeman, 2019).

1. Aleatoric (Data) uncertainty: This uncertainty arises from the natural stochasticity of observations. Aleatoric uncertainty is irreducible even when more data is provided but possibly be reduced with additional features. Aleatoric uncertainty arises mainly from labelling noise (e.g. human disagreement), measurement noise (e.g. imprecise tools), and missing data (e.g. partially observed features, unobserved confounders).

2. Epistemic (model) uncertainty: This refers to the uncertainty of the model and is often due to a lack of training data. This type of uncertainty is "reducible". Theoretically, it vanishes with infinite data (subject to deep learning model identifiability). Epistemic uncertainty can arise in areas with fewer samples for training, underrepresented groups in a facial recognition dataset, new disease classes in the test dataset or the presence of rare words in a language modelling context.

Bayesian Neural Networks provides a natural and principled way of modelling uncertainty. Bayesian deep learning covers everything from inference in Bayesian neural networks

(MacKay, 1992b; Neal, 1996) to deep generative models (Goodfellow et al., 2016; Rezende et al., 2014). Therefore Bayesian deep learning is at the intersection of Bayesian techniques and Deep Learning. Neural networks with prior distributions placed over the weights as random variables that is equivalent to a probabilistic alternative to Gaussian processes have been studied extensively.

Although both methods are simple to formulate, inference in BNNs can not scale to modern neural network architectures due to the computational complexity induced by the high dimensionality of the weight space and the posterior over these parameters (and thus the loss surface) is often highly non-convex. As a result, exact inference is analytically intractable, and hence the approximate inference has been applied instead. Thus the current research in Bayesian Neural Networks has shifted to techniques to approximate the posterior distribution, leading to approximate BNNs, which is primarily divided into Variational Inference (VI) methods and Monte Carlo (MC) methods.

Markov Chain Monte Carlo (MCMC) was the gold standard method for Bayesian learning in neural networks, through the Metropolis-Hastings or Hamiltonian Monte Carlo (HMC) (Neal, 1996), gradient-based Monte Carlo sampling algorithms to estimate a unified predictive uncertainty. However, this batch-oriented method is computationally intractable in calculating the likelihood on large datasets. An extension of HMC framework, Stochastic gradient HMC (SGHMC) (Chen et al., 2014) provides for both scalability and generalization. Stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011) is an iterative optimisation technique that uses first-order Langevin dynamics in the stochastic gradient-based algorithms. To accelerate convergence, the second-order gradient algorithms, such as stochastic gradient Riemannian Hamiltonian Monte Carlo (SGRHMC) (Ma et al., 2015) and stochastic gradient Riemannian Langevin dynamics (SGRLD) (Li et al., 2019), have been developed. In practice, problem-specific tuning parameters such as the step size and the number of integration steps and the full gradient is quite difficult and often exhibit pathological curvature and saddle points. Finite learning rates introduces approximation errors in Markov Chain Monte Carlo (MCMC) inference in neural networks.

One alternative to Markov Chain Monte Carlo (MCMC) inference in neural networks is the Laplace approximation by (MacKay, 1992a) for a finite number of parameters, Kronecker Factored Block Diagonal Laplace Approximation (Martens and Grosse, 2015; Ritter et al., 2018). However, the Laplace approximation requires the computation of the inverse Hessian of the log-likelihood, which can be infeasible to compute for large networks. Diagonal approximations to the Hessian are possible, but performance can deteriorate considerably.

Variational Inference: Bayesian approximation techniques such as variational inference (VI), a popular technique that recasts intractable Bayesian integration as an optimisation

problem was first applied to neural networks by Hinton & Van Camp (Hinton and Van Camp, 1993). The true posterior distribution is approximated with a simpler variational distribution. Thus this approximation is biased but is often faster than sampling methods.

Almost two decades later, Graves (Graves, 2011) proposed a practical, scalable variational inference (VI) approach with fully factorised Gaussian variational posterior approximation over the weights of neural networks, which implemented a simple but biased gradient estimator. This method maximises a lower bound on the marginal likelihood of the neural networks, which is then optimised using a second approximation for stochastic gradient descent (SGD). The computation of this bound requires computing the expectation of the log of the numerator of the exact posterior under a factorised Gaussian approximation. This technique was generalised by Kingma and Welling (Kingma et al., 2015) which proposed the local reparameterisation trick for training deep latent variable models. Bayes By Backprop (BBP) Blundell et al. (Blundell et al., 2015) introduced an unbiased gradient estimator, leveraging on the generalised reparametrisation trick presented by Kingma & Welling (Kingma et al., 2015). More recent works, sampling-based variational inference and stochastic variational inference focus on modelling correlations between weights to capture the posterior dependencies (Welling and Teh, 2011).

An unbiased, differentiable, and scalable estimator was proposed, i.e., the reparameterisation trick for the ELBO in variational inference. Probabilistic Backpropagation (PBP) for scalable learning to large dataset (Hernández-Lobato and Adams, 2015) incorporates each likelihood factor in a single step, which is expected to be more accurate, Dropout-based VI (Gal and Ghahramani, 2016; Kingma et al., 2015) proposed to measure uncertainty. However, variational inference achieved excellent performance for medium-sized networks but was difficult to train on larger architectures such as deep residual networks (Gal et al., 2017a).

Dropout Variational Inference: Gal and Ghahramani (Gal and Ghahramani, 2016) used a spike and slab variational distribution to neural networks with dropout at test time as approximate variational Bayesian inference. Concrete dropout (Gal et al., 2017a) extends this idea to optimise the dropout probabilities as well. From a practical perspective, integrating flexibility, scalability, and predictive performance - all built-in characteristics of neural networks - with principled Bayesian uncertainty modelling. These approaches only require ensembling at test time dropout predictions. Following Gal (Gal, 2016), Ghoshal et al.(Ghoshal et al., 2019a) also showed similar results for neural networks with MC-Dropweights.

An alternative area of research re-interprets noisy versions of optimization algorithms: for example, noisy Adam (Hernández-Lobato and Adams, 2015) and noisy KFAC (Niculescu-Mizil and Caruana, 2005), as approximate variational inference. Residual estimation with an

I/O kernel (RIO) based framework to estimate uncertainty in any pre-trained standard neural network (Qiu et al., 2019) is quite useful in modelling the uncertainties.

Non-Bayesian approaches such as bootstrapping (Osband et al., 2016) and ensembling (Dusenberry et al., 2020; Lakshminarayanan et al., 2017; Perrone and Cooper, 1992) have shown modelling the uncertainties by evaluating the predictions of ensembles of classical networks. This comes at the expense of training and evaluating multiple deep learning models for the same task. A broad study of ensembling techniques to measure uncertainty is performed by (Ashukha et al., 2020).

Decomposition of predictive uncertainty by placing a distribution over the model output and epistemic uncertainty by placing a prior distribution over model's parameters (Depeweg et al., 2017; Kendall and Gal, 2017). Liu (Liu et al., 2019a) developed a principled Bayesian nonparametric augmentation framework for ensemble learning (BNE) to separate uncertainties in posterior predictive distribution from different sources (aleatoric, parametric, structural) for a continuous outcome with complex observational noise.

In Figure 2.4, an overview of the different types and methods is given.

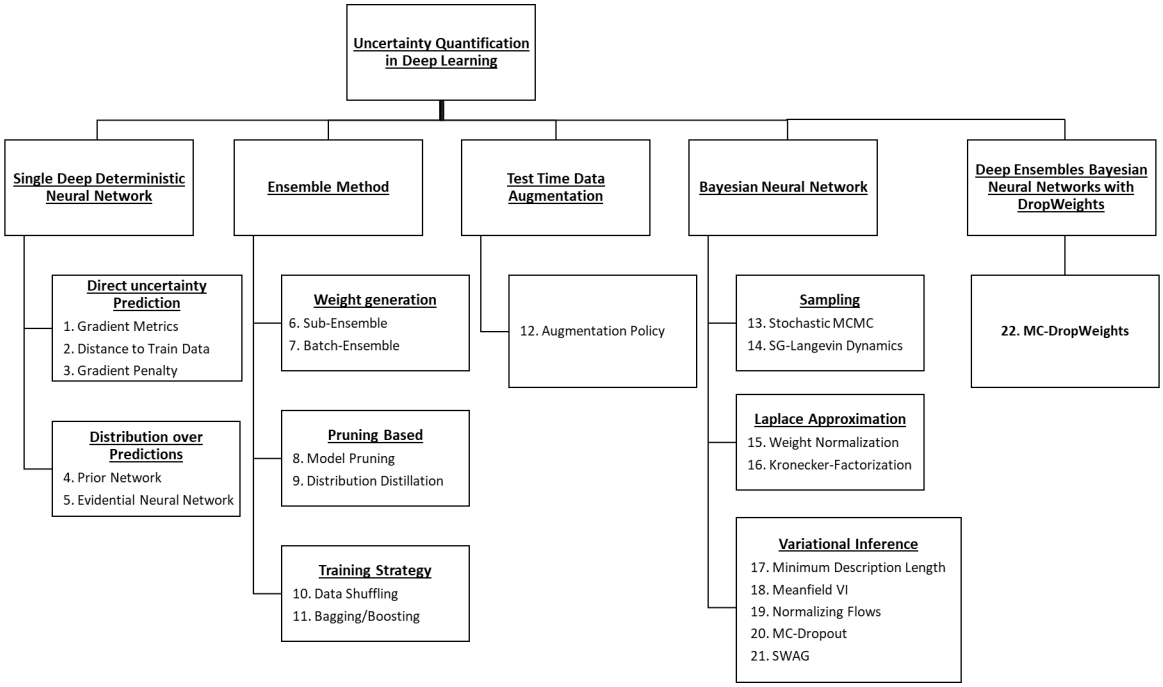


Fig. 2.4 Uncertainty quantification methods

1:[Oberdiek et al. (2018)], 2:[Lee and AlRegib (2020)], 3:[Ramalho and Miranda (2020)], 4:[Malinin and Gales (2018)], 5:[Sensoy et al. (2018)], 6:[Van Amersfoort et al. (2020)], 7:[Shorten and Khoshgoftaar (2019)], 8:[Wen et al. (2020)], 9:[Lindqvist et al. (2020); Malinin et al. (2020)], 10:[Lakshminarayanan et al. (2017)], 11:[Achrack et al. (2020)], 12:[Lyzhov et al. (2020); Wang et al. (2019)], 13:[Neal (1996); Nemeth and Fearnhead (2021)], 14:[Welling and Teh (2011)], 15:[Salimans and Kingma (2016)], 16:[Ritter et al. (2018)], 17:[Barber and Bishop (1998); Hinton and Van Camp (1993)], 18:[Blei et al. (2017)], 19:[Rezende and Mohamed (2015)], 20:[Gal and Ghahramani (2016)], 21:[Maddox et al. (2019)], 22:[Ghoshal et al. (2020, 2019a)]

Other techniques: A number of uncertainty quantification methods in traditional machine learning algorithms have extensively been studied that do not fall into any of the previous categories. Moreover, conformal prediction is a distribution-free and model agnostic framework that determines the level of confidence (degree of uncertainty) of a new prediction in machine learning models-is based on past experience (Hüllermeier and Waegeman, 2019; Papadopoulos et al., 2007; Shafer and Vovk, 2008). In particular, given an input, conformal prediction estimates a prediction interval in regression problems and a set of classes in classification problems. Conformal Prediction outline:

1. Select an apriori error rate $\alpha \in (0,1)$ depending on the application and leave a calibration set that the model hasn't seen yet.

2. Define a non-conformity score function $s(x,y) \in \mathbb{R}$ that encodes a heuristic notion of uncertainty.

3. Compute $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the calibration scores $s_1 = s(x_1,y_1),...,s_n = s(x_n,y_n)$ on the calibration dataset D_{cal}.

4. Compute confidence intervals using $\hat{q}$

$$C(x_{val}) = \{y : s(x,y) < \hat{q}\}$$

5. Conformal Prediction guarantees coverage property, i.e. $\mathbb{P}(y_{val} \in C(X_{val})) \geq 1-\alpha$ Assumption: Data must be independent and identically distributed.

Machine learning algorithms such as Support Vector Machines (SVMs) or k-Nearest Neighbours (k-NN) can be used to construct a conformal predictor. Therefore, it amounts to defining a non-conformity measure. Furthermore, the training data set is explored using scoring functions to detect the points most contributing to a given prediction and adjust its prediction correspondingly rather than the other way around.

2.6 Uncertainty Quantification in Medical Image Analysis

Recently, Deep learning (DL) has achieved remarkable success in medical image segmentation, classification and disease prediction. However, most of these methods provide overconfident prediction without quantifying uncertainty in model prediction, mainly when it is applied to an image that comes from different techniques used in radiology to form pictures, or from poorly taken scans, or rare phenotype of the disease that is not well represented in training data, etc. (Tanno et al., 2017b). On the other hand, Bayesian Deep Learning (BDL) methods provide a principled way to quantify uncertainty in medical imaging to improve the reliability of predictions. However, quantified uncertainty in deep learning does not well represent the model error and is prone to miscalibration (Laves et al., 2019b). Hence, calibrated uncertainty is essential as miscalibration can lead to decisions with fatal consequences in the safety-critical application domains.

Convolutional Neural Networks (CNN) have been highly successful in various applications. However, the accuracy of probabilistic model predictions is not sufficient because of variability in the imaging modality, variety of anatomical structures of diagnostic, pathologies and the expert annotators. Therefore, quantified aleatoric uncertainty (Epstein and Wang, 1995) and model uncertainty (Draper, 1994) estimation are needed to improve interpretability and would potentially allow clinicians to understand better the limits of the models, flag uncertain predictions, and highlight the cases that are not well covered in the training data. L. Joskowicz et al. (Joskowicz et al., 2019) conducted a study that quantifies the inter-observer variability with standard volume metrics of manual delineation of lesions and organ - 3193 contours of liver tumours (896), lung tumours (1085), kidney contours (434), and brain hematomas (497) on 490 slices of CT scans to establish a reference standard and for the evaluation of segmentation algorithms.

Variational approximation for Bayesian neural networks and ensemble learning techniques are two of the most widely-used methods estimating uncertainty in deep learning. For example, Nair et al. (Nair et al., 2020) explored MC-dropout to quantity four types of uncertainties, including variance of MC samples, predictive entropy, and Mutual Information (MI), in a 3D multiple sclerosis (MS) lesion segmentation CNN augmented to the voxel-based uncertainties within detected lesions from MRI sequences; showed that small lesions and lesion-boundaries are the most uncertain regions, which is consistent with human-inter-observer variability. In similar work by Eaton-Rosen et al. (Eaton-Rosen et al., 2018, 2019) proposed a method to convert voxel-wise brain tumour semantic segmentation uncertainty into volumetric uncertainty and calibrate the accuracy and reliability of confidence intervals from variance from MC-dropout to provide meaningful error bars over tumour volumes estimates to provide a form of quality control and quality assurance for clinical use. The

estimated uncertainty based on MC dropout has successfully demonstrated the benefits, applicability and limitations in disease grading in diagnosing diabetic retinopathy (DR) from retinal fundas images (Leibig et al., 2017), and an extension based on test-time augmentation was introduced by (Ayhan and Berens, 2018). Wang et al. (Wang et al., 2019) analysed a general aleatoric uncertainty estimation method based on test-time augmentation for deep CNN-based 2D fetal brain segmentation and 3D brain tumour Magnetic Resonance Images (MRI) segmentation tasks at both pixel and structure levels. Specifically, the distribution of the prediction was estimated by MC sampling with prior distributions of parameters of the output segmentation with image transformations and noise. Bragman et al. (Bragman et al., 2018) studied the value of uncertainty modelling for multi-task learning in the context of MR-only radiotherapy treatment planning where the synthetic CT image and the segmentation of organs at risk are simultaneously predicted from the input MRI image in the regression and segmentation of prostate cancer scans. Combalia et al. (Combalia et al., 2020) explored the MC dropout uncertainty estimation techniques for dermoscopic (skin lesion) image classification on data from ISIC Challenges and showed that uncertainty metrics could be used to detect difficult and out-of-distribution samples. Dahal et al. (Dahal et al., 2020) compared three ensembles based Monte Carlo Dropout as Bayesian Approximation, Horizontal Stacked Ensemble (HSE), test time augmentation (TTA) uncertainty techniques utilising three existing metrics - sample variance, predictive entropy mutual information and probabilistic atlas based uncertainty metrics to achieve an insight of uncertainty modelling for left ventricular segmentation from Ultrasound (US) images using two publicly available datasets in echocardiography - Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) of 2D apical four-chamber and two-chamber view sequences of 500 patients and Dynamic-Echonet dataset consists of 10,030 different echocardiography videos with the corresponding number of End Diastolic(ED) and End Systolic(ES) frame of the left ventricle structures - endocardium, epicardium, and left atrium for the experiments. They further demonstrated how uncertainty estimation could be to reject poor quality images and improve segmentation results. Do et al. (Do et al., 2020) proposed Monte Carlo dropout U-Net (Ronneberger et al., 2015) to segment myocardial arterial spin-labelled perfusion imaging and measure uncertainty. Specifically, adapt the Tversky loss function to adapt the model to obtain the most desirable performance. Ng et al. (Ng et al., 2018) compared Bayes by Backprop (BBB), Monte Carlo Dropout (MCDO), and Deep Ensembles (DE) in terms of accuracy, probability calibration, uncertainty on out-of-distribution images, and quality control in automated cardiac magnetic resonance imaging segmentation on a UNet model.

An alternative approach to these works, Araujo et al.(Araújo et al., 2020) proposed a novel Gaussian-sampling approach on a Multiple Instance Learning (MIL) framework in

deep learning model (DR|GRADUATE) to quantify uncertainty as a measure of trusted confidence for grading diabetic retinopathy (DR) using fundus images. Eaton-Rosen et al. (Eaton-Rosen et al., 2019) proposed a means to estimate prediction intervals as an output of a multi-task network on histopathological cell counting and white matter hyperintensity counting.

Li et al. (Li and Alstrøm, 2020) explored uncertainty calibration within an active learning framework for medical image segmentation. The model learns to acquire annotation of pixels in the most uncertain regions leads to a well-calibrated model. They showed that choosing regions to annotate instead of full images significantly reduce the labelling effort and improves the effectiveness of active learning. Łukasz Rączkowski (Rączkowski et al., 2019) utilised uncertainty for the selection of new samples for an accurate, reliable and active (shortly, ARA) learning framework for the classification of histopathological images of colorectal cancer. We provide further information about UQ methods applied in different medical application tasks in Table 1.1.

Application	Architecture	ML Task	Author	UQ Method
Diabetic Retinopathy	CNN	Classification	Leibig (Leibig et al., 2017)	MCDO
Brain Tumor Segmentation	ResNet	Segmentation	Jungo (Jungo et al., 2020)	MCDO
Pulmonary Nodule Detection	BCNN	Segmention	Ozdemir (Ozdemir et al., 2017)	VI
Brain Tumor	CNN	Classification	Tanno (Tanno et al., 2017a)	VI
Brain Tumor Cavity Segmentation	CNN	Segmention	Jungo (Jungo et al., 2018b)	MCDO
Fundas Images	ResNet50	Data Augmentation	Ayhan & Berens (Ayhan and Berens, 2018)	MCDO
Brain Tumor Segmentation	UNet	Segmention	Jungo (Jungo et al., 2018a)	MCDO
Brain Tumor Segmentation	CNN	Segmention	Wang (Li et al., 2017)	Weighted Loss function (Softmax variance)
Surgical Data of various diseases	CNN	Classiffication	Moccia (Moccia et al., 2018)	Superpixel
Brain Segmentation	CL	Segmentation	McClure (McClure et al., 2019)	Distributed weight consolidation
Brain Tumor	CNN	Segmentation	Wang (Wang et al., 2017)	Ensemble

Disability progression	CNN	Classification	Tousignant (Tousignant et al., 2019)	MCDO
Whole Brain Segmentation	B QuickNat	Segmentation	Roy (Roy et al., 2019)	MC samples for voxel-wise model
Brain Tumor	UNet	Segmentation	Jungo & Reyes (Jungo and Reyes, 2019)	Softmax entropy, MCDO & Ensembles
Optical Coherence Tomography (Retinal OCT) Scans	Unet	Sgmenttaion	Orlando (Orlando et al., 2019)	MCDO
Thoracic Disease	DenseNet-121	Classification	Ghesu (Ghesu et al., 2019)	MCDO
Colorectal Cancer	BCNN	IHC Classification	Raczkowski (Rączkowski et al., 2019)	VI
Skin Cancer	ResNet-101	Classification	Xue (Liu et al., 2019b)	Ensemble
Histopathological cell and white matter hyperintensity counting	UNet	Segmentation & Regression	Eaton-Rosen (Eaton-Rosen et al., 2019)	MCDO
Axon Myelin	Unet	Segmentation	Di Scandalea (Di Scandalea et al., 2019)	MCDO
Diabetic Retinopathy	BDL	Classification	Filos (Filos et al., 2019)	MCDO, VI, Ensembles
Cardiovascular Disease	Conditional GAN	Semantic segmentation	Ravanbakhsh (Ravanbakhsh et al., 2019)	adversarial discriminator
Brain tumor, cell membrane and chest Radiograph organ	BCNN	Segmentation	Jena (Jena and Awate, 2019)	MCDO
Brain tumour (Glioma) and Multiple Sclerosis (MS)	CNN	Classification	Tanno (Tanno et al., 2019)	VI
Organ Segmentation (Pancreas)	CNN & GCN	Segmentation	Soberanis-Mukul (Soberanis-Mukul et al., 2020)	MCDO
Lung nodule CT dataset and MICCAI2012 prostate MRI	Probabilistic UNET	Segmentation	Hu (Hu et al., 2019)	VI

Knee and Brain MRI	UNet & Adaptive CSNet	MRI reconstruction and Curve fitting	Hu (Hu et al., 2020)	MCDO and ensembles
Cardiovascular Disease	DCN	Segmentation	Luo (Luo et al., 2019)	MCDO
Lung Disease	UNet	Segmenttaion	Hoebel (Hoebel et al., 2020)	MCDO
Retinal OCT scans	BUnet	Segmentaion	Bogunovic (Bogunovic et al., 2020)	MCDO
Breast Cancer	BNN & UNet	Classification	Hiasa (Zhou et al., 2020)	MCDO
Material data (CT scans)	3D BCNN	Segmentation	LaBonte (LaBonte et al., 2019)	VI
Cardiovascular diseases	DenseNet, LTSM	Regression	Liao (Liao et al., 2021)	MCDO
Diabetic Retinopathy (DR)		Classification	Raghu	DUP (Direct Uncertainty Prediction)
Knee MRI	Resnet	Reconstruction and classification	Zhang	Active acquisition
Brain dMRI (diffusion magnetic resonance imaging) scans	LSTM	Tissue microstructure estimation	Ye (Ye et al., 2020)	Residual bootstrap strategy
Pancreas and Liver Tumor		Segmentation	Xia (Xia et al., 2020)	UMAT
Lung and prostate	Revesible PHiSeg	Segmentation	Gantenbein (Gantenbein et al., 2020)	VI
Cardiovascular and Retinal OCT	CNN	Segmentation	Bian (Bian et al., 2020)	Uncertainty Estimation and Segmentation Module (UESM) + Uncertainty aware Cross-Entropy (UCE) lo
COVID-19	Hierarchical Bayesian network	Classification	Donnat (Donnat and Holmes, 2020)	Stochastic Expectation-Maximization
Brain, Heart and Prostate	UNet	Segmenttion	Mehrtash (Mehrtash et al., 2020)	Ensembles
Colorectal cancer	CNN	Segmentation	Wickstrøm (Wickstrøm et al., 2020)	MCDO
PolyP	ResNet & DenseNet	Classification	Carneiro (Carneiro et al., 2020)	MC Integration

Brain Tumor	DenseUnet, ResUnet, SimUnet	Segmentation	Natekar (Natekar et al., 2020)	TTD (test time dropout) for VI
Colon and Skin cancers	Resnet	Segmentation	Li and Alstrøm (Li and Alstrøm, 2020)	MCDO
Cardiovascular disease	Resnet	Segmentation	Dahal (Duan et al., 2020)	TTA, HSE (Horizontal stacked ensemble), MC dropout
Autism	MLP	Segmentation	Li (Niu et al., 2020)	DistDeepSHAP (Distribution-based Deep Shapley value explanation)
Glands and infant brain tissues	ag-FCN (attention gated fully convolutional network)	Segmentation	Zheng (Martel et al., 2020)	dd-AL (Distribution discrepancy-based AL)
Lung disease and DR	Resnet	Classification	Wang (Wang et al., 2020a)	DRLA (Deep Reinforcement AL)
Esophago Gastro Duodenoscopy(EGD)	Resnet	Classification	Quan (Shu et al., 2019)	Bayesian uncertainty estimates and ensemble
Brain cell	DNN	Classification	Yuan (Yuan and Bar-Joseph, 2019)	Bayesian uncertainty
Prostate Lesion	CycleGAN	Segmentation	Chiou (Chiou et al., 2020)	Gaussian sampling
Left Atrium and kidney segmentation	TeacherStudent Model	Segmentation	Wang (Wang et al., 2020b)	Double-uncertainty weighted
Organ and skin lesion segmentation	ResUnet	Segmentation	Li (Li et al., 2020)	Self-loop uncertainty
Catheter segmentation	Deep Q learning and Dual-UNet	Segmentation	Yang (Yang et al., 2020)	Hybrid constraints
Brain volumes and healthy pregnant females	Unet	Segmentation	Venturini (Ritelli et al., 2013)	test-time augmentation and test-time dropout
Glaucoma Detection	Resnet	Classification	Yu (Yu et al., 2019)	FusionBranch
OCT	RelayNet	Segmentation	Huang (Huang et al., 2019)	MCDO
Breast cancer	BCNN	Classification	Khairnar (Khairnar et al., 2020)	BCNN

MRI scans, Prostate, Lung, Colorectal and Ovarian Cancer	Encoderdecoder network, U-Net, CNN)	Classification & Segmentaion	Ghesu (Ghesu et al., 2021)	Uncertainty-driven Bootstrapping and Dempster-Shafer evidence theory
Bone age prediction	BCNN	Regression	Eggenreich (Eggenreich et al., 2020)	VI and BCNNs
Organ segmentation	UNet	Segmentation	Soberanis- Mukul (Soberanis-Mukul et al., 2020)	GCN
Volume segmentation	Random walker	segmentation	Prassni (Prassni et al., 2010)	Guided probabilistic volume segmentation
Clinical data		Classification	Ulmer (Ulmer et al., 2008)	OoD detection
Automated breast ultrasound	Debse Unet	segmentation	Cao (Cao et al., 2020a)	Temporal ensembling

Table 2.1 A summary of various UQ methods applied in medical application tasks

2.7 Cost-Sensitive Calibrated Model Uncertainty

In regular safety-critical computer vision tasks, e.g. medical imaging, it is essential to obtain reliable predictive uncertainty from deep learning models. A well-calibrated model should represent the model error well, i.e. indicate high uncertainty when it is uncertain about its prediction. However, it tends to be miscalibrated (Guo et al., 2017). Uncertainty calibration is a challenging problem as there is no ground truth available for uncertainty estimates.

Calibration of deep neural networks involves accurately representing predictive probabilities with respect to true likelihood. Existing research on calibration models generally fall into one of three categories: (i) post-processing on model calibration, (ii) training the model with data augmentation (iii) probabilistic methods with Bayesian neural networks.

In practice, predictions of neural networks have been addressed by several nonparametric and parametric post-hoc calibration approaches on the pre-trained model, such as isotonic regression (Zadrozny and Elkan 2002) or Platt scaling (Platt 1999), temperature scaling (Guo et al., 2017) and Dirichlet calibration (Kull et al., 2019). However, they do not explicitly account for the quality of predictive uncertainty while training the model. Temperature scaling pushes the softmax output to slightly fewer confidence regions. Though existing post-hoc calibration methods perform well under in-data domain distribution in non-Bayesian deep neural networks but the model calibration performance degrades with data shift or

adversarial situations (Ovadia et al., 2019). Trainable calibration measures (Kumar et al., 2018) have been proposed that represent confidence calibration during training by optimising maximum mean calibration error.

Data augmentation methods such as Mixup (Thulasidasan et al., 2019) and AugMix (Hendrycks et al., 2020) improve model robustness. It is practically impossible to represent the wide spectrum of perturbations and corruptions during training conditions.

Deep-ensembles (Lakshminarayanan et al., 2017) has been shown to provide calibrated confidence (Ovadia et al., 2019) in the non-Bayesian ensemble of neural networks. However, it introduces the additional overhead of training multiple models and complexity during test time.

Approximate Bayesian inference methods such as variational inference (Blundell et al., 2015; Graves, 2011; Kingma et al., 2015), stochastic gradient variants of MCMC (Chen et al., 2014; Welling and Teh, 2011), Monte Carlo dropout (Gal, 2016) and SWAG (Maddox et al., 2019) do not capture the complete true posterior (Foong et al., 2019; Heek, 2018; Smith and Gal, 2018). This causes the model to produce overconfident predictions and fail to provide calibrated uncertainty.

In regular deep learning classification tasks, the aim is to improve model accuracy, i.e. to minimise the misclassification error, and thus all types of misclassification errors are assumed equally severe. However, in real-life medical diagnosis settings, the cost of false-positive is not the same as the cost of false-negative, i.e., misclassification errors are asymmetric, and the probability of different disease classes is not equal. So calibrated uncertainty in cost-sensitive deep learning application is crucially important in medical image analysis where safety is critical, and prediction problems are asymmetric, in the sense that different types of misclassification errors incur different costs or significant losses, which may result in the loss of life in some circumstances.

The costs depend on the predicted and true class label in case of class-dependent costs. The costs $c(k,l)$ of predicting class label k if the true label is l are usually organized into a KxK cost matrix where K is the number of classes. In general, the cost of predicting the correct class label y is minimal i.e. $c(y,y) \leq c(k,y)$ for all $k = 1,\ldots,K$. In the multi-class settings, cost matrices of a certain structure where $c(k,l) = c(l)$ for $k,l = 1,\ldots,K, k \neq l$. This means that the cost of misclassification is independent of the predicted class label. While some existing works have studied cost-sensitive neural networks (Elkan, 2001; Kukar et al., 1998).

In multi-class classification, Max-Heinrich (Laves et al., 2019b) investigated classwise logit scaling in addition to temperature scaling for calibrated uncertainty in Bayesian deep learning. Calibration of estimated uncertainty from Bayesian neural network in regression

has been addressed by (Kuleshov et al., 2018) auxiliary recalibration model using a technique inspired by Platt scaling. David et al. Ruhe et al. (2019) derived the bounds analytically on cross-entropy loss with respect to predictive uncertainty and showed that uncertainty could mitigate performance risk and loss.

Predictive probabilities from Bayesian Neural Network depend on prior assumptions about the behaviour of a random process (Fortuin, 2021). So, estimated uncertainty using MCMC methods (Neal, 1993) and Variational Inference (Gal and Ghahramani, 2016) by approximating the posterior distribution of the model tends to be poorly calibrated, i.e. predictive uncertainty is underestimated, and it does not represent well with the model error. Therefore, estimated uncertainty in deep learning is not of sufficient quality. Moreover, there was no work focused on cost-sensitive calibration of confidence or uncertainty in deep learning to the best of our knowledge. Therefore, we expect a better-calibrated uncertainty in variational Bayesian inference for cost-sensitive safety-critical applications.

2.8 Discussion

This chapter discussed the main ideas underlying the Bayesian neural network and how these tools are used in medical image analysis. The practicality of the approximate Bayesian inference techniques in deep learning above is mixed. We concluded that a technique should:

1. scale to large problems in high dimensional medical image data analysis such as segmentation, multi-class and multi-label classification and disease prediction

2. easily adapt to applications where misclassification cost is asymmetric in nature,

3. necessitates the evaluation methods on quantified uncertainty to explore the quality of uncertainty and the relationship between uncertainty estimation and model prediction error.

Chapter 3

Modelling Uncertainty in Neural Networks: DropWeights

"More data means that we need to be even more aware of what the evidence is actually worth." Sir David John Spiegelhalter (1953 -).

This chapter introduces the fundamentals of modelling uncertainty in deep learning. First, we will start with the basics of Bayesian modelling, the core concepts of variational inference as one form of approximate Bayesian modelling. Then, we will develop approximate Bayesian inference together with a deep learning generalised version of stochastic regularisation techniques (SRTs) such as DropWeights. Finally, we show extensive experiments in regression that show a significant and consistent improvement of the proposed algorithm when applied over all known uncertainty estimation methods.

3.1 Bayesian Modeling

Bayesian modelling provides a robust, mathematically grounded framework to combine prior knowledge with observations to enable the inference of probability distributions over model parameters to estimate all uncertainty within the model, both the uncertainty regarding the predictive inference and the uncertainty regarding the input parameters of the model.

Bayes' theorem: Back in the 18th century, Reverend Thomas Bayes (1702-1761) showed how to do inference about hypothesis (i.e. uncertain quantities) from data (i.e. measured quantities). Bayes' theorem is formulated as:

$$P(hypothesis|Data) = \frac{P(Data|hypothesis)P(hypothesis)}{P(Data)} \tag{3.1}$$

A Bayesian defines a model, selects a prior, collects data, computes the posterior, and makes predictions. This process is called inference. Being Bayesian in machine learning means dealing with parameters uncertainty (Neal, 1993).

Given a dataset $D = \{X, Y\}$, where $X = \{x_1, \ldots, x_n\}$ denotes a set of training examples and $Y = (y_1, \ldots, y_n)^T$ holds the corresponding class labels, instead of point estimates, the neural network learns **posterior distribution over the space of weight parameters** $p(\mathbf{w}|\mathbf{X}, \mathbf{Y})$ in accordance with the following formula:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{Y}|\mathbf{X})} \tag{3.2}$$

In deep learning, each distribution is commonly referred to as:

- $p(\mathbf{w}|\mathbf{X}, \mathbf{Y})$: the posterior over the set of random variables w. This distribution captures the most probable model parameters given our observed data.
- $p(\mathbf{w})$: the prior distribution over the space of parameters. This distribution represents subjective beliefs about a parameter before observing any data points or new evidence is introduced.
- $p(\mathbf{Y}|\mathbf{X}, \mathbf{w})$: the likelihood distribution of the parameters, which is a function of $\mathbf{w}$. This is the probability of the data given the set of random variables. The likelihood indicates how well the $\mathbf{w}$ parameters explain the observed data. We may assume either a Gaussian likelihood for regression or sigmoid or softmax likelihood for classification tasks.
- $p(\mathbf{Y}|\mathbf{X})$: this quantity measures how appropriate the model is for the data, such that it guarantees that the posterior distribution is a valid probability distribution, also called model evidence. Calculating this integration for **predictive distribution** is also referred to as marginalising the likelihood over all possible model parameter values $\mathbf{w}$.

$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w} \tag{3.3}$$

- **Inference**: To make predictions for a new input x^*, we obtain the predictive posterior probabilities by integration (i.e. by averaging over all the possible parameters' configurations). The posterior predictive distribution $p(y^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y})$ reflects the belief in a class label y^* for a given test sample $\mathbf{x}^*$ after observing data $\mathbf{D}(\mathbf{X}, \mathbf{Y})$:

$$p(y^*|\mathbf{x}^*,\mathbf{X},\mathbf{Y}) \quad = \int p(y^*|\mathbf{x}^*,\mathbf{w})p(\mathbf{w}|\mathbf{X},\mathbf{Y})\, d\mathbf{w} \tag{3.4}$$

Equation (2.4) represents a Bayesian model average (BMA). Instead of a single configuration of parameters w, the model evidence is an integral over the likelihood and prior for all possible parameter configurations of w, weighted by their posterior probabilities. This process is called the marginalisation of the parameters w since the predictive distribution of interest no longer conditions on w. Ideally it is expected marginalization overall uncertain quantities - i.e. average w.r.t. all possible model parameter values w, each weighted by its plausible prior $p(w)$.

Exact Bayesian inference is computationally intractable due to the integrals (the intricate form of the posterior and the vast number of parameters) in the marginal likelihood in equation (2.4). Hence, work around this problem by approximating the universal approximation capacity of large neural networks and available computing power. Bayesian inference for neural networks is typically performed via Monte Carlo estimation, stochastic variational inference or ensemble methods (Perrone and Cooper, 1992). There are multiple challenges in Markov chain Monte Carlo (MCMC) methods in deep learning attributed to the non-linear hierarchical structure of neural networks, including the associated covariance between parameters, non-identifiability, highly correlated samples arising from weight symmetries, lack of a priori knowledge about the parameter space, lack of convergence and ultimately the lack of scalability in high-dimensional data (Papamarkou et al., 2019). Stochastic optimisation was introduced to overcome the scalability issue in the Monte Carlo method. However, by increasing the mini-batch size, the Monte Carlo Batch Normalisation (MCBN), batch normalisation layers become deterministic and unable to capture model uncertainty. Later in the chapter, we review methods stochastic variational inference technique for practical approximate inference in BNNs. The sections thereafter move to the extensions to Monte Carlo dropweights.

3.2 DropWeights in Neural Network

Convolutional Neural Networks (CNN) in deep learning have shown outstanding performances in biomedical image processing. However, CNNs are prone to over-fitting when trained with small datasets. Several techniques have been developed for regularising neural networks, such as adding an l2 penalty on the network, Bayesian methods (MacKay, 1992b),

weight elimination and early stopping of training (Caruana et al., 2001). In deep neural networks, co-adaptation means that some neurons are highly dependent on others, significantly impacting model performance. Overfitting can be reduced by using Dropout (Srivastava et al., 2014) and DropConnects (Wan et al., 2013), to prevent complex co-adaptations on the training data. Network pruning by dropping connections has been widely studied to compress pre-trained, fully connected neural network models. It can also reduce the network complexity and over-fitting (Hassibi and Stork, 1993; LeCun et al., 2015). Bayesian Neural Networks (BNNs) is used to mitigate overfitting and can be trained with small datasets (MacKay, 1992b; Neal, 1993).

The number of neurons in a human brain stays constant throughout its life, but synapse connectivity changes dramatically over time (Stiles and Jernigan, 2010). Using this fact, I have developed a technique called "DropWeights", which randomly drops connections, i.e. incoming or outgoing weights are set to zeros, including drop neurones. DropWeights can be considered as the combination of a generalised version of Dropout and DropConnects, and this comprises the method used for regularising deep neural networks. DropWeights is a kind of ensemble (Perrone and Cooper, 1992) and approximates the output by a moment matched Gaussian, and it produces even more possible models since there are almost always more connections than units. Figure 1 below illustrates the DropWeights strategy. This DropWeights method converts a dense, fully-connected neural network to dynamically sparse representations on the weights during training and test time when DropWeights are turned on (Goodfellow et al., 2016; Olshausen and Field, 1996).

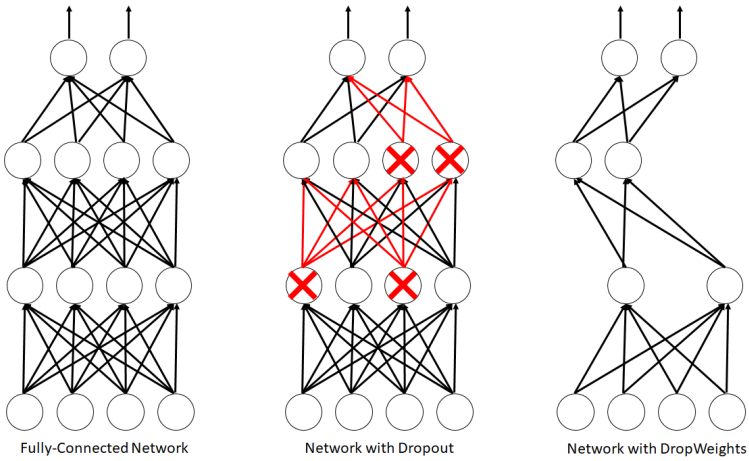


Fig. 3.1 A graphical illustration of DropWeights strategy

Considering DropWeights applied to a single fully-connected layer of deep neural network K_{i-1} dimensional input $X = \{x_1, x_2 \ldots x_N\}$, i^{th} layer of neural network K_i units would output a K_i dimensional activation vectors $a_i = \sigma(W_i x)$ where W_i is the $K_{i-1} x K_i$ weight parameters

including biases and $\sigma(.)$ the nonlinear activation function. When DropWeights is applied to the outputs of a full-connected layer, different neurons allocate different drop probabilities to enable the model to adjust the drop probability of the weight dynamically, eventually leading more sparse features of network model extraction.

The feed-forward operation of neural networks with DropWeights can be described as:

$$\rho_{ij}^{(l)} = p_{drop}(y_j^{(l-1)}) \tag{3.5}$$

$$\mathrm{M}_{ij}^{(1)} \sim \mathrm{Bernoulli}(\rho_{ij}^{(l)}) \tag{3.6}$$

$$\tilde{W}_{ij}^{(l)} = \tilde{W}_{ij}^{(l)} \odot (M_{ij}^{(l)} > \rho_{ij}^{(l)}) \tag{3.7}$$

$$z_i^l = \sum_{j=1}^{n} \tilde{W}_{ij}^{(l)} y_j^{(l-1)} + b_i^{(l)} \tag{3.8}$$

$$y^{(l)} = f(z^l) \tag{3.9}$$

For a DropWeights layer, the output activations can be written as:

$$\sum_M f((M \odot W)x) \approx f(\sum_M (M \odot W)x) \tag{3.10}$$

Where M is a binary mask encoding the connection information drawn independently from a Bernoulli distribution with probability p, $W_{ij}^{(l)}$ is for the connection weight between the j neuron in the $l-1$ layer and the i neuron in the l layer; $\rho_{ij}^{(l)}$ is for the drop probability of the weight associated with the weight $W_{ij}^{(l)}$ being set to 0; $p_{drop}(.)$ is for the calculation function of weight drop probability. Hadamard product $\odot$ denotes the element-wise product of matrices. The input of the activation function is a weighted sum of Bernoulli variables and can be approximated by a Gaussian distribution. During test time, we drew samples of $z^{(l+1)}$ and fed the samples into the activation function f. This represents the mixture model interpretation of DropWeights, where the output is a total of 2^M different network architectures possible, each with weight $p(M)$. Each of these corresponds to some of the connections being present and some being dropped. DropWeights is functionally equivalent

to an ensemble rather than a single model. The output value of the sparsity formula is in the range [0, 1.0].

3.3 Deep Ensembles Bayesian Neural Networks with DropWeights - Measuring Uncertainty in Deep Learning

I have proposed a novel technique, "Deep Ensembles Bayesian Neural Networks with DropWeights" (DE-BNN), for estimating uncertainty in Deep learning that yields high-quality predictive uncertainty estimates and outperforms existing methods (e.g., MC-Dropout and our MC-DropWeights). I present the first approach (to the best of our knowledge), a stochastic ensemble of MC-DropWeights models characterised by a different set of DropWeights rate probabilities for estimating uncertainty in Bayesian Deep Learning.

3.3.1 Bayesian Neural Networks for uncertainty estimation

A standard neural networks training via optimization could be interpreted as Maximum Likelihood Estimation (MLE) or Maximum A Posteriori (MAP) for the weights. However, MLE or MAP compute a single estimate, instead of a full distribution so ignores uncertainty in the model. Bayesian neural networks promise improved predictions by directly modelling the uncertainty by specifying prior distributions over the weights (MacKay, 1992b; Neal, 1996). The motivation of quantifying uncertainty along with the predictive probability comes from the neural network's uncertainty estimates under a function approximation, $f^{w(x)}$. The placement of a prior $p(w_i)$ over each weight $w_i \in W$ leads to a distribution over a parametric set of functions. In fact, as shown by Neal (Neal, 1996), the prior provided by a fully connected neural network with a single hidden layer tends to a Gaussian process prior as the number of neurons of the hidden layer tends to infinity. This highlights the strong relationship between Gaussian processes and BNNs.

Given a neural network model with L layers parametrized by weights $w = \{W_i\}_{i=1}^{L}$ are the random weights of layer i and a dataset $D = (X,Y)$, Bayesian inference calculates the posterior distribution of the weights given the data, $p(w|D)$. Assuming, we place a prior distribution $p(w)$ on weights and bias vectors in a neural network. The predictive posterior distribution defines a distribution of such predictions over class probabilities of an unknown label y^* of a new test sample x^* of an infinite number of neural networks with all possible configuration of the weights is given by:

$$p(y^*|x^*,X,Y) = \int p(y^*|x^*,w)p(w|X,Y)dw \tag{3.11}$$

In practice, the above equation $p(w|X,Y)$ is computationally intractable.

We adopted a popular method called variational inference to approximate the posterior distribution of the weights. We define simpler tractable distribution $q_\theta(w)$ with known form with variational parameters θ, whose structure is easy to evaluate. We want our approximating distribution to be as close as possible to the posterior distribution obtained from the original model. The variational parameters θ are fitted so that $q_\theta(w)$ approximates the desired posterior $p(w|X,Y)$. This fitted variational distribution is used for model predictions rather than the true posterior. To measure the difference between two probability distributions q(x) and p(x) over the same dataset (x) called Kullback-Leibler divergence, or KL-divergence, which is not symmetric, can be defined as:

$$KL(q(x) \,||\, p(x)) = \mathbb{E}_{q(x)}[\log \frac{q(x)}{p(x)}] = \int q(x) \log \frac{q(x)}{p(x)} dx \tag{3.12}$$

To make the variational distribution $q(w)$ close to the posterior $p(w|D)$, we want to minimise the KL-divergence between these two distributions. Applying Bayes' rule to $p(w|D)$ we obtain:

$$KL(q(w|\theta) \,||\, p(w|D)) = \mathbb{E}_{q(w|\theta)} \log \frac{q(w|\theta)}{p(D|w)p(w)} p(D) \tag{3.13}$$

This approximation process can be formulated as an optimization problem where divergence measures the discrepancy between q and p. The most prominent divergence measure is the Kullback-Leibler (KL) divergence between $q_\theta(w)$ and the true posterior $p(w|D)$, which is widely used in machine learning and can be defined as:

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

The KL divergence can not be computed directly due to $\log p(D)$. However, by rearranging we can obtain:

$$
\log p(D) = KL(q(w) \,||\, p(w|D)) + \mathbb{E}_{q(w)}[\log p(D|w)] - KL(q(w) \,||\, p(w)) \tag{3.14}
$$

The log marginal likelihood $\log p(D)$ doesn't depend on θ so it is constant. In order to minimise the divergence of $q_\theta(w)$ and the true posterior $p(w|D)$ i.e. KL divergence $KL(q(w) \,||\, p(w|D))$ is equivalent to maximize $\mathbb{E}_{q(w)}[\log p(D|w)] - KL(q(w) \,||\, p(w))$. The latter expression is also known as evidence lower bound (ELBO).

As shown by (Kingma et al., 2015), a variational lower bound that can be maximised with respect to the variational parameters θ is derived on the (conditional) marginal log-likelihood (i.e. log model evidence) as:

$$
\log p(D) = \int p(D|w)p(w)dw = \log \int \frac{q(w)}{q(w} p(D|w)p(w)dw \tag{3.15}
$$

Applying Jenson's inequality:

$$\geq \int q(w) \log \frac{p(D|w)p(w)}{q(w)} dw \tag{3.16}$$

$$\tag{3.17}$$

$$= -\int q(w) \log \frac{q(w)}{p(w)} dw + \int q(w) \log p(D|w) dw \tag{3.18}$$

$$\tag{3.19}$$

$$= -KL(q(w) \,||\, p(w|D)) + \mathbb{E}_{q(w}[\log p(D|w] \tag{3.20}$$

Thus we have that:

$$\log p(D) \geq -KL(q(w) \,||\, p(w|D)) + \mathbb{E}_{q(w}[\log p(D|w] \tag{3.21}$$

However, due to the non-negative properties of KL-divergence, we conclude that the ELBO is less than or equal to the log probability of the data. Therefore, finding variational parameters θ to minimise the KL divergence between the variational distribution $q(w|\theta)$ and the true posterior $p(w|D)$ is the same as maximising the ELBO. In effect, variational inference translates the problem of inference over the weight distribution into the optimisation problem of maximising the ELBO. Once the ELBO objective is defined, we can sample from $q(w)$ and use backpropagation, like in DNNs, to find optimal values of the variational parameters that maximise the ELBO.

The loss function used to train the Bayesian neural network corresponds to the negative ELBO. It is simple to evaluate as opposed to the exact one. We use the variational distribution $q(w)$ instead of $p(w|X,Y)$. This variational distribution is chosen close to $p(.|X,Y)$, as it minimises the Kullback Leibler divergence between the approximate posterior an the prior over w:

$$L_{VI} := \underbrace{\int q(w|\theta) \log p(Y|X,w) dw}_{\text{Log Likelihood}} - \underbrace{\int q(w|\theta) \log \frac{q(w|\theta)}{p(w)} dw}_{\text{KL divergence}} \tag{3.22}$$

$$\tag{3.23}$$

$$= E_{q(w|\theta)}[\log p(Y|X,w)] - D_{KL}(q(w|\theta) \,\|\, p(w)) \tag{3.24}$$

Hence, the ELBO can be decomposed into two terms: Log-likelihood and KL divergence. The prior KL term approximates the true distribution $p(w)$ by $q(w)$, can be evaluated

analytically while the integral in the first term cannot be computed exactly for a non-linear neural network.

3.3.2 DropWeights approximation in deep learning

Recently, Gal & Ghahramani (Gal, 2016) shown that a neural network with any number of layers and arbitrary non-linearities, with dropout applied after every weight layer is mathematically equivalent to approximate variational inference in the deep GP model. In this section, following Gal & Ghahramani, we then approximate neural network with DropWeights applied on fully connected layers via variational inference to obtain uncertainties in deep learning, for which weights are dropped by drawing from a Bernoulli prior with probability p dropweight rate for setting a weight to zero.

We have to define a variational distribution on weight parameters and to develop the objective of maximisation on the log evidence lower bound. In neural networks, like Dropout (Gal and Ghahramani, 2016), we can consider approximating distribution is DropWeights. This means weights are drawn from the Bayesian neural network with DropWeights, $W_i = M_i \odot Z_i = M_i \odot diag([z_{i,j}]_{j=1}^{K_i})$, where $w = \{W_i\}_{i=1}^{L}$ and M_i is the matrix of variational parameters i.e: weight matrix multiplied by a diagonal matrix formed by binary random vector Z_i, whose elements are distributed as: $M_{ij}^{(1)} \sim \text{Bernoulli}(\rho_{ij}^{(l)})$ for $i = 1,\ldots,L$ and $j = 1,\ldots,K_{i-1}$. Therefore, $M_{ij}^{(1)} = 0$ corresponds to the weight in the networks are being dropped out.

The main idea of this technique is to keep DropWeights enabled in the networks by performing multiple model calls during prediction so that different weights are dropped to zero across different model calls. It can be considered as Bayesian sampling from a variational distribution of deep learning models. We start with rewriting the first term of equation (3.23) as a sum over all samples and re-parametrise the integral term so that it only depends on the Bernoulli distribution instead of weights (w) directly. In summary, approximate variational inference performing DropWeights with a Bernoulli approximating distribution can be interpreted as trained weights of the neural network by minimizing the cross-entropy loss between the distribution of the true class labels and the softmax network output. Thus the loss is:

$$L_{DropWeights} = -\sum_{i=1}^{N} \log \frac{e^{f(x_i)}}{\sum_j e^{f(x_i)}} + \lambda \sum_{l=1}^{L} (W_l)^2 \tag{3.25}$$

Note that Monte Carlo sampling from $q(w)$ is equivalent to performing DropWeights during training, hence we get the Bayesian network perspective as well for already trained models.

3.3.3 Measuring the uncertainty at test time

Given dataset $X = \{x_1, x_2 \dots x_N\}$ and the corresponding labels $Y = \{y_1, y_2 \dots y_N\}$ where $X \in R^d$ be a d-dimensional input vector and $Y \in \{1 \dots C\}$ with $\mathrm{y_i} \in \{1 \dots \mathrm{C}\}$, given C class label, a set of independent and identically distributed (i.i.d.) training samples size $\mathrm{N}\{x_i, y_i\}$ for $i = 1$ to N, the task is to find a function $f : X \to Y$ using weights of neural net parameters θ as close as possible to the original function that has generated the outputs Y. We assume probability distributions over weights are Gaussians, we have a mean μ and a variance σ^2.

The DropWeights layers are kept active during inference to model uncertainty over weights for a given input sample and performing multiple predictions. We inferred using equation (3.26) that, after multiple forward passes (for T repetitions), we approximate the posterior distribution of class probabilities from the trained network with DropWeights.

$$\hat{\mu}_{pred} \approx \frac{1}{T} \sum_{t=1}^{T} p(\hat{y}|\hat{x}, \theta(\hat{w}_t)) \tag{3.26}$$

Practically, the expectation of an unknown input sample prediction label $\hat{y}$ of test sample data $\hat{x}$ by marginalizing the parameters is called the predictive mean of the model. For each test sample $\hat{x}$, the class with the largest predictive mean μ_{pred} is selected as the output prediction.

The law of the total variance for random variables X and Y on the same probability space, and finite variance of Y:

$$Var(Y) = E[Var(Y|X)] + Var[E[Y|X] \tag{3.27}$$

The variance is the average squared difference between any given output $\hat{y}$ and the expected value of any given input $\hat{y}$. The variance can be considered as a measure of uncertainty. The predictive uncertainty can be decomposed into aleatoric and epistemic uncertainty (Kwon et al., 2018).

$$Var_q(p(\hat{y}|\hat{x})) = E_q\left[(y - E[y])^2\right] = E_q\left[yy^T\right] - E_q[y]\,E_q[y]^T$$

$$= \underbrace{\int_w \left[diag\left\{E_{p(\hat{y}|\hat{x},w)}[\hat{y}]\right\} - E_{p(\hat{y}|\hat{x},w)}[\hat{y}]\, E_{p(\hat{y}|\hat{x},w)}[\hat{y}]^T \right] q_{\hat{\theta}}(w)dw}_{\text{aleatoric}}$$

$$+ \underbrace{\int_w \left[E_{p(\hat{y}|\hat{x},w)}[\hat{y}] - E_{q_{\hat{\theta}}(\hat{y}|\hat{x})}(\hat{y})\right] \left[E_{p(\hat{y}|\hat{x},w)}[\hat{y}] - E_{q_{\hat{\theta}}(\hat{y}|\hat{x})}(\hat{y})\right]^T q_{\hat{\theta}}(w)dw}_{\text{epistemic}}$$

*hspace*5(3.28).

- w is the space of all possible values for network weights w, denoted $w \in w$.
- diag is the diagonal matrix. There would have only variances if diagonal matrix were the variance-covariance matrix of weights .
- $E[\hat{y}]$ is the expected output of the input $\hat{x}$.
- $q(w)$ is the variational posterior distribution which approximates the intractable posterior distribution $p(w|D)$.

The uncertainty score for the prediction can be estimated as:

$$\text{Aleatoric uncertainty: } \frac{1}{T}\sum_{t=1}^{T} \text{diag}(\hat{y}_t) - \hat{y}_t \hat{y}_t^T \tag{3.29}$$

$$\text{where } \hat{y}_t = y(\hat{w}_t) = \text{Softmax}\left\{f^{\hat{w}_t}(\hat{x})\right\}$$

The main idea of the proposed tractable method is to decompose the variability of the predicted probability directly from Softmax predictive probabilities by performing T times predictions. The Softmax output multiplied by its transpose is subtracted from the diagonal matrix and finally we calculate the average over the variability of the output coming from the data set by dividing it all by T.

We can estimate epistemic uncertainty as:

$$\text{Epistemic uncertainty: } \frac{1}{T}\sum_{t=1}^{T} (\hat{y}_t - \bar{y}_t)(\hat{y}_t - \bar{y}_t)^T \tag{3.30}$$

where $\bar{y}_t = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t$

The epistemic uncertainty is the average of the variability of the predictive probabilities in the model and is inversely-proportional to the validation accuracy.

Epistemic uncertainty represents uncertainty over configurations of weights for a given limited data, whereas aleatoric uncertainty comes from inherent randomness or noise in the

measurement process. In regression, epistemic uncertainty is the spread of the predictive distribution along the x axis. It grows as we move away from the data.

Intuitively, the mean of the predictive posterior corresponds to the point estimates, and the predictive posterior's width reflects the predictions' reliability. We call this approach MC-DropWeights which is a generalisation over the previous work referred to as MC-Dropout (Gal, 2016). Our method is computationally efficient for large neural networks and uses a simple heuristic to choose the threshold to drop weights. A summary of these steps is provided in Algorithm 1.

Algorithm 1: Training a Bayesian Neural Network with DropWeights

Input: Dataset: $D = \{(x_i, y_i)_{i=1}^{N}\}$, given C number of classes ;
Learning rate β; Number of epoch e; Dropweights rate p
Initialization: Model weights θ in neural network by He initialization (He et al., 2015)
Result: Model with updated weights θ
for $i \leftarrow 1$ ***to*** e **do**
  Forward Pass:
  # g is Convolutional Neural Network (CNN), with W_g being the CNN filters (and biases)
  Extract Features from multi-layered CNN: $v \leftarrow g(x; W_g)$
  Random sample M mask: $M_{ij} \leftarrow Bernoulli(p)$
  # W is a fully connected weight matrix, a is a non-linear activation function and M is the binary mask matrix
  Compute activations: $r = a((M * W)v)$
  # Softmax function s takes input r and uses parameters W_k to map to a C dimensional output
  Compute output: $o = r(s; W_s)$
  Backpropagation:
  Differentiate loss L_θ wrt to θ
  Update softmax layer: $W_s = W_s - \beta * L_{W_s}$
  Update Weights in DropWeights Layer: $W = W - \beta(M * L_w)$
  Update Weights in hidden Layer: $W_g = W_g - \beta L_{W_g}$

3.3.4 Ensembles Method

The ensemble method efficiently aggregates the predictions over a collection of models, i.e. average predictions over a diverse set of functions and uses all the training data without over-fitting to obtain a more robust model to improve predictive performance when the actual model does not lie within the hypothesis class (Cao et al., 2020b; Dietterich, 2000; Ganaie

et al., 2021; Perrone and Cooper, 1992). In general neural networks often suffer from high variance or the tendency to over-fit. The ensemble model improve the bias-variance trade-off of the overall model, spread of prediction, enhances reliability and robustness.

The basic ensemble method output is to average predictions of several independently trained models with network weights given by the posterior probability of each model given the training data. Ensemble method reduces the variance of neural network predictions and reduces generalization error. One drawback, however, is that it is computationally more expensive than the other methods.

3.3.5 Deep Ensembles Bayesian Neural Networks with DropWeights

DropWeights is an efficient way to average a large number of neural nets that gives us an alternative to doing the correct Bayesian thing the alternative probably doesn't work quite as well as doing the correct Bayesian thing but it's much more practical.

Deep Ensembles is another sampling-based approach to quantify the predictive uncertainty of DNNs (Hansen and Salamon, 1990; Lakshminarayanan et al., 2017). As with other ensembling methods, multiple models having the same basic architecture are trained. Their softmax outputs are then averaged to obtain the mean and variance of predictive probability. Although bagging and bootstrapping are often employed in other ensemble learning methods, deep ensembles generally perform better with large data; however, training takes a significant amount of time. So it is desirable to train each ensemble member in parallel and on the entire dataset. Data augmentation with adversarial examples has been used (Lakshminarayanan et al., 2017) to smooth the predictive distribution of deep ensembles and make them more robust to adversarial attacks. Even though Deep Ensembles have no Bayesian grounding, empirically, they often outperform Monte-Carlo dropout, even requiring significantly fewer samples, as, e.g. (Beluch et al., 2018; Lakshminarayanan et al., 2017) have shown. (Beluch et al., 2018) examined why Deep Ensembles generally perform better and suggest that it is mainly due to an increased model capacity, as Deep Ensembles require no dropout at inference time, and due to different weight initialisations, which cause each network to converge to a different local minimum.

The current state-of-the-art Bayesian neural networks learn a distribution over weights for estimating predictive uncertainty; however, they suffer from the "mode in collapse" problem in deep convolutional neural networks (CNNs) when dealing with complex high-dimensional image data such as medical images (X-Rays, PET/CT, SPECT, MRI, Ultrasound, EEG, ECG etc.). Moreover, to estimate uncertainty in deep learning, the quality of Bayesian posterior distribution depends on prior specification and posterior approximation to translate weight uncertainty to predictive uncertainty. Therefore, adversarial examples can easily fool deep

learning models, e.g. small perturbations in the input images, resulting in overconfident predictions in variational inference.

The selection of the "optimal" parameters (w) is an optimisation problem with many local minima which depends on hyper-parameters. This randomness on network weights tends to differentiate the errors of the networks. The ensemble method is an alternative strategy to aggregate the predictions over a collection of independently trained models to improve the overall results. The collective decision made by the ensemble method reduces the generalisation error by reducing either the bias or variance of the error or both than any single network. So we can average to estimate the model uncertainty by training several models and calculating the variance of their output prediction by approximately marginalising over model parameters using MC-DropWeights as:

$$q(y|\mathbf{x}) = \mathbf{E}_{q(w)}[\log p(y|\mathbf{x},\mathbf{w})] \approx \frac{1}{M}\frac{1}{N}\sum_{m=0.0}^{M}\sum_{i=1}^{N}\log p(y|\mathbf{x},\mathbf{w}^{(i)}); \tag{3.31}$$

Interestingly, DropWeights in a neural network can be considered an ensemble technique, where the predictions are averaged over multiple networks trained by "dropping" certain weights. This can be seen as drawing from an infinite ensemble of networks with N number of forwarding passes, M the number of network models with DropWeights rate between 0 and 1.0 to estimate uncertainty. The ensemble method results in better uncertainty estimates from the stochastic ensemble.

3.4 Performance of Deep Ensembles BNN with DropWeights

We assess the models' confidence quantitatively from our approach in deriving practical inference techniques in Bayesian neural networks (BNNs) with DropWeights by empirical experiments. We evaluate the quality of quantified uncertainty based on two standard metrics: predictive probability distributions as measured by Predictive Log-Likelihood (PLL) and root mean square error (RMSE).

3.4.1 Bayesian neural networks for regression

We evaluate the predictive distribution obtained by MC-DropWeights in toy data (Hernández-Lobato and Adams, 2015). Dataset was generated by uniformly sampling 20 inputs x at random intervals [-6, 6]. For each input value of x obtained, the corresponding target y is computed as $y = x^3 + \varepsilon_n$, where $\varepsilon_n \approx (N)(0,9)$. We trained a neural network with one layer and 100 hidden units to these data using MC-DropWeights. We compare MC-DropWeights

with MAP, MC-Dropout, Hamilton Monte Carlo (HMC), Bayes by Backprop (BBB) and Deep ensembles, using 40 training epochs in all these methods. This example (Fig. 3.2) illustrates the issue of low predictive uncertainty in unseen regions in deep learning methods and the more reliable uncertainty of function approximations in BNN. The observations are shown as black dots, the true data generating function is displayed as a blue dashed line, and the mean predictions are shown as a dark continuous line. Credible intervals corresponding to ±3 standard deviations from the mean are shown as a light shaded area.

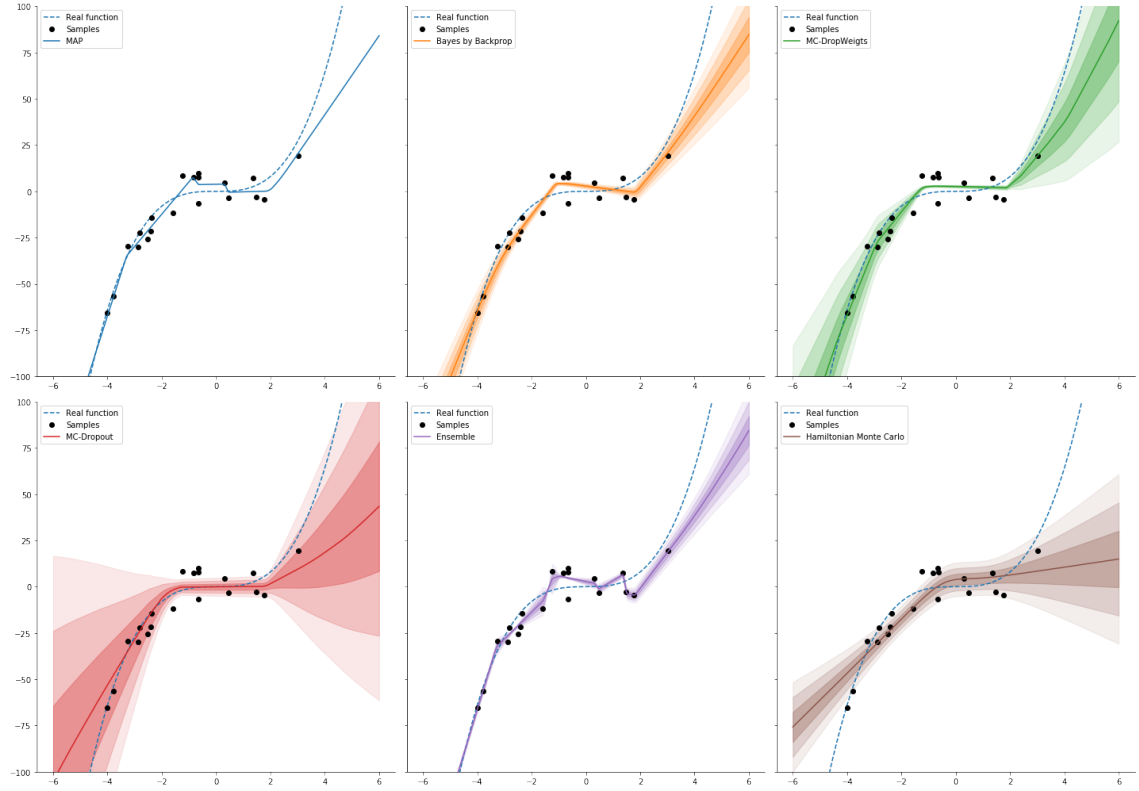


Fig. 3.2 Toy example inspired by (Hernández-Lobato and Adams, 2015): Predictions made by each method on the toy data set.

The proposed "Deep Ensembles Bayesian Neural Networks with DropWeights" exhibits the best trade-off between predictive uncertainty and regression fit. Hamilton Monte Carlo (HMC), MC-Dropout produce a good fit but underestimate the predictive uncertainty. MAP, Deep ensembles, and Bayes by Backprop (BBB) achieve a slightly worse fit and predictive uncertainty than MC-DropWeights.

3.4.2 Model Uncertainty Performance

We next evaluate the uncertainty in models' prediction quantitatively to understand how much confidence in our DropWeights based method while deriving practical inference

techniques in Bayesian neural networks. We replicate the experiment set-up in Yarin Gal (Gal and Ghahramani, 2016) and Hernandez-Lobato & Adams (Hernández-Lobato and Adams, 2015) and compare the RMSE and predictive log-likelihood of DropWeights (referred to as "MCDW" in the experiments) to that of Probabilistic Back-propagation (referred to as "PBP", (Hernández-Lobato and Adams, 2015)), "MCDO", (Gal and Ghahramani, 2016) and to a popular variational inference technique in Bayesian NNs (referred to as "VI", (Graves, 2011)). This experiment aims to assess the uncertainty quality, and the results are shown in Table 3.1. DropWeights significantly outperforms all other models in terms of RMSE and test log-likelihood on all datasets apart from 'Wine Quality Red' and 'Yacht Hydrodynamics', for which MCDO obtains better results. All experiments were averaged on 20 random splits of the dataset taken from the UCI machine learning repository.

	Avg. Test RMSE and Std. Errors(↓)				Avg. Test LL and Std. Errors(↑)			
Dataset	**MCDO**	**VI**	**PBP**	**MCDW**	**MCDO**	**VI**	**PBP**	**MCDW**
Boston Housing	2.97 ±0.85	4.32 ±0.29	3.01 ±0.18	2.81 ±0.16	-2.46 ±0.25	-2.90 ±0.07	-2.57 ±0.09	-2.51 ±0.08
Concrete Strength	5.23 ±0.53	7.19 ±0.12	5.67 ±0.09	4.88 ±0.55	-3.04 ±0.09	3.39 ±0.02	-3.16 ±0.02	-3.04 ±0.03
Energy Efficiency	1.66 ±0.19	2.65 ±0.08	1.80 ±0.05	0.61 ±0.02	-1.99 ±0.09	-2.39 ±0.03	-2.04 ±0.02	-1.34 ±0.01
Kin8nm	0.10 ±0.00	0.10 ±0.00	0.10 ±0.00	0.07 ±0.01	0.95 ±0.03	0.90 ±0.01	0.90 ±0.01	1.23 ±0.08
Naval Propulsion	0.01 ±0.00	0.01 ±0.00	0.01 ±0.00	0.01 ± 0.00	3.80 ±0.05	3.73 ±0.12	3.73 ±0.01	4.48 ±0.00
Power Plant	4.02 ±0.18	4.33 ±0.04	4.12 ±0.03	3.91 ± 0.03	-2.80 ±0.05	-2.89 ±0.0	-2.84 ±0.01	-2.78 ±0.01
Protein Structure	4.36 ±0.04	4.84 ±0.03	4.73 ±0.01	3.72 ±0.01	-2.89 ±0.01	-2.99 ±0.01	-2.97 ±0.00	-2.72 ±0.01
Wine Quality Red	0.62 ±0.04	0.65 ±0.01	0.64 ±0.01	0.63± 0.01	-0.93 ±0.06	-0.98 ±0.01	-0.97 ±0.01	-0.97 ±0.01
Yacht Hydrodynamics	1.11 ±0.38	6.89 ±0.67	1.02 ±0.05	1.42± 0.09	-1.55 ±0.12	-3.43 ±0.16	-1.63 ±0.02	-1.60 ±0.04

Table 3.1 Average test performance in RMSE and predictive log likelihood

We used DropWeights following the same way the method used in current research - without adapting model architecture. As a result, we expect the MC-DropWeights method to give better quality uncertainty estimates experimenting with different neural network architectures to that of specialised methods developed to capture uncertainty.

3.4.3 In Summary

Different approaches to Bayesian deep learning have a variety of advantages, independently of their accuracy and uncertainty quantification. The gold standard method for Bayesian neural networks is Hamilton Monte Carlo (HMC) (Neal, 1996) but at limited scalability on large data sets and task-specific hyperparameters such as the length and number of integration steps. One alternative to MCMC inference in neural networks is the Laplace approximation (MacKay, 1992b; Ritter et al., 2018) but it requires the computation of the inverse Hessian of the log-likelihood, which is not always feasible to compute for large networks.

A scalable Variational inference (VI) has received much attention both explicitly modelling parameters with distributions (Blundell et al., 2015; Graves, 2011; Hinton and Van Camp, 1993) and expectation propagation (Hernández-Lobato and Adams, 2015). MC Dropout (Gal and Ghahramani, 2016), MC-Batch Normalization (Teye et al., 2018) and other stochastic regularisation techniques (SRT) have all been proposed as alternatives and are scalable to large models. Deep ensembles method (Lakshminarayanan et al., 2017) as a Non-Bayesian alternative for uncertainty estimation, trains multiple independent networks and aggregates their predictions provides very good results but is limited by its computational cost. All of the above BNN methods require multiple forward passes to produce uncertainty estimates. Robust uncertainty estimates from existing methods in deep learning remains a challenge due to:

1. Implementation complexity in neural network architecture and choice of hyperparameters

2. Computational cost to converge than regular networks or train multiple networks. At test time, averaging the predictions from multiple models is often required.

3. Quality of quantified uncertainty depends on approximations to achieve scalability

Our objective is to define a framework for quantifying uncertainty in Deep Learning and evaluate the usefulness of predictive uncertainty in medical image analysis to avoid over-confident, incorrect predictions during decision making in computer-based medical systems rather than achieving state-of-the-art accuracy in Deep Learning and validate performance with respect to 'implementation complexity' or 'computational cost'.

3.5 Discussion

The purpose of this chapter has been to introduce deep ensemble neural networks models that are able to model or quantify uncertainty into their model prediction function approximation for a given input. However, providing posterior distributions over predictions is still an open problem. Moreover, there is an inherent challenge of ensuring that model uncertainty is appropriately captured so that we can trust a model when it is confident. We have developed a computationally tractable approximate inference technique for Bayesian neural networks, deep learning stochastic regularisation techniques (SRTs), DropWeights and proposed practical techniques to obtain model uncertainty, even from existing models. Empirically, we observed that the MC-DropWeights captures improved model uncertainty to its prediction and yields better prediction accuracy.

The two case studies show examples of how model uncertainty can be measured in regression and classification tasks. These examples highlight the need for trustworthy calibrated uncertainties. We have quantitatively compared the approximating distribution (corresponding to DropWeights) to two existing inference methods. We observed that deep ensemble neural networks with DropWeights can obtain an improved quality of quantified uncertainty in predictive log-likelihood and root mean square error (RMSE) techniques.

All BNN models must include appropriate priors over every parameter. Nevertheless, this is computationally expensive and often manifests itself in the form of an intractable integral, as found in Equation (3.11) for BNNs. All other models also come with their own limitations. It is still extremely challenging to ensure that large neural networks have correctly calibrated uncertainties. Therefore it is essential to explore what it means to measure quality uncertainty for applications that require Bayesian deep learning.

Chapter 4

Quantifying Uncertainty in Image Segmentation

> "Awareness of ignorance is the beginning of wisdom." Socrates

Image segmentation is the process of partitioning a digital image into a multiple set of coherent pixels into a meaningful subject. Segmented image assists doctors to diagnose and making decisions such as border detection, tumour detection/segmentation, and mass detection. Medical image segmentation of pathologies and anatomical structures is an inherently ambiguous task. The majority of current state-of-the-art methods do not account for such ambiguities. In this chapter, we demonstrate that the uncertainty estimates obtained from DropWeights using the Bayesian Residual U-Net (BRUNet) provide clinicians additional insight on semantic segmentation tasks with help from deep learners. I will demonstrate how the quantified model uncertainty can be used to choose which unlabelled image to annotate.

4.1 Medical Image Segmentation

Medical Image segmentation plays a vital role in image analysis to identify the pixels of organs, detection of boundaries within a 2D or 3D image, mass detection, or lesions from background medical images from many different modalities such as CT, X-ray, MRI, dermoscopy, microscopy, Endoscopy, OCT positron emission tomography (PET), single-photon emission computer tomography (SPECT), and many more.

In the segmentation process, we partition an image into multiple non-overlapping regions using a set of rules such as a set of similar pixels or intrinsic features such as colour, contrast

and texture. Segmentation reduces the area to be searched in an image by dividing the original image into two classes, such as object and background, in a meaningful form to be used and analysed. This meaningful form extracted using segmentation process involves shape, volume, relative position of organs, and abnormalities (Jungo et al., 2018a; Jungo and Reyes, 2019; Lê et al., 2016; Ma et al., 2018; Nair et al., 2020; Saad et al., 2010; Wang et al., 2019; Zhang and Ji, 2011).

Medical image segmentation of pathologies and anatomical structures is an inherently ambiguous task due to factors such as partial volumes and variations in anatomical definitions, poor contrast or various restrictions in the image acquisition, unclear structure borders and variations in annotation "styles" between different experts (Baumgartner et al., 2019; Warfield et al., 2002). The majority of current state-of-the-art deep learning methods provide point estimates of segmentation - meaning treat the problem as a one-to-one mapping from a single image to output mask per image and do not account for such ambiguities, limiting clinicians' ability to quantify the uncertainty of said segmentation for validation and interpretability. Estimating pixel-wise for the aleatoric (inherent) and epistemic (modelling) uncertainties from distributions over segmentations without sacrificing accuracy are therefore of substantial interest to the medical imaging community (Jensen et al., 2019; Wilson and Spann, 1988) to reduce misdiagnosis.

4.2 Nuclei of cells in Microscopy Image

Millions of people die due to diseases like cancer, heart disease, chronic obstructive pulmonary disease, Alzheimer's, and diabetes every year. Most of the human body's 30 trillion cells contain a nucleus full of DNA, the genetic code that programs each cell. Identifying the cells' nuclei is the starting point for most analyses - to identify each individual cell in a sample, and by measuring how cells react to various treatments, the researcher can understand the underlying biological processes at work.

It is very complex to detect and eventually cure even in primary stages due to its invasive nature. Hence, the only method to survive this disease completely is via forecasting by analysing the early mutation in cells of the patient biopsy. Cell Segmentation can be used to find the cell which has left its nuclei. Thus, automated nucleus detection enables speeding up research for almost every disease, from lung cancer and heart disease to rare disorders, faster cure, and a high survival rate - from rare disorders to the common cold. Manual cell counting is prone to error, yet an extremely tedious process that would greatly benefit from the accurate segmentation of cells.

Currently, 40% of deaths are caused by heart disease and cancer. In addition, 75% of rare diseases affect children, and 30% die before their fifth birthday. Automating the nucleus detection is important as it allows to locate cells in various conditions more efficiently. This means more time can be invested in developing and testing new research ideas, which leads to a shorter time to market for new drugs in the long term. Overall, it means patients can access the latest treatments as soon as possible, which helps improve their quality of life.

4.2.1 Segmentation of Microscopy Data for finding Nuclei

Microscopy is fundamental to medical imaging - counting cells, tracking moving populations, localising proteins, classifying phenotypes, or profiling treatments. In biological and biomedical applications, segmenting the nuclei of cells in two-dimensional light microscopy images of stained nuclei is often the first step in the quantitative analysis of imaging data. Most existing bio-image analysis tools identify nuclei using classical segmentation algorithms such as thresholding, watershed, or active contours fail to correctly segment due to the heterogeneity of biological samples or can be sensitive to technical settings. Also, these need to be configured for each experiment to account for different microscopy modalities, scales and experimental conditions, often requiring great expertise to select the algorithm that suits the problem and to adjust its parameters (Caicedo et al., 2019).

The recent success of Deep learning has shown great potential for various medical image segmentation problems. In medicine, for cell detection or localisation to be meaningful, tolerance must typically be much tighter (e.g.,$>$ 50% overlapping with the actual bounding box). Then, the predictive posterior distribution indicates a network with high or less confidence about its decision based on the input medical image. However, predictive uncertainty in deep learning actually results from three separate forms of uncertainty (Lewandowski, 2017):

1. Model uncertainty or epistemic uncertainty accounts for uncertainty in the model parameters due to the lack of training data. Epistemic uncertainty associated with the model reduces as the training data size increases.

2. Data uncertainty or aleatoric uncertainty accounts for noise inherent in the observations due to class overlap, label noise, homoscedastic and heteroscedastic noise, which cannot be reduced even if more data were to be collected unless it is possible to observe all explanatory variables with increased precision.

3. Distributional uncertainty happens when there is a mismatch between the training data and test data distributions.

Van Valen et al. developed DeepCell (Valen et al., 2016) - Convolutional Neural Network-based methods that treat segmentation of single cells in microscopy images as a pixel-wise classification problem, produce low-resolution segmentation masks and to solve the cell counting problem. Unfortunately, the model does not predict the uncertainty associated with the machine learning task. Uncertainties in medical diagnosis are so pervasive that deep learning for handling variability of the linear predictors is no longer sufficient.

4.3 Image Segmentation Dataset:

We have used the dataset provided in the Kaggle Data Science Bowl Challenge 2018 (Hamilton, 2018) to demonstrate the merit of our proposed method. It consists of microscopy images of a large number of segmented nuclei images. The images were acquired under various conditions and varied in cell type, magnification, and imaging modality (brightfield vs fluorescence). In CNN architecture, it is necessary to convert all images to the same size. Therefore, all images were cropped to a square-centre region and resized to 128 x 128 pixels to be standardised and uniform. This ensured the aspect ratio avoided distortion to speed up the process. There are 670 train samples and around 4000 test samples.

4.4 Uncertainty quantification in segmentation

Deep learning provides a framework for a powerful class of flexible, rich non-linear models for classification and prediction for scalable learning using stochastic approximations and typically generate predictors with a deterministic result. Bayesian inference provides a unified framework for model building, prediction, inference, and decision making and provides uncertainty and variability of outcomes via probability density over outcomes which is robust to overfitting. It is important to know how confident a model is for each prediction and what a model does not know, especially when making life-threatening diagnosis decisions in medical applications. We have effectively utilised the strength of each of these frameworks in our method. Currently, deep learning models provide normalised score vectors, which do not truly represent uncertainty in the parameters, inherent stochastic noise and model specification. For example, bayesian deep learning (MacKay, 1992b; Neal, 1993) approaches - Bayesian neural networks replace the weight parameters of deterministic networks with distributions over these parameters and, instead of optimising the network weights directly, averaged over all possible weights (referred to as marginalisation). In deep learning, we can model the predictive probabilities in the outcome class as a function of the mean connected

with the activation function for combined Epistemic and Aleatoric uncertainties without calculating variance estimates separately in a classification problem.

Given a dataset $X = \{x_1, x_2 \cdots, x_N\}$ and the corresponding labels $Y = \{y_1, y_2 \cdots, y_N\}$ where $X \in R^d$ be a d-dimensioned input vector and $Y \in \{1 \ldots \ldots C\}$ with $y_i \in \{1 \ldots \ldots C\}$, C class label, a set of independent and identically distributed (i.i.d.) training samples size $N\{x_i, y_i\}$ for $i = 1$ to N , the task is to find a function $f : X \rightarrow Y$ using some parameters θ as close as possible to the original function that has generated the outputs Y. Parameters, θ, are the weights of the neural net.

Gal and Ghahramani (Gal, 2016) proved that the dropout neural network is equivalent to a variational approximation of the posterior of the network's weights and presented a simple, practical method to estimate predictive uncertainty by training a dropout network and taking Monte Carlo samples of the prediction using dropout at test time.

Kendall and Gal (Kendall and Gal, 2017) derived a unified Bayesian deep learning framework for both classification and regression on pixel-based semantic segmentation, by decomposing uncertainty into aleatoric - modelled by placing a distribution over the output of the model - and epistemic uncertainty. It does this by placing a prior distribution over the model's parameters. The last layer in the network has extra nodes before activation, consisting of the mean and variance of logits. Disentangling these two sources of uncertainty can be useful to capture richer diversity of realistic segmentation.

During training, the variance estimate is sampled and added to the probability logits, which are used to calculate the training loss in the network (Gal and Ghahramani, 2016). In the above approach for estimating uncertainty, there are mainly two limitations (Kwon et al., 2018):

1. The predictive uncertainty captures the variance of predictive probabilities from sample T stochastic feedforward. Which essentially quantifies the uncertainty of the variability in the specification of the probability distribution of the linear predictor function instead of the predictive probabilities in the outcome class of the model.

2. The network produces two outputs, prediction probability logits and a variance estimate, which are computed per output pixel by explicitly modelling the variability of the last layer of neural network outputs. It requires additional parameters.

To address the above limitations, we introduce a predictive uncertainty estimator, which averages the standard deviations of the predictive probabilities as:

1. Aleatoric uncertainty (AU) or Data uncertainty accounts for inherent stochasticity in the data due to class overlap, label noise, homoscedastic and heteroscedastic noise,

always leading to predictions with high uncertainty. Aleatoric uncertainty cannot be reduced even if more data were to be collected unless it is possible to observe all explanatory variables with increased precision.

2. Epistemic uncertainty (EU), also known as Model uncertainty, is a consequence of insufficient learning of model parameters due to a finite set of training data, which leads to broad posteriors. It is impossible to determine a model's parameters exactly with limited observations. This uncertainty captures 'what the model does not know'. Epistemic uncertainty associated with the model reduces as the training data size increases. We can compute epistemic uncertainty as information available - information expressed.

$$\text{Epistemic uncertainty: } \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - \bar{y}_t)(\hat{y}_t - \bar{y}_t)^T \tag{4.1}$$

where $\bar{y}_t = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t$
In practice, the predictive probability is estimated as follows:

I Repeat the stochastic forward pass T times through the Neural Networks with Drop-weights

II For each stochastic forward pass, a different network is making predictions because DropWeights randomly switched off units.

III As a result, each stochastic forward pass returns different vectors of class predictions, which is equivalent to stochastic variational inference drawing new independent predictions.

IV Finally, average the predictions to get the final prediction as an uncertainty estimator associated with the sample in the prediction exercise.

The above method reduces the required hyperparameters and improves computation. In addition, it considers the model uncertainty associated with every class prediction.

4.5 Bayesian Residual U-Net (BRUNet)

To get a better result in semantic segmentation, it is crucial to use low-level details while retaining high-level semantic information (Olshausen and Field, 1996; Stiles and Jernigan, 2010). Therefore, we used deep Bayesian Residual U-Net (BRUNet) architecture to take

advantage of strengths from deep residual learning and U-Net architecture. The U-Net and residual networks have a simple structure and faster training speed. However, U-Net's accuracy of the experimental results, due to its lack of depth, is insufficient, and the residual network effectively addresses the problem of degeneration of deep convolutional neural network. Therefore, we have effectively combined the strengths of two networks in our artificial neural network to implement with Monte Carlo DropWeights layers, as shown in figure 4.1 below, to estimate model uncertainty.

We have designed BRUNet using a Convolution layer, an Activation layer, a Pooling layer and a Fully Connected Layer with a combination of max-pooling and batch normalisation. DropWeights are applied to the network as an approximation to the Gaussian Process (GP) and to cast as approximate Bayesian inference.

Deep learning models require to be initialised with the right weights to avoid vanishing / exploding gradients problems. "He" initialisation (Goodfellow et al., 2016) draws samples from a truncated normal distribution centred on 0 with stddev = sqrt(2 / fan-in) where fan-in is the number of input units in the weights, to asymptotically preserve variance of activations in the forward pass and variance of gradients in the backward pass. The Exponential Linear Unit (ELU) is a recently introduced activation function in Deep Learning. It computes the function $f(x) = x$ if x ≥ 0 (identity function) and $f(x) = \alpha \cdot (e^x - 1), \alpha$ is a positive constant number, if x < 0. ELU tends to converge mean activations closer to zero, causing faster learning, convergence and producing more accurate results. The last layer in the fully connected network holds the scores for each class from the sigmoid function.

The entire network is still in the form of a U-shaped structure, which involves down-sampling first, followed by upsampling; the down-sampled features are merged with the corresponding up-sampled features. Finally, the result is obtained through the fully connected layer. The intuition behind this is that extracting low-level features to correspondingly high levels creates a path for information propagation between low and high levels much easier. This facilitates backward propagation during training and compensates low-level finer details to high-level semantic features in dense prediction tasks. The contraction and expansion layers are convolutions and deconvolution layers. Hence the image is recreated with segmented masks, like the image input size.

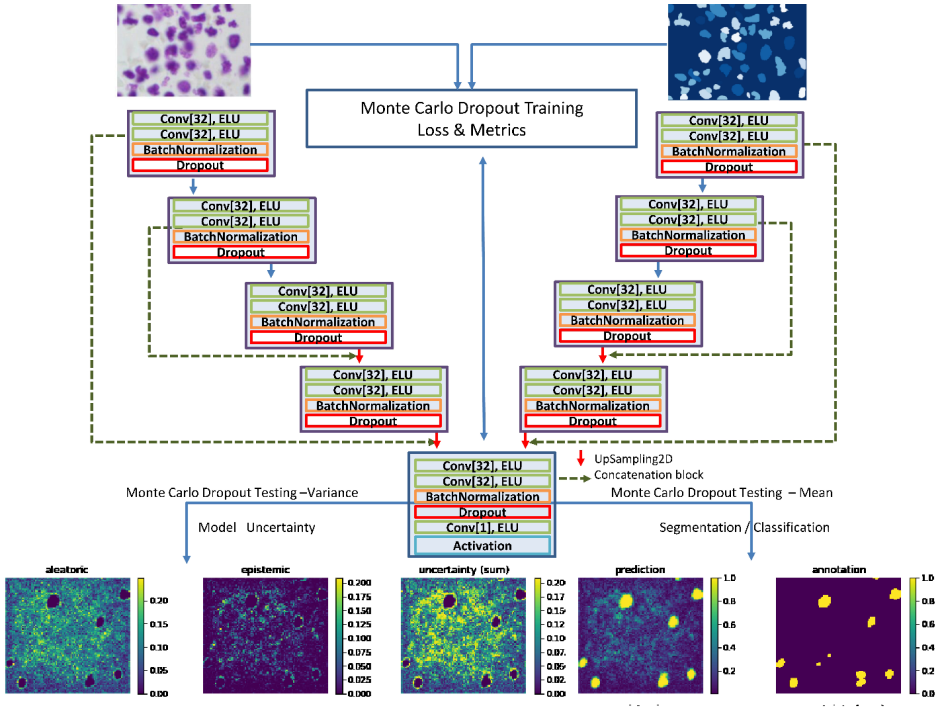


Fig. 4.1 Bayesian Residual U-Net (BRUNet) Architecture

The performance of our model is evaluated on the mean average precision at different Intersection over Union (IoU) thresholds and Dice similarity coefficient (DSC).

1. Intersection over Union (IoU), also known as the Jaccard index, is an evaluation metric for pixel-level image segmentation. The IoU is the percent overlap between the area of ground truth and the predicted area. The higher IOU to a certain threshold, the more accurate is the prediction.

2. The Dice similarity coefficient (DSC) is the most widely used measure of reproducibility as a validation of manual annotation where clinicians repeatedly annotated the same image and the pair-wise spatial overlap accuracy of automated probabilistic segmentation of images. It ranges between 0 and 1.

3. Model accuracy is used to judge the performance of the model and is similar to a loss function. The loss function is set to 1 - Dice coefficient loss between the predicted and true labels as follows:

$$\mathscr{L} = 1 - \frac{2\sum y_{true}\hat{y}_{pred}}{\sum y_{true} + \sum \hat{y}_{pred}} \tag{4.2}$$

We evaluated the validation accuracy after every epoch and saved the model with the best prediction accuracy (lowest loss) on the validation set (Badrinarayanan et al., 2017; Chollet et al., 2015).

4.5.1 BRUNet Parameters details

All models are trained and evaluated using Keras with Tensorflow backend. We used the following hyper-parameters. All Nuclei images were resized to the same dimension 128 x 128 x 3. The Bayesian Residual U-Net (BRUNet) was trained by Stochastic Gradient Descent (SGD), with weights initialised using the "He" activation. The SGD optimiser with the Dropout rate of 0.10, 0.20, 0.25, 0.50, 0.75 and 0.95 and early stopping rule with 25 epoch patience was applied, with a batch size of 16. The total parameters of the network were 4,452,097. The binary-cross entropy function was used as a loss function to calculate the validation loss of various models for comparison. The variational inference with DropWeights variational distribution was used. The number of realised sets T used in Monte Carlo integration were 10.

4.5.2 Experimental results

In the previous sections, we have discussed modelling different aspects of predictive uncertainty and presented measures of quantifying it. This section evaluates our method when applied to the problem of discovering nuclei in divergent microscopy data images. This application is receiving much attention from the deep learning community (Hamilton, 2018; Irshad et al., 2013).

Network Performance

We have observed high accuracy in the case of isolated Cells when compared with overlapping cells, as shown in Figures 4.2 and 4.3. The highest areas of aleatoric uncertainty occurred on class boundaries, and epistemic uncertainty increases for complex pixels.

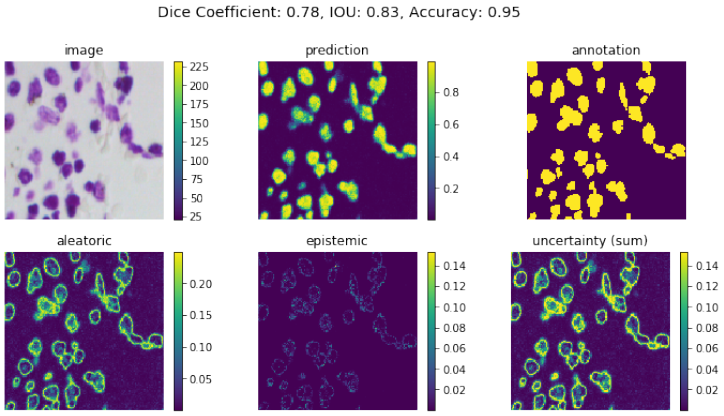


Fig. 4.2 The segmentation was performed by the model on images such as those shown above with overlapping cells with uncertainty

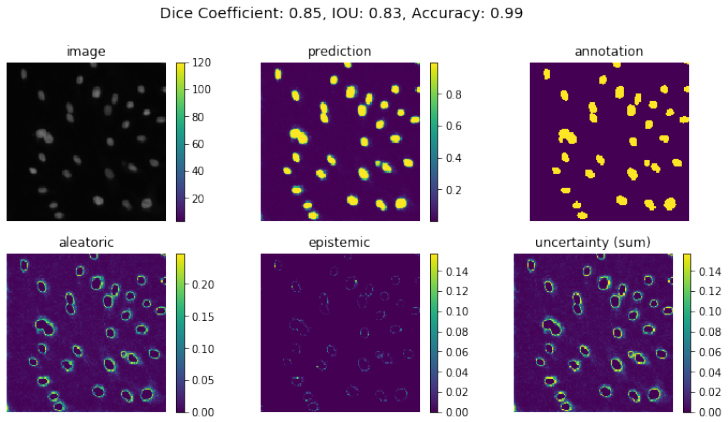


Fig. 4.3 The segmentation was performed by the model on the image with isolated cells with uncertainty

Using a simple CNN, the nuclei were spotted with model accuracy of 91% for stage 1 train results with a mean Intersection Over Union (mean IoU) score of 62%. Using the BRUNet model described above, we obtain a mean accuracy of 96.55% with a mean IoU of 83%.

Distribution of Uncertainty Estimates

The distribution of aleatoric uncertainty appears to be multi-modal, with peaks close to 0.13, as shown in below figure 4.4. The incorrect classifications greatly contribute to the multi-modality due to data's irreducible homoscedastic and heteroscedastic noise.

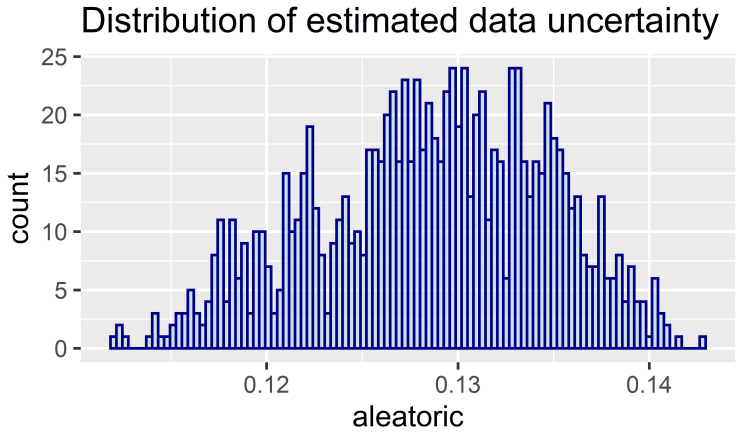


Fig. 4.4 Distribution of estimated Aleatoric uncertainty

The distribution of epistemic uncertainty appears to be multivariate normal. The incorrect predictions are centred around a higher uncertainty, whereas far more of the correctly

predicted classes are concentrated around a low uncertainty value, as shown in below figure 4.5.

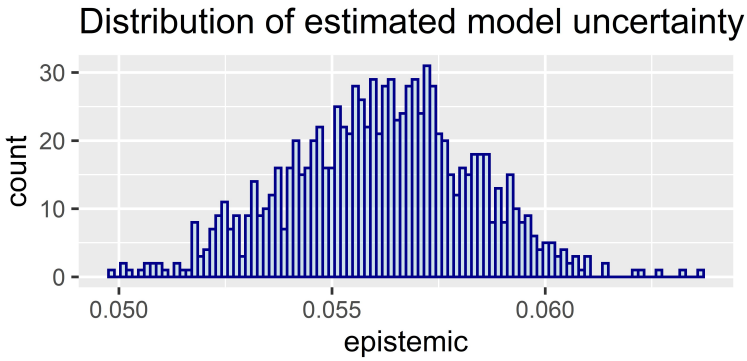


Fig. 4.5 Distribution of estimated Epistemic uncertainty

Correlation between Aleatoric Uncertainty and Epistemic Uncertainty with Predictive Probabilities

To study the correlation between model uncertainty and data uncertainty, we measured the estimated conditional expectations of the epistemic uncertainty and aleatoric uncertainty given ranges of the predictive probabilities, respectively as shown in figure 4.6.

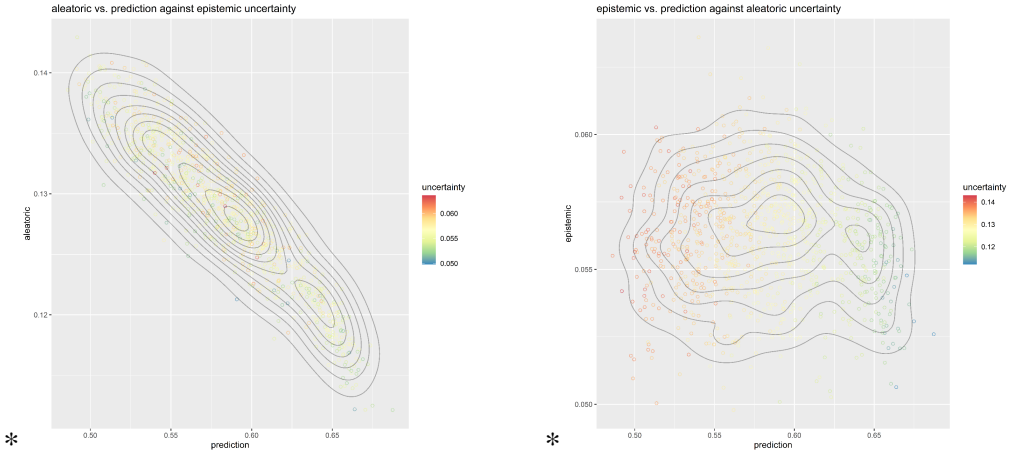


Fig. 4.6 The joint distribution of between Aleatoric uncertainty Epistemic uncertainty vs Prediction (2D Kernel Density Estimate)

As expected, uncertainty decreases as the predictive probabilities increase. A blue point corresponds to a prediction with a low value of uncertainty. A red point corresponds to observation with a high value of uncertainty. It confirms that for the higher uncertainty predominately due to incorrect classifications, most of the points are concentrated around the area of maximum entropy.

This correlation between aleatoric uncertainty and epistemic uncertainty with predictive probabilities indicates that the approximated uncertainty estimates indeed contain valuable information in incorrect cases.

The Effect of Varying Stochastic Feed Forwards

Dropweights randomly drop connections between the last two layers of neural networks, where the Bernoulli dropping is applied directly to each weight to regularise large neural network models. Each forward pass then generates a Monte Carlo sampling from different smaller sub-networks through the trained network with active Dropweights. Therefore, the model could predict different values each time for a given input. The primary goal of Monte Carlo (MC) Dropweights is to generate random predictions and interpret them as samples from a probabilistic distribution. Several of such forward passes are needed to approximate the posterior distribution of softmax class probabilities. Then, the mean of these samples can be interpreted as the network prediction.

In practice, MC-DropWeights is equivalent to performing T stochastic forward passes through the BRUNet and averaging the results. We have observed from the below figure 4.7 that the aleatoric uncertainty decreases with the increase in the number of stochastic forward pass (T), whereas the rate of change of range for the epistemic uncertainty is not very significant with increases in the number of stochastic feed forwards because the capacity of the models increases with the increase in the number of stochastic forward pass (T).

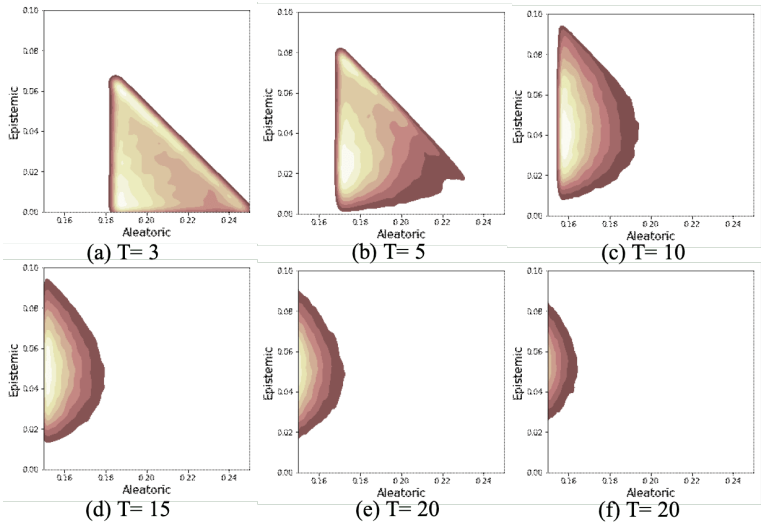


Fig. 4.7 Bivariate density plot for aleatoric and epistemic uncertainties against fixed Dropout with varied stochastic feed forwards of the model

The contribution of Uncertainty in Predictive Probabilities

The uncertainty adds complementary information to the conventional network output - for the correctly classified cases, the model uncertainty is low; however, the standard deviation of the predictive probabilities most likely class seems to be higher for the incorrectly classified images. When the prediction disagrees with the ground truth, the uncertainty identified the region missed by prediction as highlighted in figure 4.8.

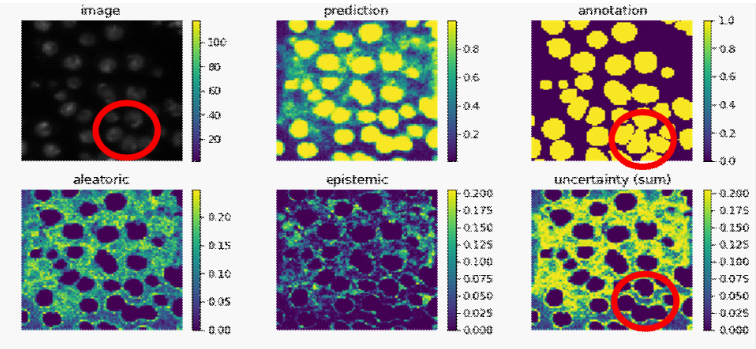


Fig. 4.8 Segmentation predictions and uncertainty maps

The model is generally highly certain or provides higher confidence in its prediction in the cases where it predicted the correct class. Uncertainty information means the model can be easily interpreted, compared to the case with no uncertainty information in Microscopy cell images of nuclei segmentation.

4.6 Application: Active Learning for Medical Image Segmentation

Previous sections discussed dropweights as stochastic regularisation techniques (SRTs) to approximate variational inference in Bayesian neural networks (NNs) to quantify model uncertainty. However, artificial Intelligence-based medical diagnosis requires a large number of many labelled images to obtain good performance. In the real world, unlabelled medical images are available in abundance; however, annotating the data with reliable class labels after careful inspection of numerous images can be very tedious, time-consuming, and expensive, as well as being subject to errors on the part of by the interpreter. Active learning is a mechanism that tries to minimise the amount of labelled data required to control the labelling process. Thus, developing active learning algorithms to learn from a small sample, high-dimensional labelled images, querying the highly informative unlabelled images, and minimising redundant examples with limited resources is of paramount practical importance.

This section will see how model uncertainty can be used in active learning to annotate unlabelled images in semantic image segmentation for medical image analysis.

There are many heuristic methods and numerous query strategies in active learning for medical image classification using traditional machine learning (Aggarwal et al., 2014; Gal et al., 2017b; Houlsby et al., 2011; Ren et al., 2020; Settles, 2009; Wang and Yeung, 2016; Wang et al., 2016). However, in deep active learning, the uncertainty based acquisition function is heavily influenced by an average of the softmax probability values and miscalibrated due to the diverse nature of the medical image samples, disease conditions and sampling bias.

We designed a novel Active Learning sample selection strategy for high dimensional image data to measure the confidence of the model uncertainty in classification and unbiased calibrated uncertainty weighted by the Euclidean Distance Transform (EDT) of the prediction for semantic image segmentation (Ghoshal et al., 2021b). A sample is selected based on the lowest uncertainty confidence score for labelling, are with highly informative and little redundancy. Using this metric can significantly reduce the number of labelled samples required compared to other selection strategies for achieving higher accuracy.

Figure 4.9 describes the Active Learning framework, which effectively trains a Bayesian residual U-Net (BRUNet) for medical image segmentation with limited labelled data during training to estimate uncertainty for each unlabelled image fed into the trained network. At each round of active learning, the algorithm computes a bias-corrected confidence score of uncertainty for all images in the unlabelled pool. Image(s) with the least score value of uncertainty confidence is selected for the clinician to label, and then the corresponding image(s) are added to the training set in the next round of the model training. Our method relies on the bias-corrected confidence score of uncertainty sampling, in which the algorithm selects the unlabelled image(s) that it finds manually hardest to annotate.

Active learning depends on the ability to select the right sample to be annotated to improve model performance and decrease model uncertainty. Therefore, defining the acquisition function is a real challenge. For example, the most popular Dropout Bayesian Active Learning by Disagreemement (BALD) (Gal and Ghahramani, 2016) maximise the mutual information between predictions and model posterior (Houlsby et al., 2011). However, it double counts Mutual Information (MI) between data points and overestimates the true MI. Estimation of entropy from the finite set of data suffers from a severe downward bias when the data is under-sampled.

The uncertainty obtained by Bayesian Neural Network (Ghoshal et al., 2019a) is prone to miscalibration, i.e. for the perfectly segmented image could have higher uncertainty. The proposed method is presented threefold: First, uncertainty obtained by dropweights

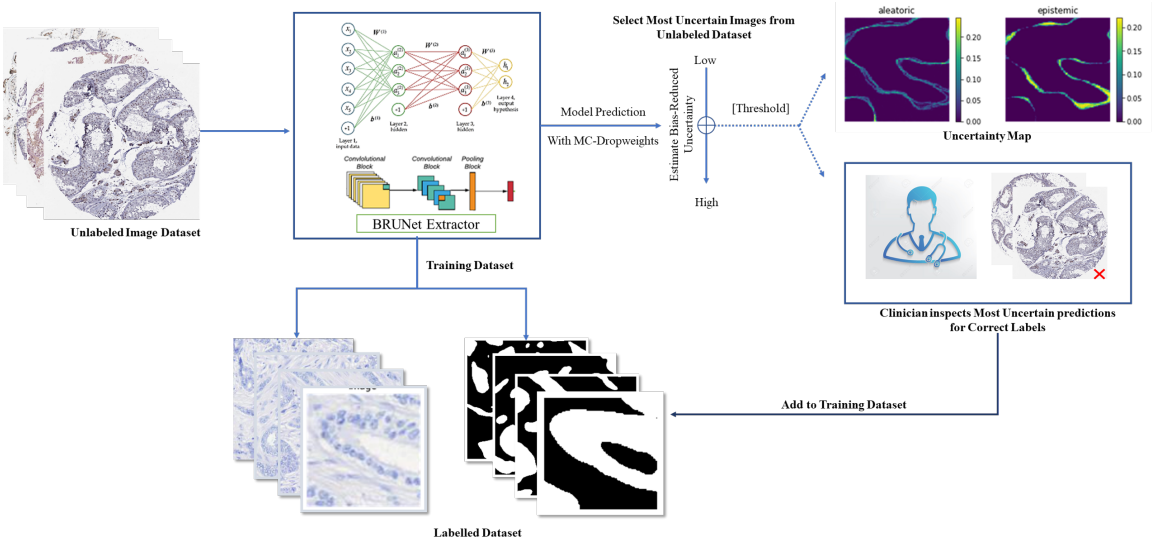


Fig. 4.9 Active Learning framework

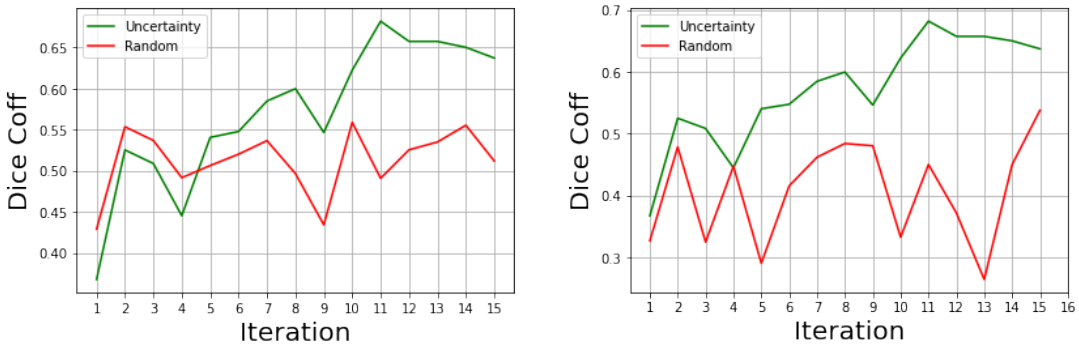
variational inference; second, pixel-wise estimated uncertainty is weighted by the Euclidean Distance Transform (EDT) (Di Scandalea et al., 2019; Ma et al., 2020) to standardise the importance of the pixel and so reduce overconfident prediction errors in dense pixels regions; and finally, calibrated uncertainty is compared with by random sample selection.

In section 4.5, I demonstrated that the estimated aleatoric and epistemic uncertainty obtained from dropweights using the Bayesian residual U-Net (BRUNet) provide additional insight for clinicians with help from deep learners (Ghoshal et al., 2019a). The value of each pixel represents the variance computed on MC samples.

The network was trained with the Dice loss function to highlight contrasting areas of the image and weighted uncertainty by distance transformation normalisation to address unreliable uncertainty mainly on class boundaries. Fig. 4.12 and Fig. 4.13 illustrate the uncertainty maps evolution over active learning iterations for the Epithelial Cells in Breast Cancers and Retina Fundus images semantic segmentation.

Figure 4.10 illustrates the segmentation performances evaluated on the test image dataset across the 15 experiments. At baseline, I compared the uncertainty-based method with random selection. The result shows a clear improvement of the proposed uncertainty method compared to the random selection. It is observed that as the number of images in the training set grows, active learning through uncertainty dominates random selection. The segmentation performances gain is noticeable after adding only five uncertainty-selected samples over the randomly-selected samples. In addition, the gap between the two Dice curves is progressively increasing as more samples are included in the training set. This is also consistent with the deep learning sensibility to the training dataset size and quality of a dataset. Therefore,

the Dice curves were expected to be sufficiently stable when running the active learning simulation multiple iterations. We also noticed that the selected samples were not always unique samples in the uncertainty-based method in all iterations, but the number of unique samples selected was much higher than random selection.



(a) Breast Cancer (b) DRIVE Dataset

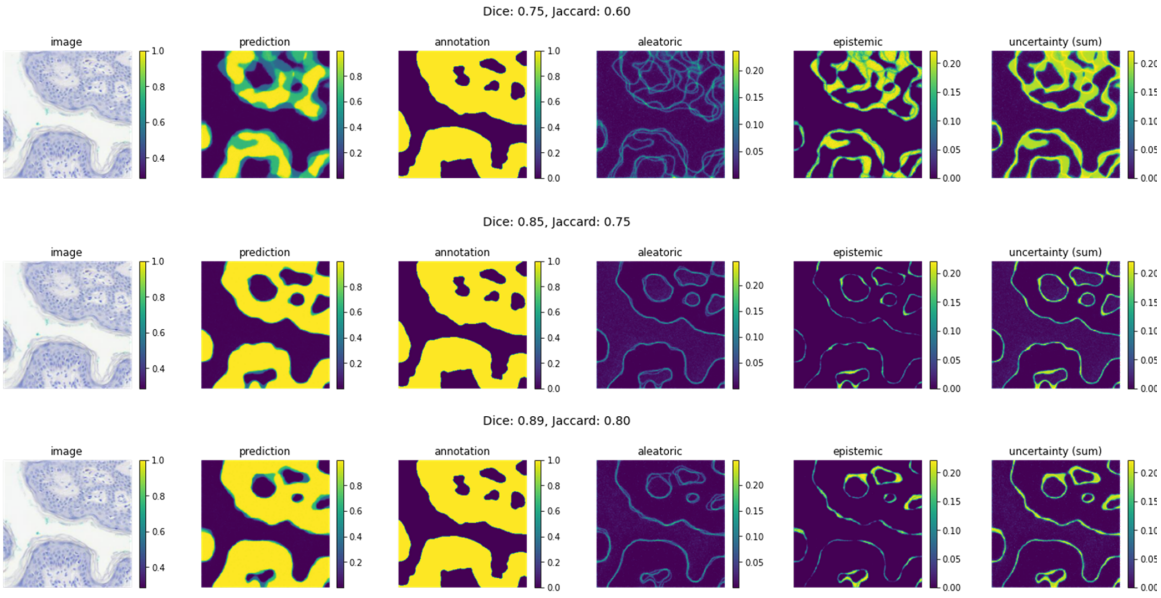
Fig. 4.10 Active Learning Performance

A. Cytokeratin-Supervised Epithelial Cells in Breast Cancers Semantic Segmentation:

Immunohistochemistry (IHC) staining's of oestrogen receptor (ER), progesterone receptor (PR), and proliferation antigen Ki-67 are routinely used for automated epithelial cell detection in breast cancer diagnostics. However, manual annotating complex IHC images to determine the proportion of non-malignant stromal or inflammatory cells in stained cells is extremely tedious and expensive and may lead to errors or inter-observer variability. Dataset included images from 152 patient samples stained with fluoro-chromogenic cytokeratin-Ki-67 double staining and sequential hematoxylin-IHC (Valkonen et al., 2019).

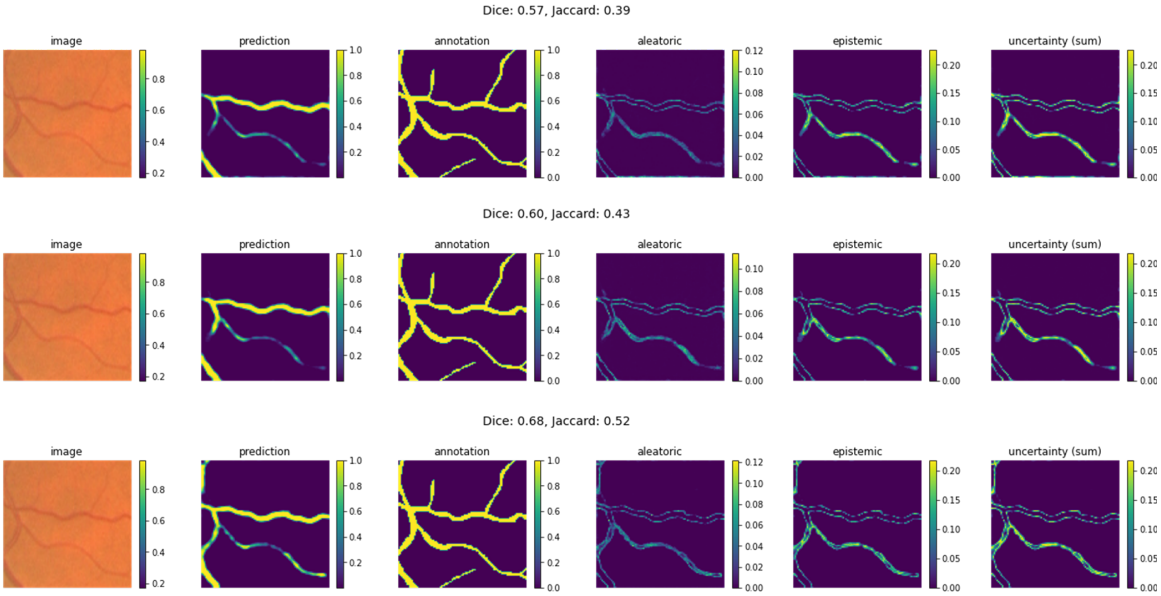
B. Digital Retinal Images for blood Vessel Extraction (DRIVE) Semantic Segmentation:

Diabetic retinopathy (DR) is one of the reasons for vision loss in diabetic patients due to retinal blood vessels damage. So, automatic detection and segmentation of retina fundus images are essential to prevent vision loss in diabetic patients. The DRIVE database contains 40 images using 400 diabetic patients (seven of them have various pathological cases) (Staal et al., 2004). All images also have corresponding manually segmented masks. The images have been divided into training and test set. Each part contains 20 images. We use testing images for performance evaluation. All images were resized to 96 x 96 pixels.



(a) Fluoro-chromogenic cytokeratin-Ki-67 Breast Cancers - AL iterations (3, 7 and 14)

Fig. 4.11 Prediction with estimated aleatoric and epistemic uncertainty maps (Ghoshal et al., 2019a). We observe that the Dice coefficient increases as active learning iterations progress with more training images.



(a) Retina Fundus images - AL iterations (1, 5 and 10).

Fig. 4.12 Prediction with estimated aleatoric and epistemic uncertainty maps (Ghoshal et al., 2019a). We observe that less variance is on thin boundary pixels, and the model seems to be more confident where it can distinguish line shapes vs round shapes.

4.7 Discussion and Perspectives

Deep learning-based medical diagnosis has to express the uncertainty of an image in the same way as a doctor may express ambiguity and ask for expert advice. We present the first approach (to the best of our knowledge) of Monte-Carlo DropWeights and BRUNet to model Bayesian neural networks as a reliable, variational inference method, which accurately estimates the models' aleatoric and epistemic uncertainties.

These techniques are simple to implement, especially in computer vision settings. We have observed constant aleatoric uncertainties per data set, regardless of the model because it is purely dependent on the data. Aleatoric uncertainties stems from multiple sources such as missing information, labeling noise (for example: human disagreement), measurement noise (for example: imprecise tools), missing data (for example: partially observed features, unobserved confounding variables). Another interesting observation is that epistemic uncertainty decreases with increasing accuracy - exactly how it should be because certainty increases with the model predictions of correct class labels.

Furthermore, we have demonstrated that medical image segmentation with uncertainty information provides additional insights into the corresponding analysis alongside point estimation, which can increase its ability to interpret the data and improve confidence, making models based on deep learning more applicable in a medical setting.

At present, the state-of-the-art medical image segmentation algorithm requires a large number of labelled images. However, such labelled images are costly to acquire in time, labour, and human expertise. Bayesian Active Learning approach using quantified uncertainty from deep neural networks with Dropweights, specifically selecting a couple of highly informative samples with very few annotated samples in a practical way, will benefit clinicians to obtain fast and accurate unlabelled image annotation with confidence. Furthermore, the heatmaps of aleatoric and epistemic uncertainty in semantic segmentation along with prediction would help clinicians better understand spatial relations in images and where the deep learning model tends to fail the most. Future research includes an evaluation in which all samples are learned and when the model reaches the optimal learning performance level.

Finally, in this chapter, we designed Monte-Carlo Dropweights as a Bayesian approximation to represent model uncertainty specifically for the task of semantic segmentation and applied it in an active learning framework for medical image segmentation for a real-world scenario. This idea was further extended to correct bias in estimated uncertainty, developed metrics to evaluate Bayesian models, and its application in Multi-Class imaging for disease detection is discussed in the next chapter.

Chapter 5

Quantifying Uncertainty in Image Classification

> "The more you know, the more you know you don't know." Aristotle

Previous chapters discussed neural networks with DropWeights to approximate inference in Bayesian neural networks (NNs) to quantify model uncertainty. However, the estimated entropy from the finite set of data suffers from a severe downward bias when the data is under-sampled. This chapter leveraged the Jackknife resampling method to correct the bias. We have applied deep ensemble neural networks with the MC-DropWeights method using the bias-corrected estimator in multi-class diseases detection. However, multi-label classification is more valuable in practical applications. We have applied our framework in recognising proteins expressed in cell types in testes based on immunohistochemically (IHC) stained images using a grid search scheme based on Matthews correlation coefficients. We next assess the quality of quantified uncertainty obtained from various approximating distributions on the task of classification.

5.1 Estimating Bias-Corrected Uncertainty using Jackknife Resampling Method

Entropy is the basic principle of information theory proposed by Shannon(Shannon, 1948). It depends on sample size and typically exhibits substantial bias. The model output in classification is a conditional probability distribution $P(y|x)$ over a discrete set of outcomes Y.

Many measures estimate uncertainty, such as softmax variance, expected entropy, mutual information, predictive entropy, and averaging predictions over multiple models. In supervised learning, information gain, i.e. mutual information (MI) between the input data and the model parameters, is considered as the most relevant measure of the epistemic uncertainty (Depeweg et al., 2017; Shannon, 1948). However, it double counts mutual information between data points and overestimates the actual MI (Houlsby, 2014). Estimation of entropy from the finite set of data suffers from a severe downward bias when the data is under-sampled. Even small biases can result in significant inaccuracies when estimating entropy (Macke et al., 2013). We leveraged the Jackknife resampling method to calculate bias-corrected entropy (Quenouille, 1956).

We have analysed two approaches to estimate uncertainty within classification: tractable view of the Mutual Information (Houlsby et al., 2011; Shannon, 1948), and Bias-Corrected Mutual Information(Quenouille, 1956). The Mutual Information (MI) between the prediction y and the posterior over the model parameters w capture model confidence.

Given a set of training dataset $X = \{x_1, x_2 \cdots, x_N\}$ and the corresponding labels $Y = \{y_1, y_2 \cdots, y_N\}$ where $X \in R^d$ be a d-dimensioned input vector and $Y \in \{1 \ldots \ \ldots C\}$ with $y_i \in \{1 \ldots \ \ldots C\}$, C class label, a set of independent and identically distributed (i.i.d.) training samples size $N\{x_i, y_i\}$ for $i = 1$ to N, a BNN is defined in terms of a prior $p(w)$ on the weights, as well as the likelihood $p(D|w)$. Consider class probabilities $p(y_{x_i} = c \mid x_i, w_t, D)$ with $w_t \sim q(w \mid D)$ with $W = (w_t)_{t=1}^T$, a set of independent and identically distributed (i.i.d.) samples, drawn from $q(w \mid D)$ where T is the number of variational samples. The entropy of the predictive distribution (H) can be defined as:

$$H[P(y^*|x,D)] = -\sum_{y^* \in \{1 \ldots M\}} P(y^*|x,D) \log P(y^*|x,D) \tag{5.1}$$

The mutual information (MI) measures the information gain about the model parameters w can be defined as the difference between entropy of the predictive distribution and the mean entropy of predictions across multiple stochastic samples:

$$I(w, y^*|x,D) \quad \simeq \quad H[P(y^*|x,D)] - \tfrac{1}{T} \sum_{t=1}^{T} H[P(y^*|x,\hat{}_t)]. \tag{5.2}$$

The first term in the MC estimate of the mutual information is called as the plug-in estimator of the entropy:

$$\hat{H} = H(\hat{p}) = -\sum_{i=1}^{C} y_{x,i} \log(p_{x,i}) \tag{5.3}$$

, where $\hat{p}_{x,i} = \frac{1}{T}\sum_{t=1}^{T}\hat{p}_{x,i}^{(t)}$ are the maximum likelihood estimates of each probability $\hat{p}_{x,i}$. It has long been known that the plug-in estimator underestimates the true entropy and plug-in estimate is biased (Basharin, 1959; Harris, 1975).

A classic method for bias correction is the Jackknife resampling method (DasGupta, 2008). In order to alleviate the bias problem, we propose a Jackknife estimator to estimate epistemic uncertainty to improve the entropy based estimation model. Unlike MC-Dropout, it does not assume constant variance. If $D(X,Y)$ is the observed random sample, the ith Jackknife sample, x_i is the subset of the sample that is a "leaves-one-out" observation $x_i : x_{(i)} = (x_1, \ldots x_{i-1}, x_{i+1} \ldots x_n)$. For sample size N, the Jackknife standard error $\hat{\sigma}$ is defined as: $\sqrt{\frac{(N-1)}{N}\sum_{i=1}^{N}(\hat{\sigma}_i - \hat{\sigma}_{(\odot)})^2}$, where $\hat{\sigma}_{(\odot)}$ is the empirical average of the Jackknife replicates: $\frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_{(i)}$. Here, the Jackknife estimator is an unbiased estimator of the variance of the sample mean.

The Jackknife correction of a plug-in estimator $H(\cdot)$ is computed as (DasGupta, 2008):

1. Given a sample $(p_i)_{i=1}^{N}$ with p_i discrete distribution on multi-class classification $1...C$
2. for each $i = 1...N$

   compute the leave-one-out estimator: $\hat{p}_c^{-i} = \frac{1}{N-1}\sum_{j\neq i} p_{jk}$

   compute the Jackknife estimator of entropy: $\hat{H}_{-i} = H(\hat{p}^{-i})$
3. then compute the bias-corrected entropy estimator $\hat{H}_{jk} = N\hat{H} + \frac{(N-1)}{N}\sum_{i=1}^{N}\hat{H}_{(-i)}$, where $\hat{H}_{(-i)}$ is the observed entropy based on a sub-sample in which the ith individual is removed.

We leveraged the following relation:

$$\mu_{-i} = \frac{1}{N-1}\sum_{j\neq i} x_j = \frac{N}{N-1}\mu - \frac{1}{N-1}x_i = \mu + \frac{\mu - x_i}{N-1}\,.$$

while resolving the i-th data point out of the sample mean $\mu = \frac{1}{N}\sum_i x_i$ and recompute the mean μ_{-i}. This makes it possible to compute leave-one-out estimators of discrete probability distribution quickly.

The above method was simple to implement and computationally cheaper than the other resampling methods such as Bootstrap. It derives an estimate of the finite sample bias from the leave-one-out estimators of the entropy and reduces bias considerably down to $O(n^{-2})$ (DasGupta, 2008).

The bias-corrected epistemic uncertainty estimation model explains regions of ambiguous data space that are hard to classify as data distribution due to noise in the inputs, or the model

was trained with different domain data. Consequently, these inputs should be assigned a higher aleatoric uncertainty. As a result, we can expect a high model uncertainty in these regions. A summary of these steps is provided in Algorithm 2.

Algorithm 2: Estimating Bias-Corrected Uncertainty with MC-DropWeights

Input: Dataset: $D = \{(\hat{x})\}$;;
Model f with optimized parameters θ;
Initialization: Dropweights rate r; Number of Inferences (/stochastic forward pass) T
Result: Mean prediction $\hat{y}$; Uncertainty σ
Reference Gal (2016) [Gal, Y. 2016 (eq. (6.3) p.109, Prop. 4 p.149)]
$p = \{\}$
for $t \leftarrow 1$ ***to*** T **do**
  # Neural network with Dropweights rate r performs stochastic variational inference
  # Independently drawn a set of weights vector $(\hat{w}_t)_{t=1}^{T}$ from $q_{\hat{\theta}}(w)$
  $\hat{p}_t = p \cup f^{\theta^t}(\hat{x}_t, r)$
Compute predictive probability: $\hat{p} = \frac{1}{T}\sum_{t=1}^{T}\hat{p}_t$
Compute prediction: $\hat{y} = argmax(\hat{p})$
for $t \leftarrow 1$ ***to*** T **do**
  Compute the leave-one-out estimator: $\hat{p}_c^{-t} = \frac{1}{T-1}\sum_{j \neq t} p_{jc}$
  Compute the plug-in entropy estimator: $\hat{H}_{-t} = -\sum_c \hat{p}_c^{-t} log(\hat{p}_c^{-t})$
Compute the bias-corrected uncertainty: $\sigma = \hat{H} + (T-1)(\hat{H} - \frac{1}{T}\sum_t \hat{H}^{-t})$

5.2 Estimating Uncertainty in Multi-Class Classification

5.2.1 Multi-Class Image Classification Dataset:

Magnetic Resonance Imaging (MRI) is the most commonly used medical imaging technique that provides informative data for diseases such as brain tumour diagnosis. However, the interpretation of medical images, including diagnosing the tumours from the MRI images that are an integral part of medical diagnosis, requires an experienced radiologist, a human whose skills are scarce and susceptible to mistakes. Therefore, we can use Artificial Intelligence (AI), the development of deep learning techniques and simple features from the images, such as intensity, contours and shapes, as means of computer-based assisting (classification and prediction) in medical diagnostic imaging. Here, deep learning-based solutions for detecting disease have been proposed with quantifying uncertainty in a decision, e.g. image-based (aleatoric) and model (epistemic) uncertainties.

To validate the effectiveness of our framework for multi-class medical image classification, we performed experiments on brain MRI scan images of 3 brain tumour types (Astrocytoma, Glioblastoma, Oligodendroglioma) with additional two categories (Healthy brain MRI and Unidentified tumour).

The MRI images, which include Astrocytoma, Glioblastoma, Oligodendroglioma and unidentified tumours, were obtained from the Repository of Molecular Brain Neoplasia Data (REMBRANDT) from Cancer Imaging Archive (Clark et al., 2013). This dataset has 65427 MRI images in DICOM format (the standard format of MRI images) categorised according to the 100 patient IDs. The images were converted into standard image formats like JPEG and categorised according to the tumour types with the help of clinical metadata. The MRI images of healthy brain images were obtained from the Brain Images of Normal Subjects (BRAINS) Image Bank repository of University of Edinburgh (Dickie et al., 2016) and Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD), a dataset used in research related to Alzheimer's disease(AD) (Malone et al., 2013). The MIRIAD dataset contains MRI images of the healthy brain ad AD group. A single pickle file was created with these images along with their labels for quick access and computation (Balasooriya and Nawarathna, 2017). The complete dataset with the number of images in each category is listed in Table I below.

This dataset contains 3,064 MRI images of 233 patients, containing 708 meningiomas, 1426 gliomas, and 930 pituitary tumours diagnosed with one of those above three brain tumour types. The most important property of this data set is that it includes both the brain images and the segmented tumours.

Data source	**Tumor type**	**No. of Images**
REMBRANDT	Astrocytoma	21307
REMBRANDT	Glioblastoma	17983
REMBRANDT	Oligodendroglioma	12460
REMBRANDT	Unidentified	13677
MIRIAD	Healthy brain	30688
BRAINS	Healthy brain	556
Total	-	96115

Table 5.1 The brain MRI dataset

The classes in our dataset are not balanced. Class imbalance is one of the most common problems in the real-world classification task. We have split the dataset into training (80%)

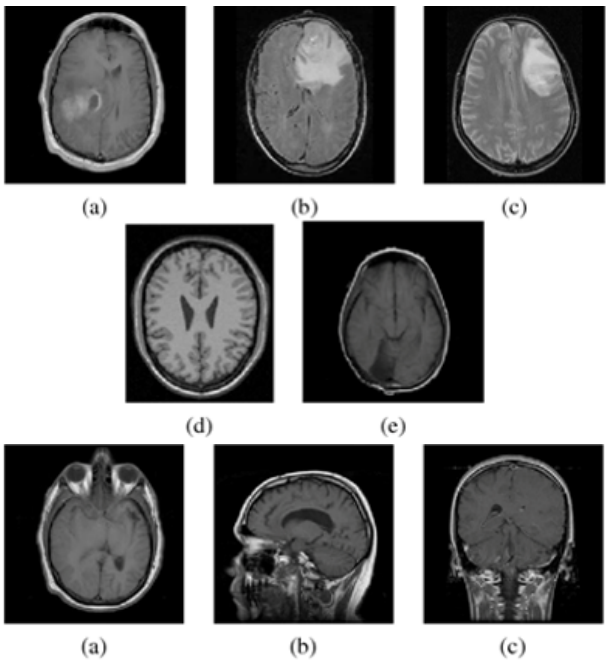


Fig. 5.1 [A] Types of brain tumors used. (a) Astrocytoma, (b) Glioblastoma Multiforme, (c) Oligodendroglioma, (d) Healthy tissue and (e) Unknown Tumor and [B] Image Planes of a brain MRI. (a) Axial Plane, (b) Sagittal Plane and (c) Coronal Plane

and testing (20%) before sampling so that data points won't be shared among the training and test dataset. We took the same number of samples from all classes to create a balanced dataset to train our models. Finally, we compared the performances between the two types of datasets, balanced, which contains 20% per class of brain tumour types and imbalanced, which have data distribution based on the abundance of classes of brain tumours in the image dataset: 22%, 19%, 13%, 32%, 14%.

5.2.2 Experiments

Our objective was to define a framework for measuring uncertainty in deep learning models and evaluate its usefulness. It was not, however, to achieve the state-of-the-art performance in deep learning, so for the DNN architecture, we used a generic building block containing the following model structure: Conv-Relu-BatchNorm-MaxPool-Conv-Relu-BatchNorm-MaxPool-Dense-Relu-[Dropout or DropWeights]- Dense-Relu-[Dropout or DropWeights]-Dense-Softmax, with 32 convolution kernels, 3x3 kernel size, 2x2 pooling, dense layer with 64 units, 32 units, and DropWeights rate probabilities ranging from 0.1 to 1.0, increasing by 0.05 to obtain models for uncertainty. One of the most useful and more robust optimisers

is Adam. Essentially Adam optimisation algorithm is an extension to stochastic gradient descent (SGD) and computes individual adaptive learning rates for different parameters. It combines the advantages of two stochastic gradient descent (SGD) extensions - Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp). We trained the network to minimise the cross-entropy loss using ADAM optimiser, which had an initial learning rate of 0.0001. The batch size was set to 32, and training was performed for a maximum of 100 epochs. After every epoch, we evaluated the validation accuracy and saved the model with the best prediction accuracy on the validation set.

All models were trained and evaluated using Keras with Tensorflow backend. Each image had three colour channels. The images are resized to 64 x 64 pixels for faster feature extraction.

5.2.3 Results and Discussion

This section proposes uncertainty estimation performance metrics in classification that incorporate the ground-truth label, model prediction, and uncertainty threshold. In addition, it analyses how the model uncertainty can help rank the model predictions by referring to uncertain MRI images of brain tumours. This will improve the overall model performance and improve clinical diagnosis.

We also compared the uncertainty-based classification performance obtained through our proposed method, using the state-of-the-art method, MC-Dropout, on balanced and imbalanced datasets, which showed considerable improvement in prediction accuracy and quality of uncertainty estimation.

Uncertainty Estimation Performance Metrics in Classification

There is no ground truth for uncertainty threshold or tolerance for evaluation of estimated uncertainty in deep learning. We leveraged the estimated uncertainty to enhance classification performance metrics (Mobiny et al., 2019). We first computed the accuracy map using the ground truth labels, model predictions and confidence map, by normalising uncertainty threshold values to develop the evaluation matrix.

Like in real-world referral situations, any medical diagnostic deep learning algorithm should be able to flag the least confident images that require more investigation by medical experts. Although the model does not necessarily require being confident for correctly predicting cases, it is expected that the estimated uncertainty will be high for incorrect predictions.

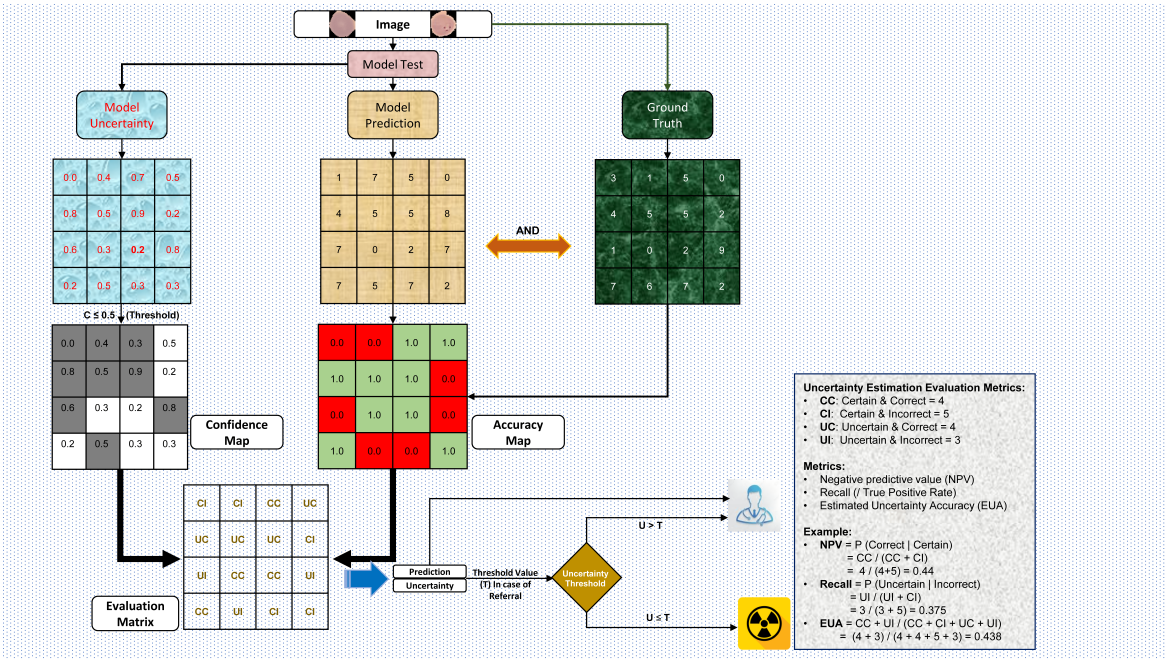


Fig. 5.2 Overview to evaluate the uncertainty quality metrics in classification task in disease detection

The evaluation matrix itself is not an estimated uncertainty performance measure. However, based on that, we can measure Uncertainty Accuracy (UA) Recall and negative predictive value (NPV), as shown in Figure 5.2, incorporating the ground-truth label, model prediction, and uncertainty value to evaluate the performance for uncertainty-aware classification.

Figure 5.2 shows the processing steps to compute the evaluation metrics using the ground truth labels, model predictions and confidence map by normalising uncertainty threshold values to develop the evaluation matrix.

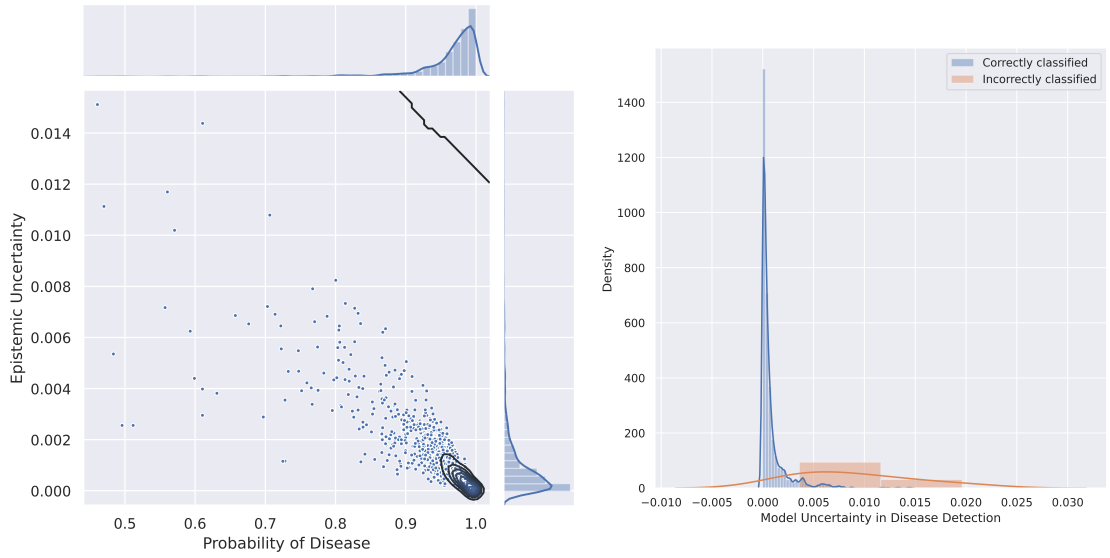
We first compute the AND operation of the ground truth labels and model predictions to compute the accuracy map. Then, we apply a threshold on the estimated uncertainty from model predictions to select the images that correspond to certain and uncertain groups for the confidence map.

In uncertainty-aware classification, we have four scenarios which are certain-correct (cc), certain-incorrect (ci), uncertain-correct (uc), and uncertain-incorrect (ui), to compute metrics (refer to Figure 5.2). The computation of negative predictive value (NPV), Recall and Uncertainty Accuracy (UA) depend on the uncertainty threshold setup. Higher the values, the model that performs better for all the proposed metrics.

Figures 5.4 and 5.5 plot each metric w.r.t various uncertainty thresholds and compare Bayesian Deep Ensembles of MC-DropWeights with the Ensembles MC-Dropout, MC-Dropout and MC-DropWeights using Mutual Information and Bias-Corrected Uncertainty Estimator(BCEU) using the area under each curve (AUC) metric.

Detecting Infected Patients with Confidence

Importantly, epistemic uncertainty as quantified by bias-corrected mutual information adds complementary information to the deep learning output. We observed with high probabilities that a diseased image is confined to lower epistemic uncertainties, as shown in figure 5.3 (a). In contrast, the uncertainty variation as seen on the scatter plot has a broader spread for healthy images.



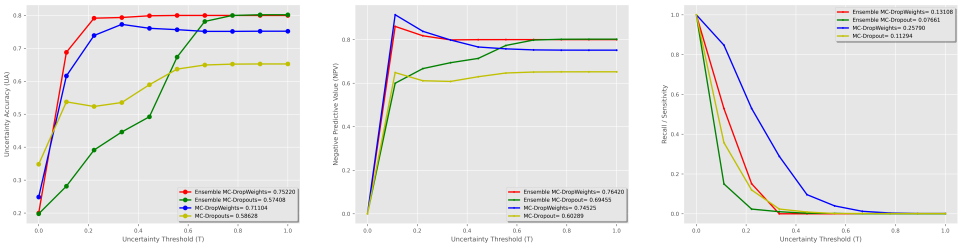
(a) a (b) b

Fig. 5.3 (A): The scatter plot between predictions and uncertainty. It shows that data with inherent noises might cause prediction errors. (B): Illustrating the distributions of model uncertainty values are plotted separately for correct and incorrect predictions

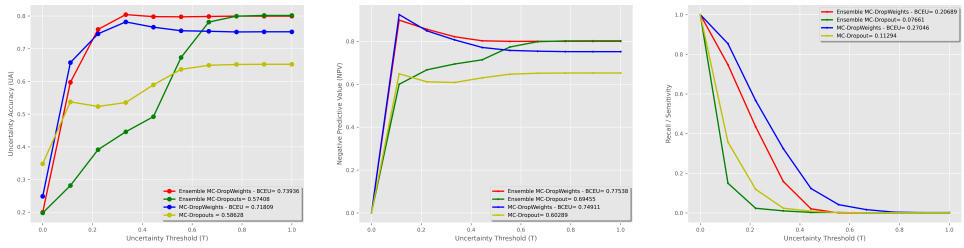
Figure 5.3 (b) shows the distribution of bias-corrected uncertainty values grouped by correct and incorrect predictions for test images. Given that a prediction is correct, there is a strong likelihood that the prediction uncertainty is also low. As a result, our model can confidently identify incorrectly classified images.

Uncertainty-based Classification Performance Comparison

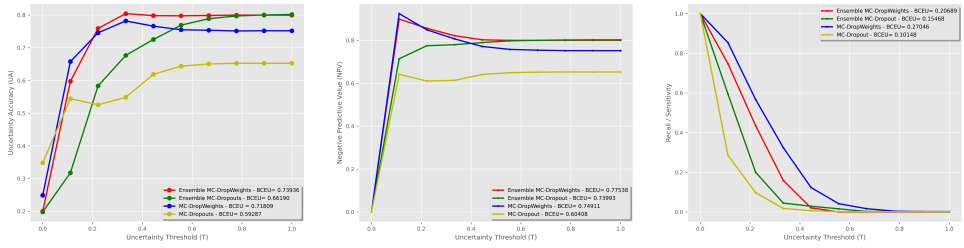
As an application to the proposed uncertainty measures, we have evaluated the uncertainty estimation performance of Bayesian Deep Ensembles of MC-DropWeights with the Ensembles MC-Dropout, MC-Dropout and MC-DropWeights using Mutual Information and Bias-Corrected Uncertainty Estimator(BCEU). Our experimental results (Fig. 5.4 and 5.5) show, that Bias-Corrected Uncertainty estimator using the Ensemble MC-DropWeights model for balanced and imbalanced dataset respectively. Note that, the uncertainty estimation metrics show a significant improvement using Deep Bayesian Neural Networks with MC-DropWeights without bias correction in measured model uncertainty when varying the uncertainty threshold.



(a)

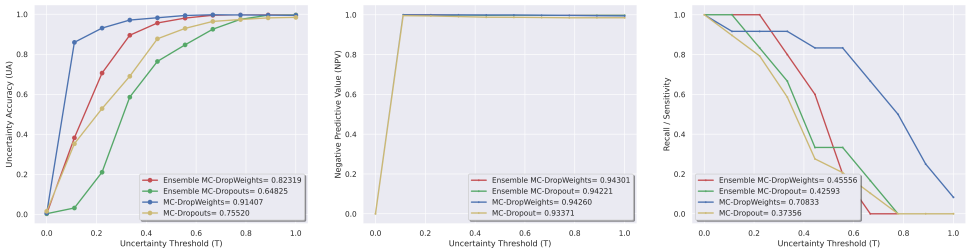


(b)

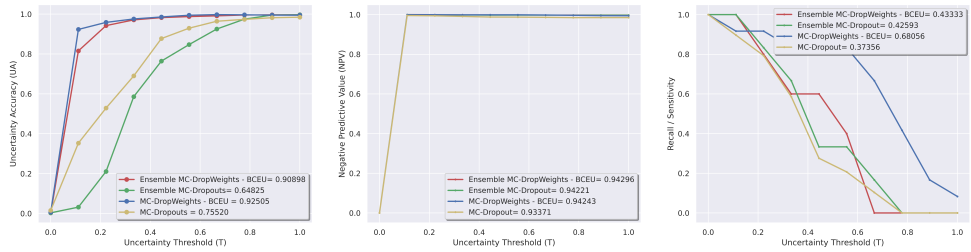


(c)

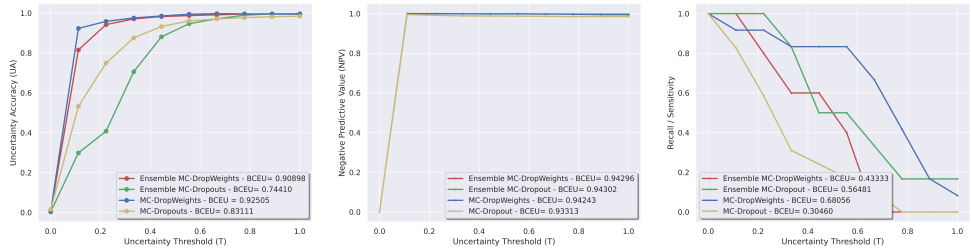
Fig. 5.4 Estimated uncertainty performance for the multi-class classification task of balanced MRI image dataset using the uncertainty evaluation metrics: Uncertainty Accuracy (UA), Negative Predictive Value (NPV), Recall/Sensitivity for (a) without bias correction of estimated uncertainty (BCEU) (b) bias correction of estimated uncertainty (BCEU) from MC-DropWeights and (c) bias correction of estimated uncertainty (BCEU) from MC-Dropout and MC-DropWeights.



(a)



(b)



(c)

Fig. 5.5 Estimated uncertainty performance for the multi-class classification task of imbalanced MRI image dataset using the uncertainty evaluation metrics: Uncertainty Accuracy (UA), Negative Predictive Value (NPV), Recall/Sensitivity for (a) without bias correction of estimated uncertainty (BCEU) (b) bias correction of estimated uncertainty (BCEU) from MC-DropWeights and (c) bias correction of estimated uncertainty (BCEU) from MC-Dropout and MC-DropWeights.

5.3 Estimating Uncertainty in Multi-Label Classification

Machine learning algorithms have been widely applied for the recognition of nuclei, that can be used for segmentation of specific cells or tissue compartments, i.e. distinguishing between epithelial and stromal cells or between benign and malignant (Blom et al., 2019; Chen and Chefd'Hotel, 2014; Stenman et al., 2020; Van Eycke et al., 2018), detection of immune cells (Aprupe et al., 2019; Swiderska-Chadaj et al., 2019), classification or quantification of certain cell states, such as mitotic cells (Tellez et al., 2018) , HER2 positive tumour cells in breast cancer (Tewary et al., 2021), or Ki67 positive proliferative cells (Feng et al., 2020; Gerad et al., 2019; Joseph et al., 2019; Saha et al., 2017). Deep learning has been used widely and with considerable success in various medical image analysis settings, including detection of disease and the localisation and estimation of affected areas for images generated by X-ray, Microscopy, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET), and ultrasound (Litjens et al., 2017). AI-driven and deep learning approaches hold much promise for efficient and accurate pattern recognition of histological images, and there have been several efforts based on IHC images over the years. For example, automated algorithms have been widely applied for the recognition of nuclei that can be used for segmentation of specific cells or tissue compartments, i.e. distinguishing between epithelial and stromal cells or between benign and malignant (Blom et al., 2019; Chen and Chefd'Hotel, 2014; Stenman et al., 2020; Van Eycke et al., 2018), detection of immune cells (Aprupe et al., 2019; Swiderska-Chadaj et al., 2019), classification or quantification of certain cell states, such as mitotic cells (Tellez et al., 2018), HER2 positive tumour cells in breast cancer (Tewary et al., 2021), or Ki67 positive proliferative cells (Feng et al., 2020; Gerad et al., 2019; Joseph et al., 2019; Saha et al., 2017). However, most of the previous studies using IHC in machine learning approaches have focused on a smaller number of markers, often well-known biomarkers. These markers have been used to train the algorithm to recognise and measure certain cell types within the tissues (Bulten et al., 2019) or to quantify the number of cells positive for a specific marker (Morriss et al., 2020). However, no previous studies suggest how such frameworks can be implemented for high-throughput annotation of complex tissue samples stained with IHC, applicable to stainings from any type of protein.

Despite impressive reported accuracy, deep learning models tend to require large training sample image sets. Whilst this can be overcome to some degree for many image tasks by using transfer-learning (Van Eycke et al., 2018), there is limited scope for this on IHC images due to the variation in protocols used to process tissue samples across different labs, though this is still a potential area for future work. Furthermore, deep learning models tend to make overconfident predictions and lack the ability to report "I do not know" for ambiguous or

unknown cases. Therefore, it is not sufficient to depend on prediction scores alone from deep learning models, but critical to estimate bias-reduced uncertainty as an additional insight to the prediction. However, no previous study has addressed the challenge presented here, training an AI model that distinguishes the cell type-specific protein expression pattern in human IHC samples, applicable to stainings from any type of protein (Long et al., 2020; Rączkowski et al., 2019).

In this section, we aim to:

1. Present the first approach in multi-label pattern recognition that can recognise various cell types-specific protein expression patterns in testes based on antibody-based proteomics images and provide information on which cell types express the protein with estimated uncertainty.

2. Show Multi-Label Classification (MLC) is achieved by thresholding the class probabilities, with the Optimal Thresholds adaptively determined by a grid search scheme based on Matthews correlation coefficient.

3. Demonstrate through extensive experimental results that a Deep Learning Model with MC-Dropweights (Ghoshal et al., 2019b) is significantly better than a wide spectrum of MLC algorithms such as Binary Relevance (BR), Classifier Chain (CC), Probabilistic Classifier Chain (PCC) and Condensed Filter Tree (CFT), Cost-sensitive Label Embedding with Multi-dimensional Scaling (CLEMS) and state-of-the-art MC-Dropout [5] algorithms across various cell types.

4. Develop Saliency Maps to increase model interpretability by visualising descriptive regions and highlighting pixels from different areas in the input image. Deep learning models are often accused of being "black boxes", so they need to be precise, interpretable, and uncertainty in predictions must be well understood.

5. Present a novel method for automated annotation of immunohistochemistry images that increased accuracy of automated image predictions leveraging an uncertainty metric called the DeepHistoClass (DHC) confidence score (Ghoshal et al., 2021a). The DHC score is cell type-specific and combines uncertainty with the predictive label probability, thereby revealing which images are reliably classified by the model, but also has the possibility to identify manual annotation errors.

5.3.1 Multi-Label Image Classification Dataset:

The HPA database based on antibody-based proteomics constitutes the largest and most comprehensive knowledge resource for spatial localization of proteins in organs, tissues,

cells and organelles. The HPA project has characterized >15,000 different proteins across >40 different normal tissues and organs, and 20 types of cancer (Hikmet et al., 2020; Uhlén et al., 2015), with the publicly available database www.proteinatlas.org containing >10 million high-resolution images, thereby constituting a major resource for machine learning algorithms.

Here, we focused on one particular organ – the testis. Based on an integrated 'omics' approach using transcriptomics and antibody-based proteomics, more than 500 proteins with distinct testicular protein expression patterns have previously been identified (Pineau et al., 2019), and transcriptomics data suggests that over 2,000 genes are elevated in testes compared to other organs. The unique nature of this tissue harbouring a large number of proteins not expressed anywhere else in the human body (Esteva et al., 2017; Jumeau et al., 2015; Morriss et al., 2020; Vandenbrouck et al., 2020). The testis has the highest number of tissue elevated genes (Djureinovic et al., 2014; Fagerberg et al., 2014) and is considered to be one of the most complex organs in the human body due to the spermatogenesis process that requires activation and suppression of thousands of genes and proteins. However, the function of a large proportion of these proteins is largely unknown, and all genes involved in the complex process of spermatogenesis are yet to be characterised (Jumeau et al., 2015; Pineau et al., 2019; Vandenbrouck et al., 2016). Spermatogenesis is built on a continuous interplay between multiple cell types and cell stages leading to sperm maturation. The process is studied in a wide variety of both primary and clinical research areas, such as toxicology (e.g., toxicants effects on germ cells or somatic cells), evaluation of male infertility in patients (e.g., maturation arrest, vacuolation, etc.), or effects on spermatogenesis as a result of different treatments (e.g., cancer therapy). For a complete understanding of the molecular mechanisms underlying normal and pathological spermatogenesis, it is necessary to study the exact localisation of all proteins related to testis specific-functions.

Nearly all cells have the same DNA, which encodes for proteins. Different cell types express different genes that dictate cells' function by the differential expression of various proteins. Different proteins are expressed in certain combinations of these cell types. Some proteins may be expressed in just one subset, while others are more ubiquitously expressed. The expression of several proteins' expression increases or decreases during differentiation, seen as a gradient in expression in cell states that undergo transformation with differences in size and shape. Therefore, the distinction of protein expression in different cell types is a multi-label image classification problem.

Previous multi-level classification studies, including a recent Kaggle challenge (Ouyang et al., 2019) have used immunofluorescence (IF) images of human cell lines, where antibody staining determined the different subcellular localisation of the protein, related to the Sub-

cellular Atlas of the HPA (Sullivan et al., 2018; Thul et al., 2017). While numerous studies are focusing on machine learning and IHC, few of these studies aim at distinguishing cell type-specific protein expression patterns using IHC, a no previous approach can be applied to any type of protein staining (Kumar et al., 2014; Long et al., 2020; Newberg and Murphy, 2008; Rączkowski et al., 2019; Xu et al., 2013). In addition to numerous research initiatives, there are several readily available commercial and open-source software supporting IHC images, such as QuPath (Morriss et al., 2020), VisioPharm (Stålhammar et al., 2016; Zhang et al., 2016), Halo (Thommen et al., 2018), Aiforia (https://www.aiforia.com/) and Definiens (https://oraclebio.com/). Some of these software require coding abilities, and others are fully operational with custom algorithms or built-in easily trained applications by which certain structures are outlined, and thresholds are set in a user-friendly interface. Tuning the software parameters for different images and staining conditions could be a tedious and time-consuming task to make such a workflow applicable to the multi-level task presented here, where each label is represented by a wide range of different staining patterns.

Manual annotation provides the standard for scoring immunohistochemical staining patterns in different cell types. However, it is tedious, time-consuming and expensive and subject to human error as it is sometimes challenging to separate cell types by the human eye (Pineau et al., 2019). Therefore, it would be extremely valuable to develop an automated algorithm that can recognise the various cell types in testes based on antibody-based proteomics images while providing information on which that cell type expresses proteins.

Our main dataset is taken from 'The Human Protein Atlas' project, which maps the distribution of all human proteins in human tissues and organs (Uhlén et al., 2015). We used IHC images from normal human testis, in which the automated model can recognize positive IHC staining in any combination of eight different cell types, stained with antibodies targeting any type of protein. We focused on generating a novel in-depth annotation dataset based on images of normal testis generated as part of the HPA project. We were careful of the potential impact of image resolution on the performance of the models. Most artificial intelligence or machine learning solutions use significantly downsampled images because of the size of neural networks, which contain millions of parameters. The size and number of images makes analysis incredibly demanding, requiring vast computational power. Given the success of deep learning models in image classification, researchers have applied the downsampled techniques used in the ImageNet competitions to medical imaging. Downsampled images are much faster to train deep neural networks. Moreover, lower-resolution images may lead to less overfitting of deep learning models that focus on important high-level features. In the present investigation, a high performance was demonstrated despite using downsampled

images, but we may see further improved performance by analyzing the full-size images, particularly for staining patterns restricted to certain cellular or subcellular level features.

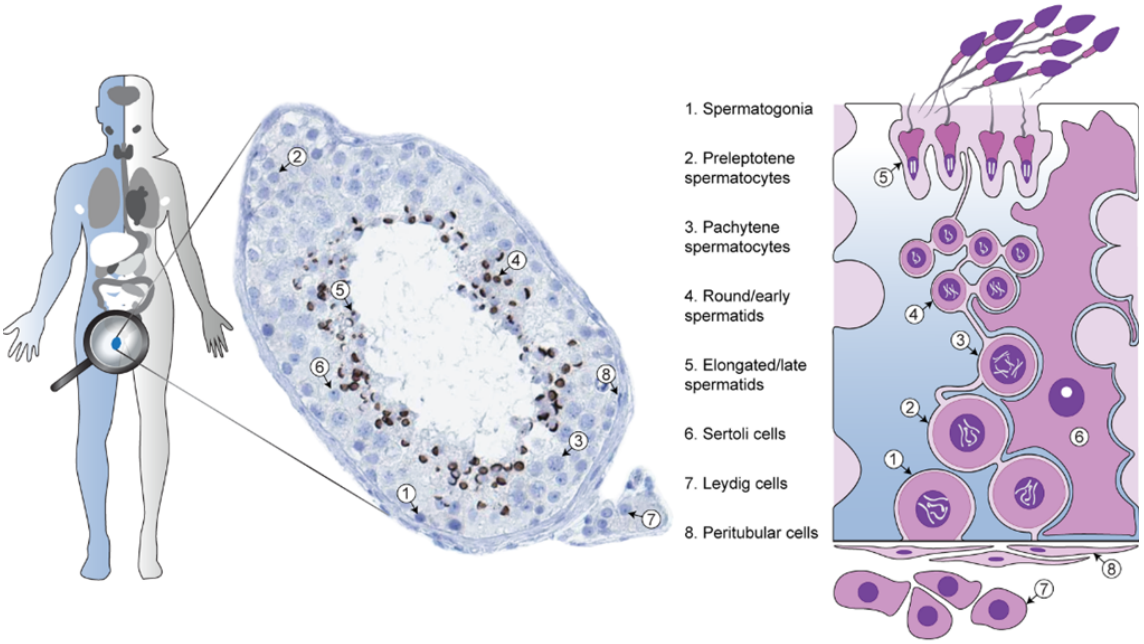


Fig. 5.6 Schematic overview: Cell Type-Specific Expression of Testis Elevated Genes (Pineau et al., 2019)

Here, we used high-resolution digital images of immunohistochemically stained testes tissue consisting of 8 cell types: spermatogonia, preleptotene spermatocytes, pachytene spermatocytes, round/early spermatids, elongated/late spermatids, sertoli cells, leydig cells, and peritubular cells, publicly available on the Human Protein Atlas version 18 (v18.proteinatlas.org), as shown in Figure 5.7:

5.3.2 Cell type-specific expression based on manual annotation

To get an overview of the protein expression pattern across the entire dataset and determine the relationship between the eight different cell types, pairwise Kendall correlation was used to create a heatmap of the protein expression correlations and the associated clusters (Figure 5.8 a) between cell types. A relationship was observed between spermatogonia and preleptotene spermatocytes cell types and between round/early spermatids and elongated/late spermatids cell types along with Pachytene spermatocytes cells. The observable pattern is that very few cell types are strongly correlated with each other.

The analysis was based on the manual annotation of staining intensity across the entire dataset of 7,848 images. As expected based on functional characteristics (Pineau et al.,

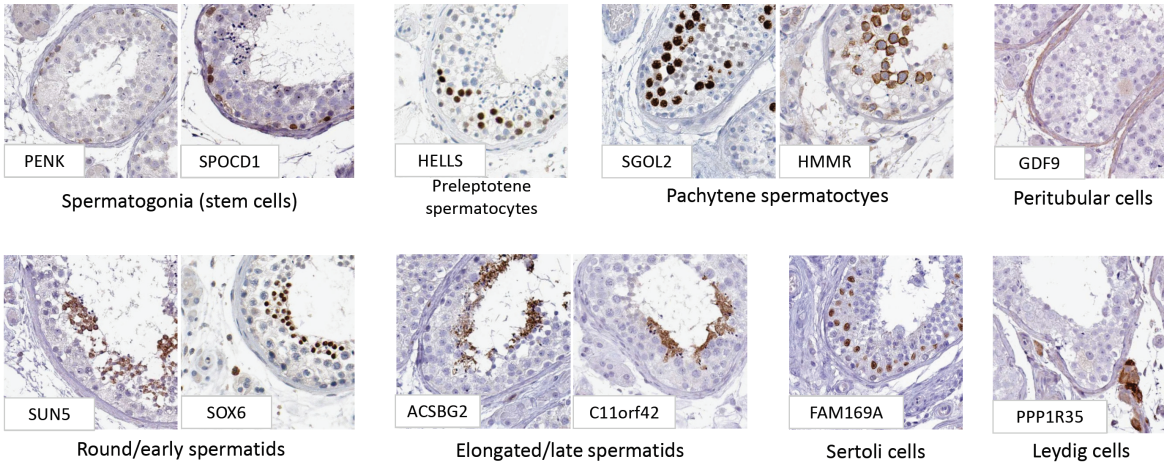


Fig. 5.7 Examples of proteins expressed only in one cell-type (Pineau et al., 2019)

2019), there were three main clusters: i) somatic cells (Sertoli cells, Leydig cells and peritubular cells), ii) premeiotic cells (spermatogonia and preleptotene spermatocytes), and iii) meiotic/post-meiotic cells (pachytene spermatocytes, round/early spermatids and elongated/late spermatids). Of the 7,848 images analysed, only 815 (10%) showed the only immunoreactivity in one cell type, while most of the images were positive in two to five cell types (Figure 5.8 b) and visualised as a waffle distribution plot. In 35 images, the human observer had marked all cell types as negative. When separated, the three different sets showed slightly different proportions of the number of positive cell types (Figure 5.8 c), where the test set consisted of more cell type-specific images and the validation set contained a higher proportion of images with five to eight cell types that had been labelled (Figure 5.8 c). There were large differences in the presence of different cell type labels (Figure 5.8 d), with Leydig cells being labelled in as many as 5,218 (66%) of the images, while peritubular cells represented the most unusual staining pattern, positive in only 755 (10%) of the images. The staining was mostly localised to the cytoplasm, the cytoplasm, the plasma membrane, or the nucleus, but there were clear differences between cell types. Sertoli cells more often showed positivity in the plasma membrane or a combination of nucleus + membrane, in most cases referred to as the nuclear membrane. A majority of the staining observed in Leydig cells was cytoplasmic (Figure 5.8 d).

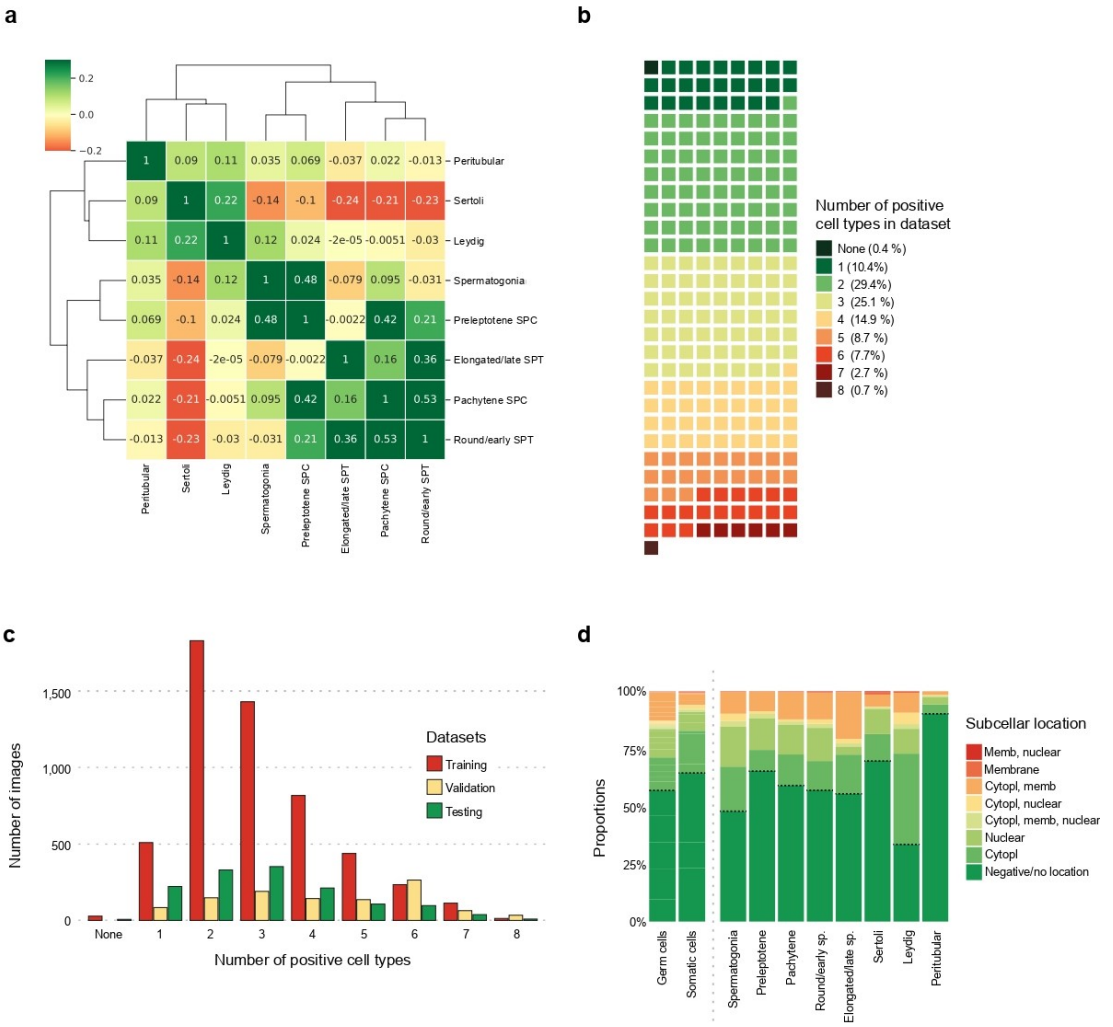


Fig. 5.8 Input image data distribution based on manual annotation. (A) Heatmap and cluster analysis of testicular cell types. (B) Waffle distribution plot. (C) The bar chart of the number of positive cell types by each dataset. (D) The distribution of subcellular location.

5.4 Multi-Label Cell-Type Recognition and Localisation with estimated uncertainty

5.4.1 Problem Definition:

Given a set of training data D, where $X = \{x_1, x_2 \ldots x_N\}$ is the set of N images and the corresponding labels $Y = \{y_1, y_2 \ldots y_N\}$ is the cell-type information. The vector $y_i = \{y_{i,1}, y_{i,2} \ldots y_{i,M}\}$ is a binary vector, where $y_{i,j} = 1$ indicates that the i^{th} image belongs to the j^{th} cell-type. Note that an image may belong to multiple cell-types, i.e., $1 <= \sum_j y_{i,j} <= M$. Based on $D(X,Y)$, we constructed a Bayesian Deep Learning model giving an output of the predictive probability with estimated uncertainty of a given image x_i belonging to each cell category. That is, the constructed model acts as a function such that $f : X \rightarrow Y$ using weights of neural net weight parameters w where $(0 <= \hat{y}_{x,j} <= 1)$ as close as possible to the original function that has generated the outputs Y, output the estimated value $(\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,M})$ as close to the actual value $(y_{i,1}, y_{i,2}, \ldots, y_{i,M})$.

5.4.2 Solution Approach:

We tailored Deep Convolutional Neural Network (DCNN) architectures for cell type detection and localisation by considering a large image capacity, binary-cross entropy loss, sigmoid activation, along with Dropweights in the fully connected layer and Batch Normalisation formulation of propagating uncertainty in deep learning to estimate meaningful model uncertainty.

Multi-label Setup:

There are multiple approaches to transform the multi-label classification into multiple single-label problems with the associated loss function (Huang and Lin, 2017). This study used immunohistochemically stained testes tissue consisting of 8 cell types corresponding to 512 testes elevated genes.

Therefore, we define a 8-dimensional class label vector $Y = \{y_1, y_2 \ldots y_N\}$; $Y \in \{0,1\}$, given 8 cell types. y_c indicates the presence with respect to according cell type expressing the protein in the image while an all-zero vector [0; 0; 0; 0; 0; 0; 0; 0] represents the “Absence” (no cell type expresses the protein in the scope of any of 8 categories).

Multi-label Classification Cost Function:

The cost function for Multi-label Classification has to be different because a prediction for a class is not mutually exclusive. So we selected the sigmoid function with the addition of binary cross-entropy.

Data Augmentation:

We used Keras' image pre-processing package to apply affine transformations to the images, such as rotation, scaling, shearing, and translation during training and inference. This reduces the epistemic uncertainty during training, captures heteroscedastic aleatoric uncertainty during inference, and improves models' performance.

Multi-label Classification Algorithm:

In Bayesian classification, the mean of the predictive posterior corresponds to the parameter point estimates, and the width of the posterior reflects the confidence of the predictions. The network's output is an M-dimensional probability vector, where each dimension indicates how likely each cell type in a given image expresses the protein. The number of cell types that simultaneously express the protein in an image varies. One method to solve this multi-label classification problem is placing thresholds on each dimension. However, different dimensions may be associated with different thresholds. If the value of the i^{th} dimension of $\hat{y}$ is greater than a threshold, we can say that the i-th cell-type is expressed in the given tissue. The main problem is defining the threshold for each class label.

A threshold based on Matthews Correlation Coefficient (MCC) is used on the model outcome to determine the predicted class to improve the accuracy of the models.

We adopted a grid search scheme based on Matthews Correlation Coefficients (MCC) to estimate the optimal thresholds for each cell type-specific protein expression (Chu and Guo, 2017). Details of the optimal threshold finding algorithm is shown in Algorithm 3.

The idea is to estimate the threshold for each cell category in an image separately. We convert the predicted probability vector with the estimated threshold into binary and calculate the Matthews correlation coefficient (MCC) between the threshold and actual values. The Matthews correlation coefficient for all thresholds is stored in the vector ω, from which we find the index of threshold that causes the largest correlation. The Optimal Threshold for the i^{th} dimension is then determined by the corresponding value. We then leveraged the Bias-Corrected Uncertainty quantification method (Ghoshal et al., 2019a) using Deep Convolutional Neural Network (DCNN) architectures with Dropweights (Ghoshal et al., 2019b).

Algorithm 3: Find Optimal Threshold

Input: Ground Truth Vector: $\{y_{i,1}, y_{i,2}, \ldots, y_{i,M}\}$;
Estimated Probability Vector: $\{\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,M}\}$;
Upper Bound for threshold = Ω, and Threshold Stride = S
Result: The Optimal Thresholds T = $(ot_1, ot_2, \ldots, ot_M)$
Initialization: The set of threshold T = $(ot_1 = 0, ot_2 = 0, \ldots, ot_M = 0)$;
for $i \leftarrow 1$ ***to*** M **do**
    $j \leftarrow 0$;
    $\omega \leftarrow 0$;
    $\pi \leftarrow 0$;
    for $j < \Omega$ **do**
        Initialize M-dimensional binary vector $\mathbf{v} \leftarrow (v_1 = 0, v_2 = 0, \ldots, v_M = 0)$;
        if $\hat{y}_i > j$ **then**
            $v_i \leftarrow 1$;
        else
            $v_i \leftarrow 0$;
        $\omega \leftarrow \omega.append(MCC(\mathbf{y}[1:i], v))$;
        $\pi = \pi.append(j)$;
        $j = j + S$
    $\hat{m} \leftarrow argmax_m \omega = (\omega_1, \omega_2, \ldots, \omega_m, \ldots)$;
    $ot_i = \pi[\hat{m}]$

Network Architecture:

Our models are trained and evaluated using Keras with Tensorflow backend. For the DNN architecture, we used a generic building block containing the following model structure: Conv-Relu-BatchNorm-MaxPool-Conv-Relu-BatchNorm-MaxPool-Dense-Relu-Dropweights and Dense-Relu-Dropweights-Dense-Sigmoid, with 32 convolution kernels, 3x3 kernel size, 2x2 pooling, dense layer with 512 units, 128 units, and eight feed-forward Dropweights probabilities 0.3. We optimised the model using Adam optimiser with the default learning rate of 0.001. The training process was conducted in 1000 epochs, with a mini-batch size of 32. We repeated our experiments three times for an algorithm and calculated the mean of the results.

Following Gal (Gal, 2016), we define the stochastic versions of Bayesian uncertainty using MC-Dropweights, where the class probabilities $p(y_{x_i} = c \mid x_i, w_t, D)$ with $w_t \sim q(w \mid D)$ and $W = (w_t)_{t=1}^{T}$ along with a set of independent and identically distributed (i.i.d.) samples drawn from $q(w \mid D)$, can be approximated by the average over the MC-Dropweights forward pass.

We trained the multi-label classification network with all eight classes. We dichotomised the network outputs using optimal threshold with algorithm 1 for each cell type, with a 1000 MC-Dropweights forward passes at test time. In these detection tasks, $p(y_{x_i} >= 0; OptimalThreshold_i \mid x_i, w_t, D)$, where 1 marks the presence of cell type, is sufficient to indicate the most likely decision along with estimated uncertainty.

5.4.3 Results and Discussions

We conducted the experiments on Human Protein Atlas datasets to validate the proposed algorithm, MC-Dropweights in Multi-Label Classification.

Multi-Label Classification Model performance:

We successfully associated deep learning-based predictions on cell type-specific protein expression patterns in histological testis sections stained with IHC. Quality metrics that are typically being used in binary classifications or single-label multi-classifications include area under the curve (AUC) or receiver operating characteristics (ROC). In multi-label classification, the predictions constitute a subset of actual class labels, and therefore, the prediction can be fully incorrect, partially correct or fully correct. As a result, AUC cannot be directly calculated for multi-label classifications but separately computed for each label. Multiple ROC analyses can be carried out through aggregation, but this does not take into account class label imbalance. Here, we assessed multi-label classification using Matthews

%Metrics	BR	CC	PCC	CFT	CLEMS	MC-Dropout	MC-Dropweights
Hamming Loss (↓)	0.2445	0.2420	0.2420	0.2375	0.2370	0.207	0.1925
Rank Loss (↓)	3.6700	3.5740	3.1580	3.2920	3.1120	2.862	2.626
F1 Score (↑)	0.5038	0.5184	0.5733	0.5373	0.5902	0.6306	0.6627
Avg. Accuracy Score (↑)	0.4236	0.4389	0.4643	0.4573	0.5052	0.6150	0.7067

Table 5.2 Performance Metrics

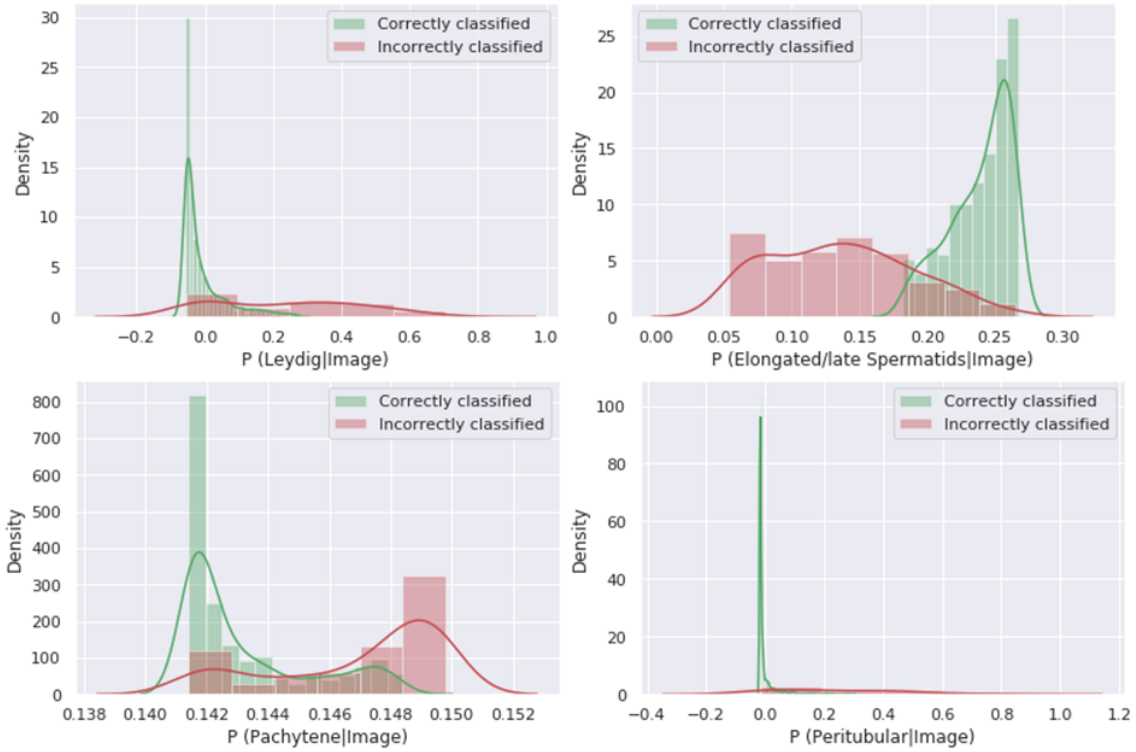
Correlation Coefficient (MCC), a common metric for analyzing such classifiers. This metric has the attractive property of dealing with imbalance and asymmetry.

Model evaluation metrics for multi-label classification are different from those used in multi-class (or binary) classification. The performance metrics of multi-label classifiers can be classified as label-based (i.e., it is assumed that labels are mutually exclusive) and example-based (Wu and Zhou, 2017). In this work, example-based measures (Accuracy score, Hamming-loss, F1-Score) and Rank-Loss are used to evaluate the performance of the classifiers.

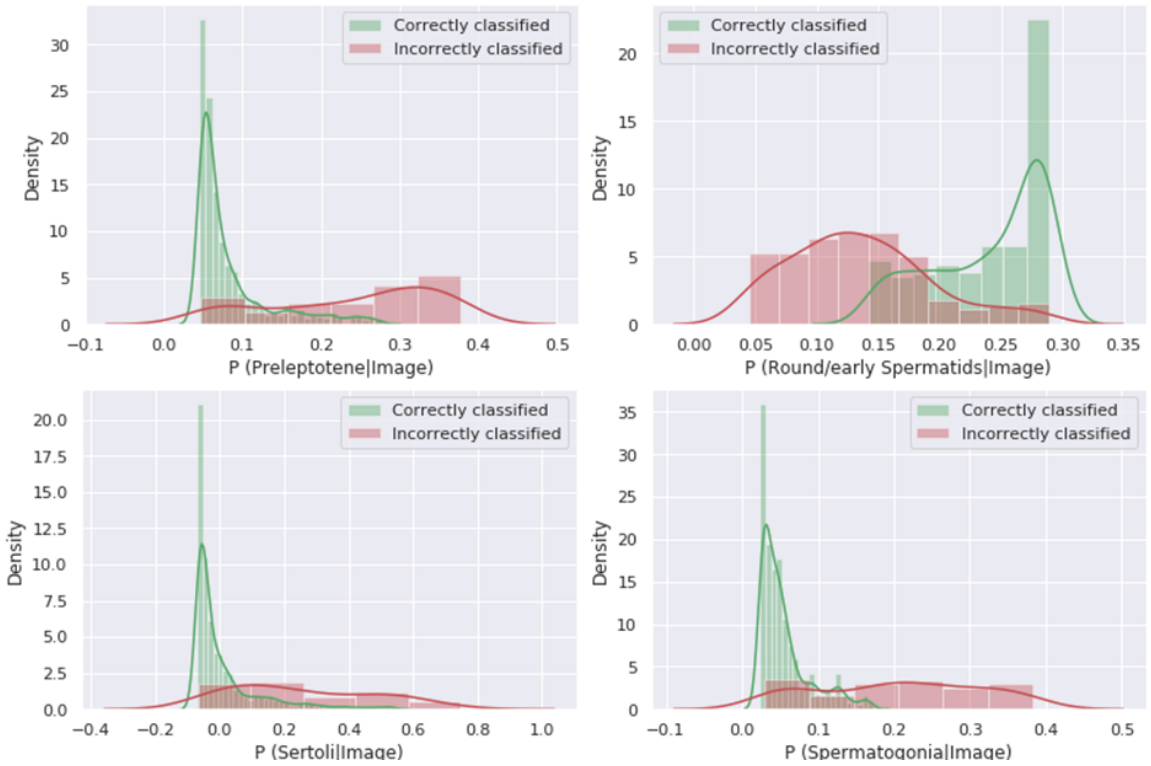
In the first experiment, we compared the MC-Dropweights neural network-based method with five machine learning multi-label classification algorithms: binary relevance (BR), Classifier Chain (CC), Probabilistic Classifier Chain (PCC) and Condensed Filter Tree (CFT), Cost-Sensitive Label Embedding with Multi-dimensional Scaling (CLEMS) and the MC-Dropout neural network model. Table 5.2 shows that MC-Dropweights exhibits considerably better performance overall the algorithms, which demonstrates the importance of considering the Dropweights in the neural network.

Cell Type-Specific Predictive Uncertainty:

The relationship between uncertainty and predictive accuracy grouped by correct and incorrect predictions is shown in Figure 5.9. It is interesting to note that, on average, the highest uncertainty is associated with Elongated/late Spermatids and Round/early Spermatids. This indicates that some feature contributes greater uncertainty to the Spermatids class types than the other cell types.



(a) Leydig, Elongated/Late Spermatids, Pachytene, Peritubular Cell Type



(b) Preleptotene, Round/Early Spermatids, Sertoli, Spermatogonia Cell-Type

Fig. 5.9 Distribution of uncertainty values for all protein images, grouped by correct and incorrect predictions. Label assignment was based on optimal thresholding (algorithm 1). For an incorrect prediction, there is a strong likelihood that the predictive uncertainty is also high in all cases except for Spermatids.

Cell Type Localization:

Estimated uncertainty with Saliency Mapping is a simple technique to uncover discriminative image regions that strongly influence the network prediction in identifying a specific class label in the image. It highlights the most influential features in the image space that affect the predictions of the model (Adebayo et al., 2018) and visualises the contributions of individual pixels to epistemic and aleatoric uncertainties separately. We calculated the class activation maps (CAM) (Zhou et al., 2016) using the activations of the fully connected layer and the weights from the prediction layer as shown in Figure 5.10.

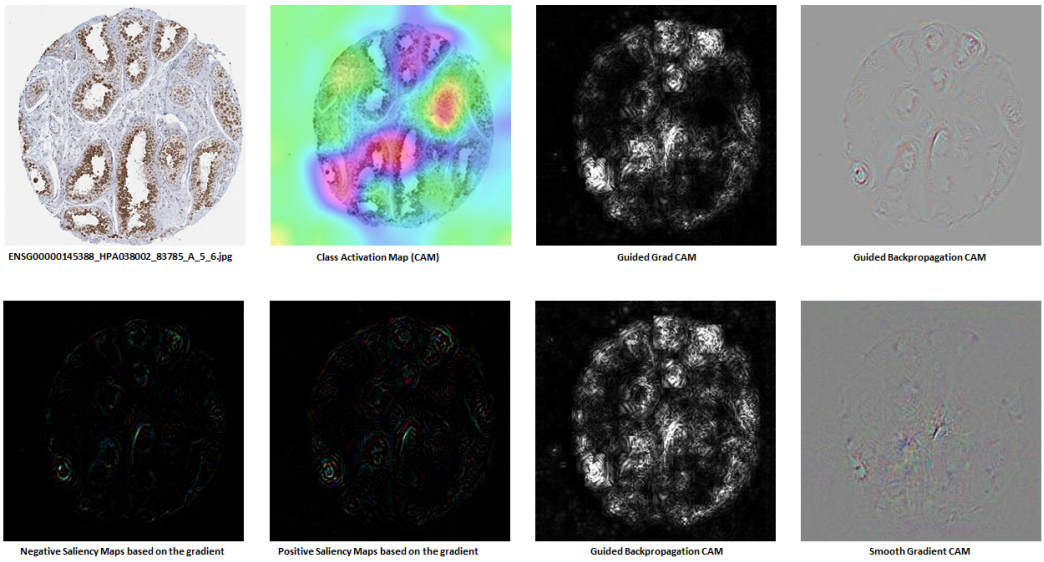


Fig. 5.10 Saliency maps for some common methods towards model explanation

5.5 DeepHistoClass: A novel strategy for confident classification of immunohistochemistry images using Deep Learning

Human physiology depends on complex processes built on intercellular interactions and cell type-specific functions unique to each tissue and organ. To fully understand the underlying mechanisms of disease, it is necessary to study tissue architecture and molecular constituents with a single-cell resolution. In the field of transcriptomics, dramatic improvements have been made in the single-cell RNA-seq (scRNA-seq) technology, which is a powerful approach due to its excellence in studying mRNAs in smaller subsets of cells that would fall below detection limits when mixed with other cell types in complex tissues samples (Regev et al., 2017). One major initiative taking advantage of this new technology is the Human Cell Atlas consortium (www.humancellatlas.org). While transcriptomics has the advantage of quantitative measurements and low abundance detection, it is important to note that validation at the protein level is necessary to understand the role in health and disease, as proteomics constitutes the functional representation of the genome. This has recently been shown for expression of the SARS-CoV-2 receptor ACE2, where low abundant measurements based on transcriptomics do not fully reveal the exact localization in tissues unless complemented with proteomics approaches (Hikmet et al., 2020).

The standard method for visualizing proteins with a single-cell resolution is antibody-based proteomics and immunohistochemistry (IHC), which allows studying the protein localization in histologically intact tissue samples. This allows for determining the localization in different compartments at a tissue, cellular, and subcellular level and provides important information in the context of neighbouring cells. IHC thus constitutes an excellent method for direct validation of cell-type-specific expression patterns identified by scRNA-seq. The most significant initiative for mapping the human proteome using IHC is the Human Protein Atlas (HPA) project (Sjöstedt et al., 2020; Thul et al., 2017; Uhlén et al., 2015; Uhlen et al., 2019, 2017), covering all major normal tissues and organs, as well as the most common forms of cancer. The open-access database visualizes the expression of >80% of all human proteins in >10 million high-resolution images, constituting an excellent resource for comparison of cell-type-specific expression patterns identified with large-scale transcriptomics approaches, which has recently been shown in the new Single Cell Type Atlas www.proteinatlas.org/humanproteome/celltype (Karlsson et al., 2021).

Despite the IHC technology having been used for decades and is a standard method in clinical pathology, the main approach for evaluation of IHC staining patterns is still the rather

subjective manual assessment. A manual observer has the advantage of identifying technical staining errors or artefacts, but it is both time-consuming and costly. Additionally, manual annotation is error-prone and poorly reproducible. It may lead to fatigue or mislabeling of images due to lack of experience in detecting the correct cell types or structures or technological challenges related to staining intensity or identification of small objects. Manual annotation is commonly faced with two types of errors, i) false negatives where true positive staining is missed or neglected, and ii) false positives where lack of protein expression is falsely interpreted as positive. Histological samples consist of a mixture of different cell types that can be challenging to distinguish even by a trained eye, and setting a manual threshold of what is regarded as negative/positive is tedious and highly difficult. This leads to challenges in large-scale approaches aiming at aligning IHC datasets with data generated by other quantitative methods, such as scRNA-seq.

To increase accuracy and speed up the process of manual interpretation, the application of Artificial Intelligence (AI) in the evaluation of medical images has received increased attention both in research and diagnostics (Bejnordi et al., 2017; Esteva et al., 2017; Gulshan et al., 2016; Jackson et al., 2020; Nagpal et al., 2019). AI-driven and deep learning approaches hold much promise for efficient and accurate pattern recognition of histological images, and there have been several efforts based on IHC images. However, most of these previous studies using IHC in machine learning focused on a smaller number of markers, often well-known biomarkers. These markers were either used to train the algorithm recognizing and measuring the presence of certain cell types within the tissues (Bulten et al., 2019), or to quantify the number of cells positive for a certain marker (Morriss et al., 2020). No previous study has addressed the challenge presented here, training an AI model that distinguishes the cell type-specific protein expression pattern in human IHC samples, applicable to stainings from any type of protein (Long et al., 2020; Rączkowski et al., 2019).

One of the challenges when implementing AI models for automated annotation of IHC is that IHC images typically consist of a complex mixture of multiple cell types of various shapes and sizes that can express a protein in different combinations. Additionally, a protein may not only be expressed in certain cell types, but could also be localized to different subcellular compartments, e.g., cytoplasm or nucleus, or be expressed at different levels. As a result, training an algorithm to distinguish cell type-specific localization of proteins based on IHC is a multi-label task. Since each class is not mutually exclusive, both the manual observer and the trained model must consider every possible label separately. Different approaches to address multi-label classification problems have been developed previously (González-López et al., 2019), but none of these have been applied to IHC images. Another challenge is correctly addressing the accuracy of automated predictions, which is especially important

when implementing algorithms in a clinical setting, and in whole-proteome approaches such as the HPA project to compare results between different proteins at a global proteome-wide level. Addressing prediction accuracy requires a large dataset of manually annotated images and a method to score the confidence in the prediction. Few existing large-scale imaging datasets are labelled in detail at a cell type-specific level, and many state-of-the-art algorithms do not currently consider methods for addressing prediction accuracy. Bayesian neural networks (BNNs) learn a distribution with a prior distribution on its weights and are currently considered state-of-the-art for estimating uncertainty in model prediction, thereby constituting an important element when building automated workflows for annotation of histological images, which was shown in a recent study (Ghoshal et al., 2020).

The goal of the present investigation was to build upon our earlier work on estimating uncertainty in deep learning (Ghoshal et al., 2020), to present a reliable and comprehensive framework for automated annotation of IHC images that addresses prediction accuracy and that can be used for large-scale approaches. As a model system, we focused on one particular organ—the testis—due to its complex histological features with as many as eight different cell types that the human eye can distinguish. These cell stages involved in spermatogenesis and sperm maturation require activation and suppression of thousands of genes and proteins, out of which a large proportion has an unknown function (Djureinovic et al., 2014; Fagerberg et al., 2014; Jumeau et al., 2015; Pineau et al., 2019; Vandenbrouck et al., 2016). As a basis, we included a large set of 7848 human testis histology images, corresponding to IHC stainings of 2794 different proteins, generated as part of the HPA project. The previous standard HPA annotation in two different testicular cell types for these images was replaced by a new manual in-depth characterization in eight different cell types, which formed the basis for model training in the present investigation. Our automated framework was built for recognizing IHC staining patterns at a cell-type-specific level in each of these eight cell types and addresses uncertainty with a novel metric—DeepHistoClass (DHC) Confidence Score. The DHC Score is cell-type-specific and combines uncertainty with the predictive label probability, thereby revealing which images are reliably classified by the model, but it also has the possibility to identify manual annotation errors.

DeepHistoClass (DHC) Confidence Score

DeepHistoClass (DHC) Confidence Score in supervised learning can be considered as a measure of representativeness, information content, and diversity on high dimensional image classification by estimating uncertainty from an approximate Bayesian Neural Networks and class predictive probability distance. For example, we identify for images with low DHC

score and choose those images to be labelled by an expert in the hope that these will improve model performance and decrease model uncertainty.

We employ the maximum class predictive probability distance (CPPD), which is the difference between the probability values of the highest and the second highest predictive probability value as a measure of a representativeness heuristic. The vector of softmax probabilities $\hat{y}_t = \text{Softmax}\, f^{\hat{w}_t}(\hat{x})$ obtained after the t th stochastic forward pass is denoted $p\left(\hat{y}_t|x^*, \hat{\theta}_\text{t}\right)$, where $\hat{\theta}_\text{t}$ denotes the sampled parameters resulting from DropWeights. Thus, the class probabilities of estimates are given by $\mu_{pred} = \frac{1}{T}\sum_{t=1}^{T} p\left(\hat{y}_t|x^*, \hat{\theta}_\text{t}\right)$. We obtain the Class Predictive Probability Distance (CPPD):

$$\text{CPPD}\,(x_i) = \arg\min_{x_i \in U}\left(\frac{1}{T}\sum_{t=1}^{T} p\left(\hat{y}_{Best}|x_i^*, \hat{\theta}_\text{t}\right) - \frac{1}{T}\sum_{t=1}^{T} p\left(\hat{y}_{NextBest}|x_i * \hat{\theta}_\text{t}\right)\right) \tag{5.4}$$

The MC-DropWeights estimate of the vector of softmax probabilities aim to decompose the source of uncertainty. We propose Bayesian deep learning framework for image classification to directly estimate the predictive uncertainty. The predictive mean can be estimated by:

$$\hat{\mu}_\text{c} = \frac{1}{T}\sum_{t=1}^{T} p\left(\hat{y} = c|x^*, \hat{\theta}_\text{t}\right); c \in \{1, \ldots, \text{C}\} \tag{5.5}$$

The predictive variance to measure uncertainty can be estimated as:

$$\text{Estimated Uncertainty}\ \left(\hat{\sigma}_\text{epistemic}\right) : \frac{1}{\text{C}}\sum_{\text{i}=1}^{\text{C}}\sqrt{\frac{1}{\text{T}}\sum_{\text{t}=1}^{\text{T}}\left[p\left(\hat{y}_\text{t} = c|\text{x}^*, \hat{\theta}_\text{t}\right) - \hat{\mu}_\text{c}\right]^2} \tag{5.6}$$

$$\text{where}\ \hat{y}_t = y\left(\hat{\theta}_t\right) = \text{Softmax}\left(f\hat{\theta}_t(\hat{x})\right)$$

The main idea is to select unlabeled samples that are not only highly uncertain but also highly representative. In our approximated uncertainty measure in prediction (i.e. equation 5.6), 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.

We can calculate DHC as below:

$$\text{DHC} = \frac{\text{CPPD}\,(x_i)}{\text{Estimated Uncertainty}\left(\hat{\sigma}_\text{epistemic}\right)} \tag{5.7}$$

The lower the DHC value, the higher the information content of the corresponding sample images which should represent uncertain predictions. In practice, the smaller DHC value and corresponding uncertainty estimate should ideally refer to the least number of images, where as many as possible should be for incorrect image classifications while simultaneously having the largest proportion is labeled for correctly classified images and subsequently added to the training set.

1. $DHC \approx 1$ means that class predictive probability distance and uncertainty are relatively similar. This happens if a) the models have failed to reach a consensus (class membership difference is small) but model uncertainty is low, or b) the models have reached a consensus (class membership difference is large) but model uncertainty is high.

2. $DHC \rightarrow 0$ means that uncertainty is much larger than class membership difference. These set of images represents uncertain predictions.

3. $DHC \rightarrow \infty$ means that uncertainty is much smaller than difference. These set represents predictions with high confidence.

In order to accelerate the learning process, it is necessary to select more than one unlabeled sample at each iteration. But batch oriented active learning methods are usually affected by out-of-domain labeling samples or redundancy between the selected samples. Therefore, the diversity of the selected samples needs to be exploited. We adapted Greedy Query Strategies. We rank all unlabeled samples in ascending order of quotient value. The formulation for the sample selection measure can be given as: $X_{DHC} = argsort\,\{DHC_x\}\,[:batchsize]$. The active learner can then start to query points for which the model has the lowest X_{DHC} for the specified batch size.

5.5.1 Generation of a semi-automated image annotation framework

We propose a semi-automated annotation framework and confidence metrics for multi-label classification of cell type-specific protein expression patterns in testis, based on a Hybrid Bayesian Neural Network (HBNet). This is the first study combining deep learning of multi-label IHC images with uncertainty measures to the best of our knowledge. The model has important implications for unbiased high-throughput annotation of IHC images and will aid in gaining important biological insights within the field of spatial proteomics, ultimately leading to further understanding of human cell biology in health and disease.

The general view and detailed image annotation framework is illustrated by Fig. 5.11. A Hybrid Bayesian Neural Network (HBNet) model was trained, considering both hand-crafted and deep learning features. The input IHC high-resolution images consisted of 1-3 human testis TMA punch-outs for each antibody, comprising a total of 7,848 images. For each antibody, eight different cell types were manually inspected with regards to staining intensity (negative, weak, moderate, strong) and subcellular location (cytoplasm, nucleus, membrane); 1: Spermatogonia; 2: Preleptotene spermatocytes; 3: Pachytene spermatocytes; 4: Round/early spermatids; 5: Elongated/late spermatids; 6: Sertoli cells; 7: Leydig cells; 8: Peritubular cells. The manual data was used as a basis for machine learning, combining hand-crafted features with standard deep learning features. The mean predictive probability and bias-corrected estimated uncertainty were used to generate a DeepHistoClass (DHC) confidence score, which allowed for dividing the images into those reliably predicted by the model and those of high uncertainty that need a manual inspection.

The proposed streamlined workflow for automated annotation of IHC images constitutes an excellent method for large-scale approaches that currently rely on manual annotation. The method can discard highly uncertain predictions, highlight which images need to be checked manually, and identify unfamiliar patterns or manual errors corresponding to outliers in the data distribution. The method has important implications for large-scale protein mapping efforts such as the HPA project or other digital pathology initiatives to save time and lead to higher accuracy in the exploration of cell-type-specific protein expression patterns in health and disease.

A total of 7,848 IHC stained high-resolution images of the human testis, corresponding to 3,046 different antibody stainings and 2,794 unique proteins there divided into three different sets: a training set (5,411 images), a validation set (1,063 images) and a test set (1,374 images). All images were annotated manually in five germ cell types (spermatogonia, preleptotene spermatocytes, pachytene spermatocytes, round/early spermatids and elongated/late spermatids), and three somatic cell types (Sertoli cells, Leydig cells and peritubular cells), taking into consideration staining intensity (negative, weak, moderate, strong) and subcellular

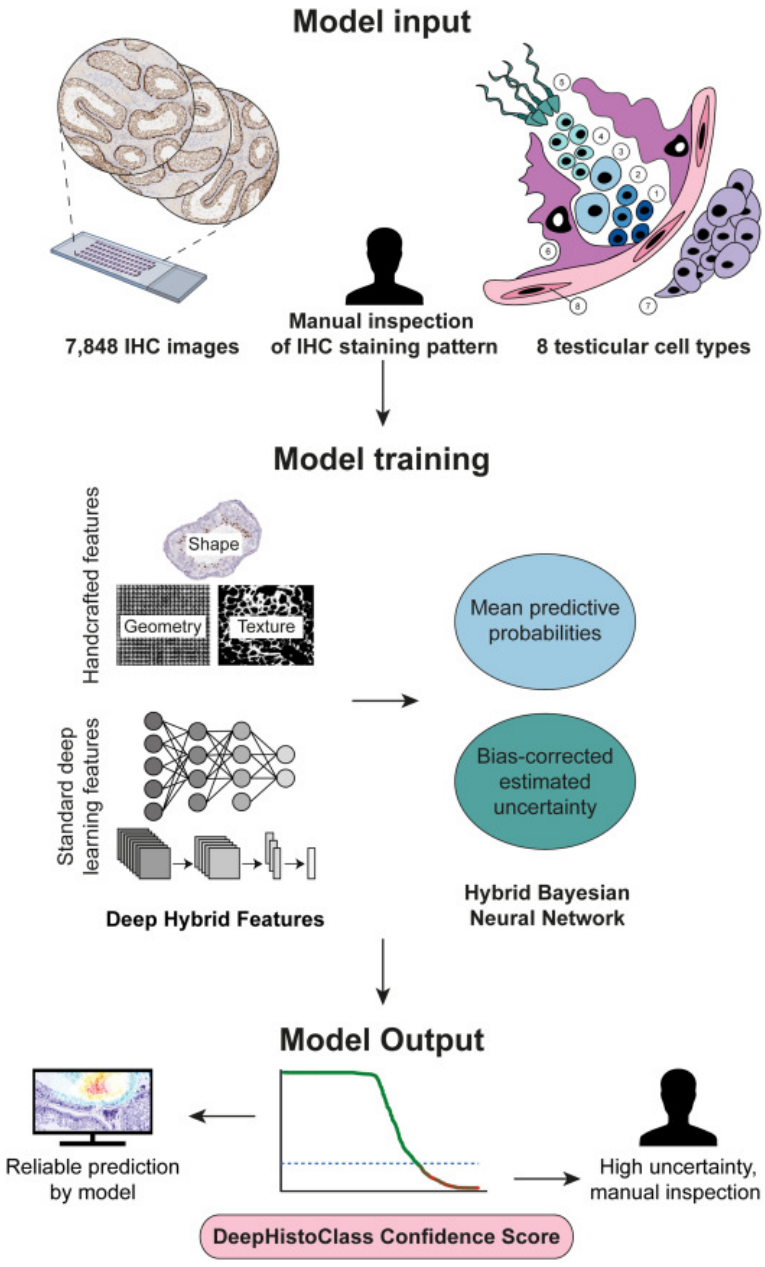


Fig. 5.11 Overview of the image annotation framework.

localisation of the staining (cytoplasm, nucleus, membrane). This novel refined scoring in eight different cell type manually scored images formed the basis for a semi-automated image annotation framework, as presented in Figure 5.11.

5.5.2 Training of neural network and overall model performance

The manually annotated images from the training set of 5,411 images and the validation set of 1,063 images were used for training a Hybrid Bayesian Neural Network (HBNet) model, exploiting DropWeights and combining the features from a standard deep neural network (DNN) with hand-crafted features. The neural network's output is an 8-dimensional probability vector, where each dimension indicates how likely each cell type in a given image expresses the protein. The neural network was then applied to the test set of 1,374 images, for which the accuracy was evaluated.

Evaluation metrics for multi-label classification performances are different from those used in binary or multi-class classification (Wu and Zhou, 2017). In multi-label classification, a miss-classification is no longer a definite right or wrong since a correct prediction containing a subset of the actual labels is considered better than a prediction containing none of them. Here, four different metrics were used for evaluating the multi-label classification performance: i) Hamming loss, ii) F1-score, iii) Exact Match ratio, and iv) mean-Average Precision (mAP). Table 5.3 presents the evaluation of classification performance for a Hand-crafted features with Neural Network, CNN features Neural Network, Hybrid features Multilabel k Nearest Neighbours, Hybrid features Random Forest Classifier, Hybrid features Support Vector Machine, Hybrid features deep neural network (DNN) and the proposed Hybrid Bayesian Neural network (HBNet), based on five different metrics. The results for each metric are shown as a percent. Hamming loss is the most common evaluation metric in multi-label classification, which considers prediction errors (false positives) and missed predictions (false negatives), normalised over the total number of classes and the total number of samples analysed. The smaller the value of Hamming loss (closer to 0), the better the learning algorithm's performance. F1 score is the harmonic mean of precision and recall, where Macro F1 score calculates the metric independently for each class label and then takes an average, and Micro F1 score aggregates the contributions of all labels when calculating the average metric. The Exact Match ratio is the strictest metric, indicating the percentage of all analysed samples with all their labels classified correctly. Mean Average Precision (mAP) considers both the average precision (AP) separately for each label and the average over the class. It provided a measure of quality across recall levels and was shown to be stable and able to distinguish between cell types. The higher the mAP (closer to 100), the better the quality. There was considerable improvement in HBNet across all metrics used (Table 1).

Based on HBNet, the Exact Match ratio showed that 67% of the 1,374 images were correctly classified in all eight cell types.

5.5.3 Cell type-specific model performance

Next, we evaluated the model's performance on a cell type-specific level. In Figure 5.12, a confusion matrix is shown, comparing the neural network's output with the manual observer, summarising the false positives and negatives of the DNN and the HBNet for each cell type. For all cell types, HBNet had a higher accuracy than DNN, with >80% overall accuracy and >90% for Sertoli cells and peritubular cells. The largest difference between DNN and HBNet was seen for pachytene spermatocytes and round/early spermatids, where the accuracy improved from 75.6 to 82.6%, and 69.3 to 80.5%, respectively. In addition, HBNet dramatically reduced the number of false negatives compared to DNN and showed a decrease in the number of false positives. The total number of false positives (n=444) across all cell types was lower compared to the number of false negatives (n=993), indicating that the model performed better at accurately detecting positive labels but more often differed with the human observer in classifying cell types as negative. This is expected due to the human observer deliberately neglecting very weak staining patterns that can be considered unspecific or being due to artefacts. However, the ratios between false positives and false negatives were opposite for Sertoli cells and peritubular cells, for which false negatives were rare. Positivity in these cell types was generally less common (Figure 5.8 D) and, to a larger extent, cell type-specific and not as often showing simultaneous staining in other cell types (Figure 5.8 A). This suggests that positivity in these cell types was mostly considered specific by the human observer.

5.5.4 Estimation of model certainty

To rank all images based on model confidence over eight cell types, each prediction included an uncertainty measurement, presented as a DHC Score. Table 5.4 shows the predictions per cell type for each of the 1,374 images in the test set, along with DHC Score, predictive probability and manual annotation. The DHC Scores ranged from zero to one for each HBNet prediction over the eight-cell types. All predictions were then plotted in confidence maps (Figure 5.13), where images for which the model agreed with the human observer, i.e. the cell type was truly positive or truly negative, were marked in green, whilst images with disagreement between the model and the human observer were marked in red. Images suggested misclassified tend to have lower DHC scores than correctly classified images. The shape of the DHC curves varies for each cell type, and the curves for Sertoli cells and

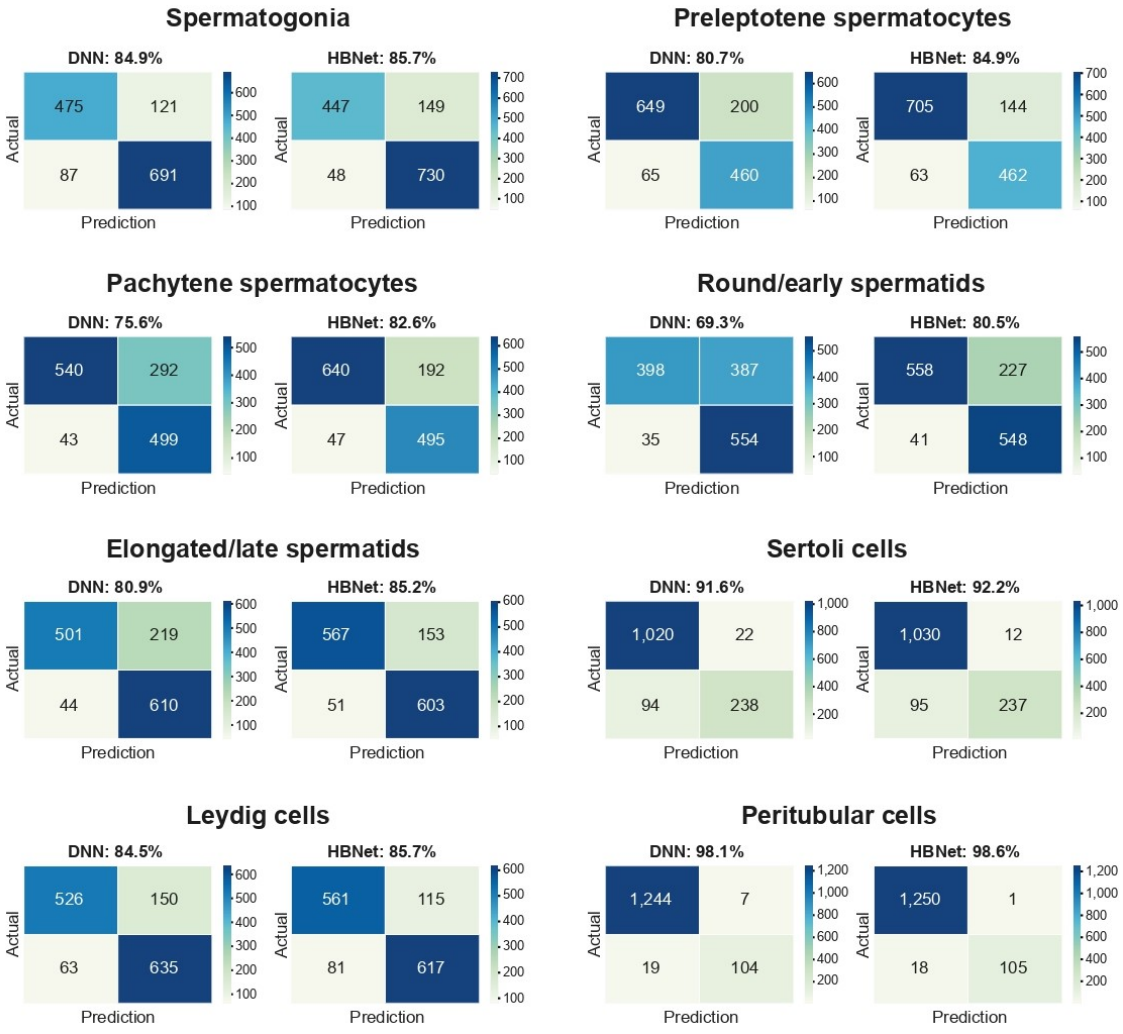


Fig. 5.12 Confusion matrix for each of the eight testicular cell types based on standard deep neural network (DNN) and hybrid Bayesian neural network (HBNet). Each quadrant shows the number of images that were true negative (upper left), false negative (upper right), false positive (bottom left) and true positive (bottom right), color-coded based on the number of images.

peritubular cells stood out as having a higher proportion of images with low DHC Scores than the other cell types. This is because staining in these cell types was less common (Figure 5.8 D), and cell types classified as lacking staining often have low DHC Scores. The spread of misclassifications determined the cut-off for reliable classification, which was marked as a blue line. Note that this cut-off was set at a DHC Score between 0.0 and 0.11 for all types except pachytene spermatocytes, round/early spermatids and elongated/late spermatids, for which it was set at 0.22, 0.78 and 0.22, respectively. The protein expression patterns of these three cell types showed a high correlation (Figure 5.8 A), suggesting that many proteins were co-expressed in these cells. Since they were not mutually exclusive, this may explain why the model would have more difficulties distinguishing these cell types from each other. Round/early spermatids are particularly challenging to distinguish manually from the transition into elongated/late spermatids. In the present investigation, there were only 67 images with expression restricted to round/early spermatids, while 254 images showed expression specific to elongated/late spermatids, and 212 images had expression in both of these two. This likely causes a particularly high DHC Score for round/early spermatids.

When only considering thresholded samples above the DHC cut-off, including classifications of high reliability, the classification accuracy of the HBNet model was substantially improved and considerably higher than all other classifiers. Table 5.4 shows model performance on a cell type-specific level. The % accuracy for predicting the labels for each cell type is shown for standard deep neural network (DNN) with only hand-crafted features, three standard classification approaches including our hybrid features (Multilabel k Nearest Neighbours, Random Forest Classifier and Support Vector Machines), our hybrid Bayesian neural network (HBNet), and DHC-thresholded HBNet (HBNet - DHC) along with the percentage of discarded images based on low DHC confidence. The standard deviation (std dev.) between each cross-validation fold is included for HBNet to indicate sampling variance.

The HBNet DHC-thresholded accuracy was >92% for all cell types except for round/early spermatids, which had an accuracy of 83.5%. For most cell types, approximately 30 to 39% of the images were below the DHC cut-off, except for peritubular cells where only 1.3% of the images were discarded, and Sertoli cells, where none were. Predictions above cut-off can be considered reliably annotated by the model, meaning manual annotation is only needed for, on average, 28.1% of the predictions. Note that there is a direct tradeoff for the choice of DHC threshold between accuracy and number of discarded images (Figure 5.14). Also note, that accuracy is an orthogonal measure of uncertainty. Similar performance to HBNet may sometimes be obtained with other deterministic classification methods, particularly if they have hybrid features as input. However, they do not provide the added value of confidence in their prediction, which enables the identification of images that can be automatically labelled.

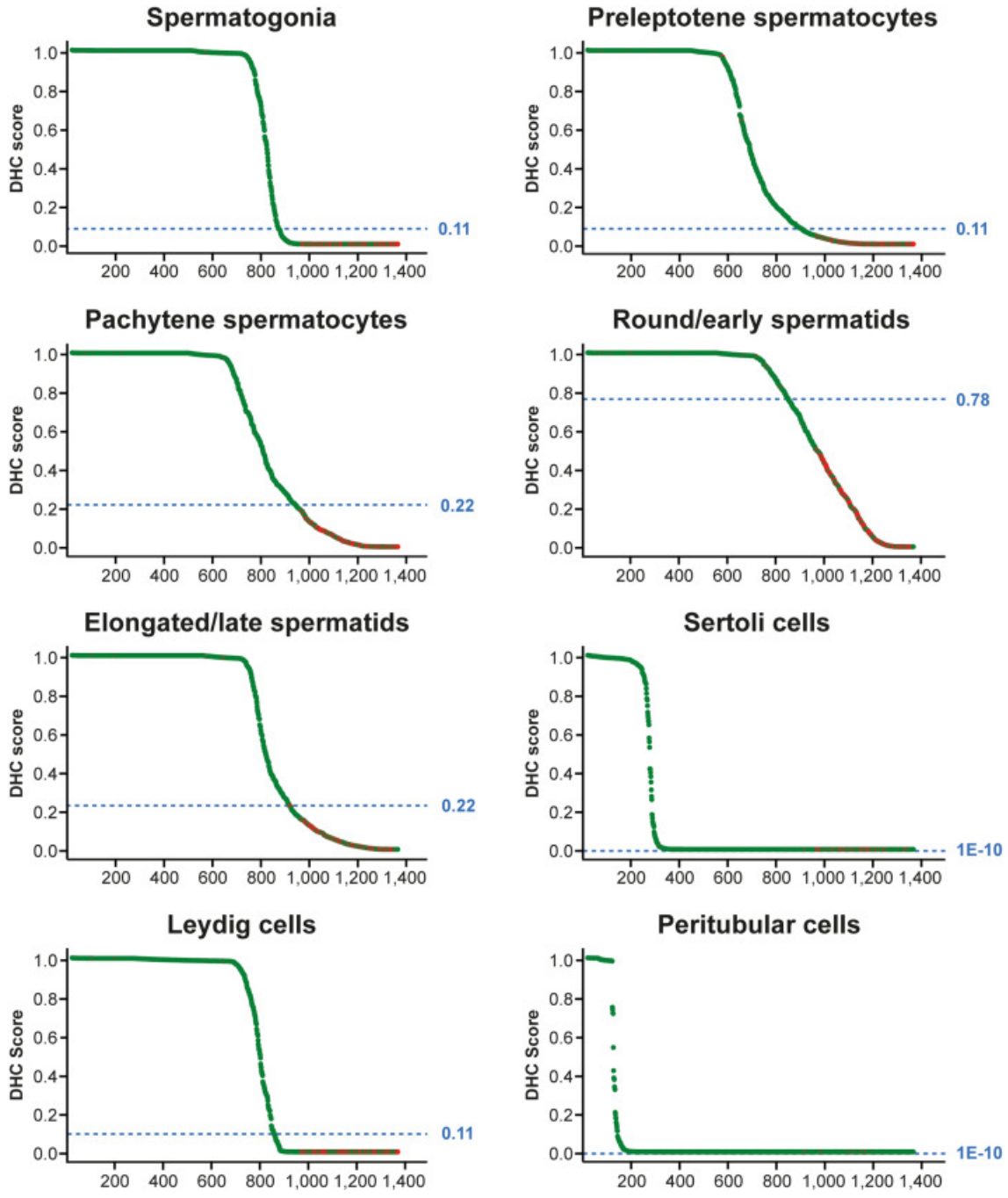


Fig. 5.13 Confidence maps of all automated predictions for each of the eight-cell types. Each dot corresponds to one prediction, with green = correct and red = incorrect. The predictions were sorted based on their DHC Score, showing the confidence in the prediction. The blue lines depict the determined cut-off for each cell type where classification is considered unreliable.

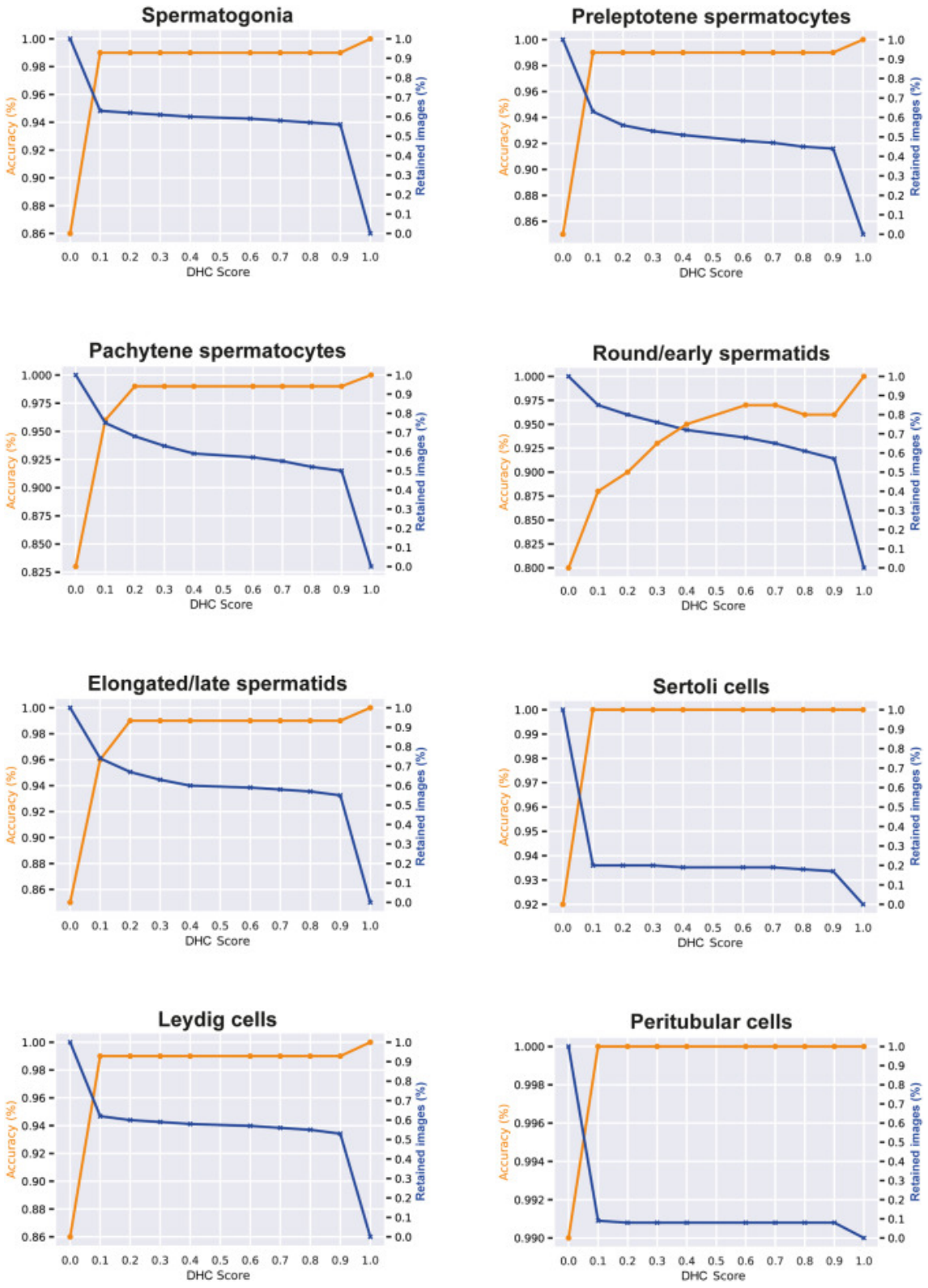


Fig. 5.14 Estimation of model certainty: Note that there is a direct tradeoff for choice of DHC threshold between accuracy and number of discarded images. Also note, accuracy is an orthogonal measure to uncertainty.

5.5.5 Evaluation of correctly classified and misclassified images

The DHC confidence metric allowed us to identify correctly classified images and images where the model disagreed with the human observer for one or several cell types.

In Figure 5.15, examples of correctly classified images are provided, i.e., these images were among the 67% that according to the Exact Match Ratio had all eight cell types annotated as either true positive or true negative - Heatmaps (left), IHC staining patterns (middle), with an overview of HBNet prediction and manual annotation of the eight different cell types (right). The colours of the heatmaps indicate where the HBNet model focuses on making a labelling decision from purple (no activation) through blue, green, yellow, to red (high activation). IHC images show positive staining in brown (a protein expressed) and counterstaining in blue (protein not expressed). Cell type names: Spermatogonia (SPG), preleptotene spermatocytes (Prel SPC), pachytene spermatocytes (Pach SPC), round/early spermatids (RE SPT), elongated/late spermatids (EL SPT), Sertoli cells (Sertoli), Leydig cells (Leydig) and peritubular cells (Peritub.). Green dots: Correct classification. Melanoma-associated antigen B18 (MAGEB18) and Synuclein beta (SNCB) showed selective expression in one cell type only, while Apoptosis associated tyrosine kinase (AATK) and T cell leukaemia translocation altered protein (TCTA) were expressed in several testicular cell types. MAGEB18 showed a speckled nuclear staining pattern in pachytene spermatocytes (arrows), with clearly visible nucleoli. SNCB was positive in elongated/late spermatids and sperm flagella (arrows), seen in the lumen of seminiferous ducts. AATK displayed cytoplasmic staining in pachytene spermatocytes (black arrows), round/early spermatids (white/black arrows) and Leydig cells (double-headed arrow). TCTA showed mainly cytoplasmic staining in Sertoli cells (arrows), Leydig cells (white/black arrows) and peritubular cells (double-headed arrows), in Sertoli cells accompanied with distinct positivity of nuclear membranes.

The images show that the model performed well both for proteins with distinct and selective staining and for more complex images where the protein was expressed in several cell types of varying intensity and staining patterns. The IHC stained images are presented along with heatmaps (Zhou et al., 2016) highlighting which area of the images that the model focused on for making the labelling decision. For the correctly classified images, it is evident that the model focused on several different areas within the image, including areas where cells were intact and well-represented.

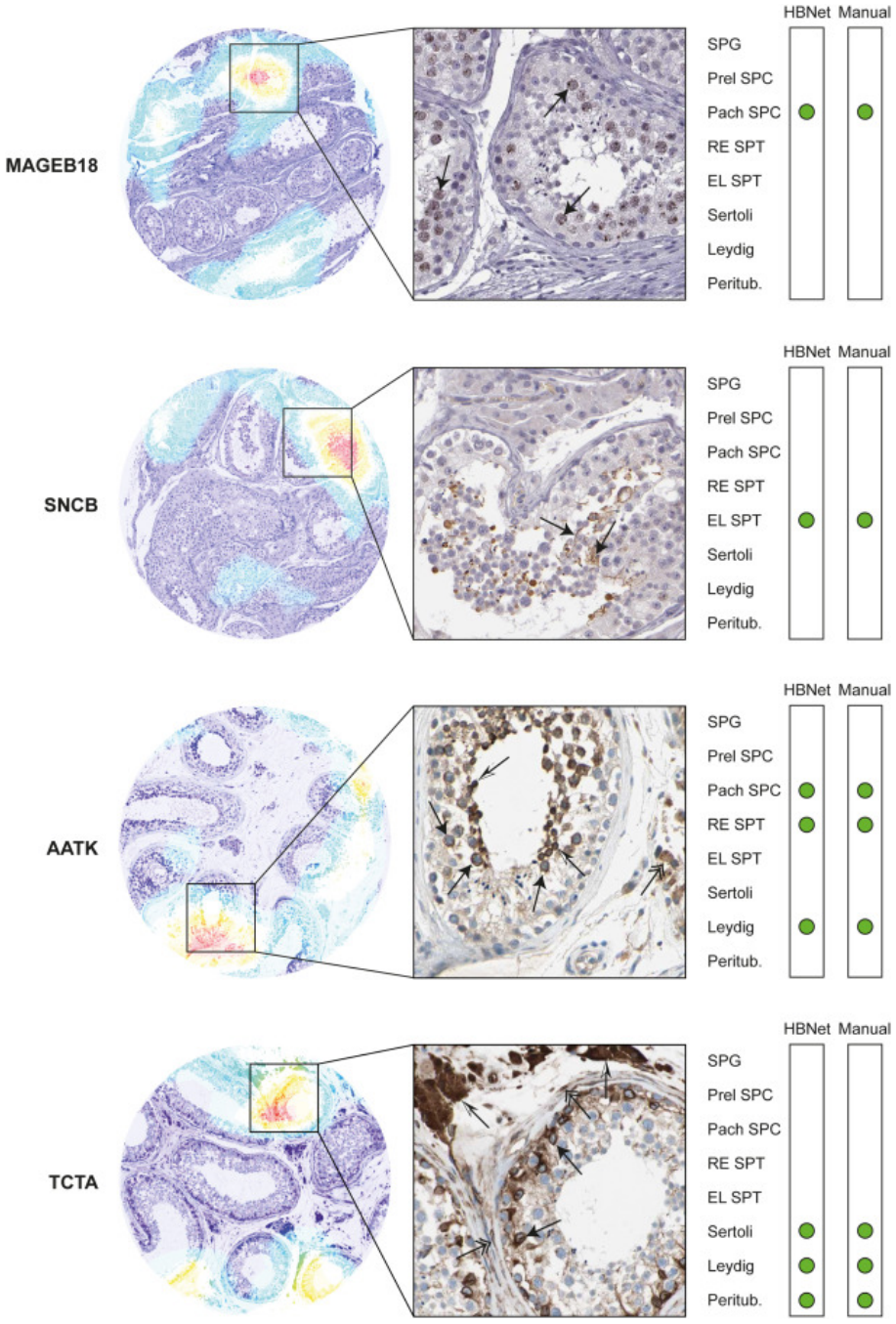


Fig. 5.15 Examples of correctly classified images.

In Figure 5.16, examples of incorrectly classified images are provided, i.e., Heatmaps (left) and IHC staining patterns (right), exemplified by one cell type each where HBNet prediction and manual annotation disagreed. The colours of the heatmaps indicate where the HBNet model focuses on making a labelling decision from purple (no activation) through blue, green, yellow, to red (high activation). IHC images show positive staining in brown (a protein expressed) and counterstaining in blue (protein not expressed). Cell type names: Spermatogonia (SPG), pachytene spermatocytes (Pach SPC), round/early spermatids (RE SPT), elongated/late spermatids (EL SPT), Sertoli cells (Sertoli) and Leydig cells (Leydig). Green dots: Correct classification. Orange dots: Correct classification but can be considered incorrect based on human knowledge. Red dots: Incorrect classification. (A) Polycomb group ring finger 3 (PCGF3) and SPANX family member D SPANXD represent manual errors. For PCGF3, the manual observer missed Sertoli cells that showed clear nuclear staining (arrows), while for SPANXD, Leydig cells had been annotated as positive, despite being completely negative (arrows). (B) FUN14 domain containing 2 (FUNDC2) and Minichromosome maintenance complex component 6 MCM6 showed staining neglected by the human observer. FUNDC2 displayed weak cytoplasmic positivity in spermatogonia (arrows), but due to strong staining in elongated/late spermatids (white/black arrow), the spermatogonia staining was considered unspecific. Similarly, MCM6 showed weak nuclear staining in pachytene spermatocytes, and was considered unspecific compared to the strongly positive preleptotene spermatocytes (white/black arrows). (C) The uncharacterized protein KIAA1324 and Spectrin repeat containing nuclear envelope family member 3 (SYNE3) were stained in small structures missed by the HBNet prediction. KIAA1324 showed positivity in small perinuclear structures of round/early spermatids, most likely representing centrosomes (arrows). SYNE3 was stained in nuclear membranes of Sertoli cells (arrows). (D) Leucine-rich repeat-containing 39 (LRRC39) and Rho related BTB domain containing 2 (RHOBTB2) correspond to images of poor quality. The area for which the HBNet model focused on for prediction of LRRC39 staining only contained unhealthy seminiferous ducts without the correct cell types. Similarly, RHOBTB2 had damaged seminiferous ducts where the cells had been separated from each other, and several cell types were missing.

Misclassified predictions included both falsely positive and falsely negative images and could be further divided into cases with high certainty (high DHC Score) and low certainty (low DHC Score). Several misclassified predictions represented clear errors made by the manual observer (Figure 5.16a). Such misclassifications often had high DHC Scores, and in these cases, the model can be used for identifying manual mistakes. Other misclassified predictions were due to unspecific staining deliberately neglected by the human observer (Figure 5.16b). Such stainings in need of further protocol optimization were often represented

by false-negative predictions with high DHC Scores, indicating that the model performed a correct prediction. However, based on experience, the positivity was interpreted as unspecific by the human observer. Some misclassified images corresponded to proteins expressed in small structures, including nuclear membranes, nucleoli or centrosomes (Figure 5.16c). Such staining patterns are rare and may be particularly challenging for the model to interpret due to limitations in the current pixel resolution. These predictions were often false positives with low DHC Scores. Finally, some misclassified images contained artefacts, such as damaged tissue sections or sections that contained areas where the testicular samples were not completely healthy (Figure 5.16d). Such misclassifications, both false positives and false negatives, often had low DHC Scores, and it was evident from the model heatmaps that the labelling decisions were mostly made on areas of the images where not all cell types were clearly represented, or the image/visible cells had poor quality.

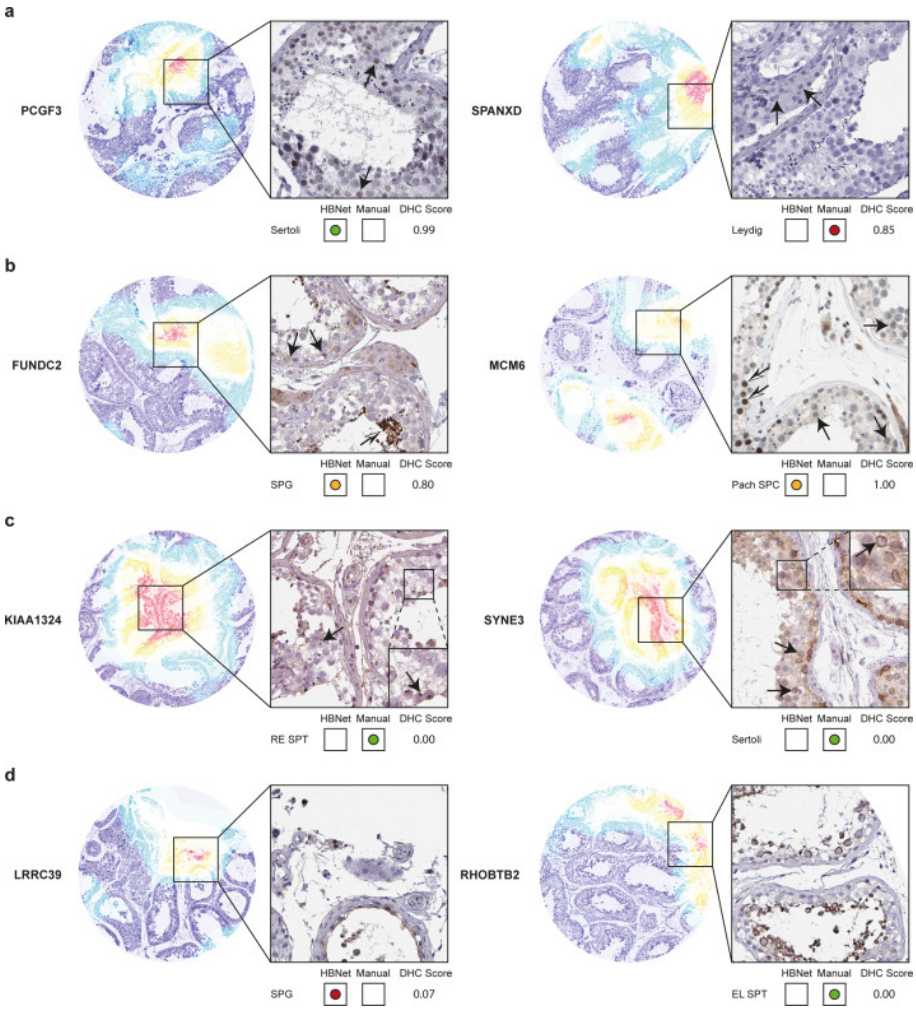


Fig. 5.16 Examples of misclassified images.

5.5.6 Model performance based on subcellular localisation and staining intensity

The manual annotation of the cell type-specific protein expression not only considered which cell types were positive but also in which subcellular organelle the staining was observed. In Table 5.5, the DHC-thresholded model performance in the test dataset is presented on a subcellular level. Similarly, as in the whole dataset (Figure 5.8 D), it was clear that some organelles were more common in certain testicular cell types, which may affect the overall accuracy, but it should also be noted that the patterns of different subcellular localisations appear differently in the various cell types based on the cell shape. In total, the best accuracy was found for staining patterns where all subcellular localisations (cytoplasmic, membranous and nuclear) were present. This is not surprising, as clearly outlining each cell structure increases the likelihood of the model identifying the correct cell types. Sertoli cells had lower accuracy of specific subcellular localisations compared to other cell types. The staining of Sertoli cells is challenging to interpret as these cells are situated in the interspace between the germ cells, and staining may be difficult to distinguish from other cell types.

In addition to cell type-specific pattern and subcellular localisation of the staining, the human observer also considers the intensity of the staining. This somewhat subjective measurement that determines the brown saturation level is considered to represent the amount of protein expression ranging from low levels (weak staining/beige colour) through moderate levels (medium brown) to high levels (dark brown/black). As shown in Table 5.6, it is evident that the DHC-thresholded accuracy did not depend on staining intensity, and there was no significant improvement in predictions performed on distinctly stained cells compared to those that showed more faint positivity.

Metrics	**Neural Network**	**Multi-label k Nearest Neighbours (Hybrid Features)**	**Random Forest Classifier (Hybrid Features)**	**Support Vector Machine (Hybrid Features)**	**Hybrid Features DNN (%)**	**Hybrid Features BNN HBNet (%)**
Hamming Loss (↓)	17	13	15	14	17	13
Macro F1 Score (↑)	77	82	77	81	81	84
Micro F1 Score (↑)	78	83	79	81	80	84
Exact Match ratio (↑)	41	70	47	61	48	67
mean-Average Precision (mAP) (↑)	70	73	69	72	71	76

Table 5.3 Overall model performance.

	Model Performance Accuracy (%)						Discard Tradeoff
Cell type	**DNN**	**MLkNN**	**RFC**	**SVM**	**HBNet (std. dev.)**	**HBNet- DHC**	**HBNet - DHC % Discarded**
Spermatogonia (0.11)	85.9	83.8	80.2	81.2	85.7 (0.24)	99.4	37.20%
Preleptotene spermatocytes (0.11)	74.8	85.5	71.9	73.1	84.9 (0.38)	99.2	37.20%
Pachytene spermatocytes (0.22)	69.9	82.4	72.1	73.3	82.6 (0.24)	99.2	31.70%
Round/early spermatids (0.78)	68.1	79	72.8	74.3	80.5 (0.55)	83.5	39.10%
Elongated/late spermatids (0.22)	77.4	79.8	76.9	76.7	85.2 (0.36)	98.7	30.10%
Sertoli cells (1.00E-10)	74.1	86.3	65.9	57.6	92.2 (0.16)	92.2	0.00%
Leydig cells (0.11)	80.4	81.6	80.3	76.2	85.7 (0.34)	99.2	38.30%
Peritubular cells (1.00E-10)	84.3	95.9	67.4	67.9	98.6 (0.09)	98.7	1.30%

Table 5.4 Model performance on a cell type-specific level.

Cell type	#DHC-thresholded labels / #actual labels with subcellular localization)	HBNet - DHC % accuracy (#labels)						
		Cyt	Cyt, Mem	Mem	Nucl	Nucl, Cyt	Nucl, Cyt, Mem	Nucl, Mem
Spermatogonia	518/521	99.5 (204/205)	97.6 (41/42)	100.0 (5/5)	99.5 (201/202)	100.0 (40/40)	100.0 (25/25)	100.0(2/2)
Preleptotene spermatocytes	357/360	100.0 (121/121)	100.0 (30/30)	100.0(2/2)	98.17 (161/164)	100.0(15/15)	100.0(28/28)	0
Pachytene spermatocytes	388/391	99.3 (145/146)	100.0 (66/66)	100.0 (4/4)	98.5 (135/137)	100.0 (9/9)	100.0 (29/29)	0
Round/early spermatids	361/362	99.2 (131/132)	100.0 (57/57)	0	100.0 (147/147)	100.0 (10/10)	100.0 (16/16)	0
Elongated/late spermatids	405/409	98.2 (215/219)	100.0 (83/83)	0	100.0 (67/67)	100.0 (22/22)	100.0 (18/18)	0
Sertoli cells	225/231	97.8 (87/89)	100.0 (31/31)	100.0 (6/6)	95.2 (79/83)	100.0 (1/1)	100.0 (11/11)	100.0 (10/10)
Leydig cells	466/470	98.9 (277/280)	100.0 (71/71)	100.0 (5/5)	100.0 (81/81)	100.0 (25/25)	100.0 (7/7)	0
Peritubular cells	105/120	89.3 (50/56)	100.0 (9/9)	83.7 (46/55)	0	0	0	0
Average all cell types	*2825/2864*	*97.78*	*99.7*	*97.28*	*98.77*	*100*	*100*	*100*

Table 5.5 Model performance based on subcellular localization.

Cell type	HBNet - DHC % accuracy (#DHC-thresholded labels / #actual labels)		
	Only weak labels (intensity =1)	**Only moderate labels (intensity = 2)**	**Only strong labels (intensity = 3)**
Spermatogonia	100.0 (27/27)	99.3 (142/143)	99.4 (349/351)
Preleptotene spermatocytes	100.0 (49/49)	100.0 (150/150)	98.14 (158/161)
Pachytene spermatocytes	100.0 (70/70)	99.3 (141/142)	98.9 (177/179)
Round/early spermatids	100.0 (53/53	100.0 (102/102)	99.6 (206/207)
Elongated/late Spermatids	100.0 (41/41)	97.3 (145/149)	100.0 (219/219)
Sertoli cells	98.6 (72/73)	86.1 (31/36)	100.0 (122/122)
Leydig cells	98.9 (172/174)	99.5 (202/203)	100.0 (92/92)
Peritubular cells	100.0 (17/17)	84.5 (49/58)	86.7 (39/45)
Average all cell types	*99.7*	*95.8*	*97.9*

Table 5.6 Model performance based on staining intensity

5.5.7 Discussion

The point predictions were combined with a confidence score (DHC), generated by a Monte-Carlo DropWeights method in conjunction with an approximate BNN with hybrid image features. The proposed model architecture showed outstanding performance in both simple images with clear cell type-specific staining and more complex images where several cell types showed positivity of varying intensity and staining patterns. In addition, the novel DHC Score adds another level of insight, particularly important for challenging cases where uncertain predictions can be highlighted.

Weaknesses of an automated algorithm may be related to the fact that manual annotation is not only based on visual examination of staining intensity but to a large extent also relies on experience, where the manual observer takes into consideration staining protocol, overall image quality, artefacts and previous literature on the protein being analyzed. As a result, unspecific staining may be neglected by the human observer, especially when accompanied with distinct staining in other structures that more likely represents the true protein expression. However, challenges related to tissue processing, IHC staining procedure, and experience in identifying artefacts are overcome in the presented framework, as uncertain predictions will be highlighted.

Our proposed HBNet showed high accuracy for all eight cell types for samples generated by the same laboratory, with increased accuracy after applying a DHC Score threshold and improving the overall accuracy from 86.9 to 96.3%, and revealing which images that the model reliably classifies. When examining images above and below this threshold, it was evident that many images for which the model faced challenges constituted images expected to be particularly difficult, often due to the reasons described above. Three cell types needed a higher DHC Score threshold for reliable prediction: pachytene spermatocytes, round/early spermatids and elongated/late spermatids. This is not surprising, as these cells correspond to the most common combination for proteins co-expressed in more than one testicular cell type, as described previously (Pineau et al., 2019).

The suggested workflow can be developed further for other organs in the future, but already now, the method can cover the entire dataset of testis images corresponding to in total >15,000 proteins stained with IHC as part of the HPA project. The workflow can also be used in other large-scale projects that focus on distinguishing between healthy and diseased tissues, widely applicable to e.g. cancer research and routine diagnostics if retrained specifically on datasets from other laboratories. The daily pathology workflow largely depends on manual microscopic evaluation of tissue sections, which may not only lead to a delayed disease diagnosis with potential worsened patient prognosis but also to a false diagnosis (Goodman

et al., 2012). Further advances in the automated annotation of histological sections are therefore clearly warranted.

5.6 Uncertainty Quality Matrices

In this subsection, we review related and commonly accepted uncertainty quality metrics. We have evaluated the quality of uncertainty estimates using four statistical matrices: Predictive Log-Likelihood (PLL), Negative Log Predictive Density (NLPD), Brier Score (BS) and Root Mean Squared Error (RMSE).

1. Negative Log Predictive Density (NLPD): Negative Log Predictive Density takes the negative logarithm of the posterior class probabilities for classification and the predictive density for regression. This predictive performance penalises both over and under-confident predictions but in general favours conservative models, that is, models that tend to be under-confident rather than over-confident. This scoring rule can only be used to compare the quality of predictive uncertainty between different models' performance on the same dataset and are not transferable.

$$\text{NLPD (L)} = -\frac{1}{N}\sum_{i=1}^{N} logp(y_i = c_i|x_i) \tag{5.8}$$

   NLPD infinitely penalises wrong predictions made with zero uncertainty.

2. Root Mean Squared Error (RMSE): The Root Mean Squared Error (RMSE) is the standard deviation of the prediction errors. The higher the value, the greater the uncertainty.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{5.9}$$

3. Predictive Log-Likelihood (PLL): Predictive Log-Likelihood is a widely accepted metric as a marker for the quality of uncertainty, used as the primary uncertainty quality metric in (Hernández-Lobato et al., 2016; Teye et al., 2018). It captures how well a model fits the data. The fundamental property is that PLL makes no assumptions about the form of the predictive distribution. PLL has no upper bound, so larger values indicate a better model fit. While PLL is an elegant measure, outliers have a negative effect on the score.

   The PLL can be defined for test image (x_i, y_i), where F is cumulative distribution function (CDF) of the prediction and $\hat{w}_j$ is the parameter from posterior distribution of T stochastic feed-forward as below:

$$\mathrm{PLL}(f_w(x),(y_i,x_i)) = \log p(y_i|f_w(x_i)) = \log \int f_w(x_i,y_i)p(w|D)dw \tag{5.10}$$

$$\approx \log \int f_w(x_i,y_i)q_\theta(w)dw \approx \log \frac{1}{T}\sum_{j=1}^{T} p(y_i|f_{\hat{w}_j}(x_i)) \tag{5.11}$$

4. Brier Score (BS): The Brier score is a score function that measures the accuracy of probabilistic predictions. It calculates the mean squared difference between a binary label and its associated predicted probability. Therefore, the lower the Brier score in multi-class classification, the better the predictions are calibrated. (Lakshminarayanan et al., 2017).

$$BS = \frac{1}{n}\sum_{j=1}^{C}\sum_{i=1}^{n}(\hat{p}_{ij} - x_{ij})^2 \tag{5.12}$$

, where C is the number of classes, $\hat{p}_{ij}$ is the estimated probability of class j in trial i, and x_{ij} is 1 or 0, depending on the occurrence of class j in trial i. The normalization by n guarantees a value in [0,2].

Metrics	**Ensemble MCDW**	**Ensemble MCDO**	**MCDW**	**MCDO**
Negative Log Predictive Density (NLPD) (↓)	43.62	3.59	123.80	1313.34
Root Mean Square Error (RMSE) (↓)	0.55	0.61	0.58	0.60
Predictive Log Likelihood (PLL) (↓)	0.25	0.27	0.27	0.28
Brier Score (BS) (↓)	0.66	0.71	0.65	0.70

Table 5.7 Quality Metrics

Our experimental results (Table 5.2 and 5.7) show that Ensemble MC-DropWeights improves prediction accuracy under estimated uncertainty. More importantly, the uncertainty quality metrics show a significant improvement when using Ensemble MC-DropWeights.

5.7 Conclusion

This chapter provides a Bayesian perspective for Neural Networks applications in multi-class image classification and demonstrates the benefits and applicability of uncertainty leveraging MC-Dropweights-based estimated uncertainty for Deep Learning in disease detection from MRI images.

In this chapter, we develop the first deep learning study (to the best of our knowledge), which quantifies uncertainty and model interpretability in multi-label classification and applies it to the problem of recognising proteins expressed in testes cell types based on immunohistochemically stained images. Multi-label classification is achieved by thresholding the class probabilities, with the optimal thresholds adaptively determined by a grid search scheme based on Matthews Correlation Coefficients (MCC). Our experimental results show that the MC-Dropweights visibly improve performance to estimate uncertainty compared to current approaches.

We present a novel method for automated annotation of immunohistochemistry images, combining the predictions with an uncertainty metric, the DeepHistoClass (DHC) confidence score, to improve the accuracy of automated image predictions and identification of manual identification annotation errors. The suggested streamlined framework constitutes an important approach for accurate large-scale efforts mapping the human proteome such as the HPA project and holds promise for both research and diagnostics, aiming at analyzing the spatiotemporal expression of human proteins in health and disease.

Chapter 6

Cost-Sensitive Calibrated Uncertainty in Medical Decision Making

"Apprehension, uncertainty, waiting, expectation, fear of surprise do a patient more harm than any exertion. Remember, he is face to face with his enemy all the time. To be 'in charge' is certainly not only to carry out the proper measures yourself but to see that everyone else does so too." - Florence Nightingale (1820 - 1910)

This chapter is focused on the methodological and algorithmic contributions for the cost-sensitive calibrated uncertainty problem. Reliable and cost-sensitive calibrated estimated uncertainty in deep learning is important in many real-world applications where safety is critical, and prediction problems are asymmetric in the sense that different types of misclassification errors incur different costs. However, uncertainty obtained by approximate inference techniques, such as variational inference cannot guarantee optimal predictions to represent the model error and is prone to miscalibration (and often poor calibration) due to the assumption of the constant cost of misclassification, which is not realistic in medical diagnosis.

This chapter proposes a variational inference with Monte Carlo Dropweights based Bayesian neural networks model, which means cost-sensitive calibrated predictive uncertainty can be estimated while minimising asymmetric cost as an expected utility function with improved accuracy. We have highlighted potential issues in commonly used performance metrics, uncertainty calibration measures, the quality of the estimated uncertainty and proposed revised evaluation metrics to mitigate them.

6.1 Introduction

Advances in deep learning have achieved state-of-the-art performance in medical image analysis such as detection and localisation, segmentation, registration, classification and prediction of treatment outcomes (Altaf et al., 2019; Leibig et al., 2017; Litjens et al., 2017). In real-world safety-critical applications such as assessing the degree of disease severity in medical image analysis, prediction problems are asymmetric as different types of misclassification errors incur different costs or significant losses. So overconfident incorrect predictions and consequently reaching a false conclusion may result in the loss of life in some circumstances.

Neural networks trained by minimising a cross-entropy loss tend to overfit based on classification accuracy. The Bayesian Neural Networks provides a natural and principled way of modelling uncertainty with a prior distribution on its weights, which is robust to over-fitting (i.e. regularisation). However, exact inference is analytically intractable, and hence the approximate inference has been applied instead. Approximate inferences such as Variational inference (Blundell et al., 2015; Gal, 2016; Ghoshal et al., 2019b; Graves, 2011) and Markov chain Monte Carlo (MCMC) (Neal, 2012) which approximates the posterior distribution in Bayesian neural networks (BNN), are prone to miscalibration, and the estimated uncertainty does not always represent the error in model prediction. This can lead to overconfident predictions, which raise concerns over its safety in critical applications (Cobb et al., 2018). For example, in practical decision-making systems, the cost of falsely misdiagnosing a disease when a patient is infected (i.e. a false negative) may be much higher than incorrectly diagnosing a disease when it is not present (false positive). Current approaches to approximate Bayesian Deep learning assumes an equal cost for classification errors. Therefore, deep learning models are poorly calibrated at quantifying predictive uncertainty, i.e. the mismatch between a model's uncertainty and its error.

In cost-sensitive classification (Elkan, 2001; Kukar et al., 1998; Turney, 1994; Zhou and Liu, 2010) considers the varying costs of different misclassification types. The goal of cost-sensitive supervised learning is to minimise the total cost or maximise utility function. A cost matrix encodes the penalty of classifying samples from one class as another. Bayesian optimal decision can help obtain the cost-sensitive prediction. The below equation shows the predicted class label that reaches the lowest expected cost:

$$y_{pred} = argmin_{1<c<=C} \sum_{i=1}^{C} P(y=i|x,W,b)K(j,i) \tag{6.1}$$

where $K(j,i)$ denotes the cost of predicting a sample from class j as class i. C is the total number of classes. The cross diagonal elements in the utility matrix are the weights of corresponding categories, others are zero. Larger value in the cost matrix impose larger penalty. The $P(y=i|x,W,b)$ is to estimate the probability of class i given x.

In practice, several non-parametric and parametric calibration approaches such as isotonic regression, Platt scaling, temperature scaling (TS) (Guo et al., 2017) or parametric multi-class Dirichlet calibration have been extensively studied in neural networks (Kull et al., 2019).

Given $\hat{y}$ is the class prediction of the model and $\hat{p}$ is its associated confidence, a model is calibrated only if confidence in a prediction matches its probability of correctness (Wenger et al., 2019):

$$E\left[1_{\hat{y}=y} \mid \hat{p}\right] = \hat{p}, \tag{6.2}$$

Expected Calibration Error (ECE) is a well-known measure of the degree of calibration to quantify the miscalibration of the difference in expectation between confidence and accuracy, i.e. weighted average over the absolute accuracy/confidence difference. (Guo et al., 2017; Naeini et al., 2015; Wenger et al., 2019).

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{accuracy}(B_m) - \text{confidence}(B_m)| \tag{6.3}$$

, where n is the number of samples and B_m is the set of indices of samples whose uncertainty falls into the interval.

The maximum calibration error (MCE) can be defined as (Wenger et al., 2019)

$$\text{MCE} = \max_{p\in[0,1]} |\text{accuracy}(B_m) - \text{confidence}(B_m)|. \tag{6.4}$$

Some existing works have studied cost-sensitive neural networks (Elkan, 2001; He and Garcia, 2009; Kukar et al., 1998) but none of them has focused on cost-sensitive uncertainty estimation in deep learning to the best of our knowledge.

Bayesian decision theory provides a principled approach for optimal decision making under uncertainty given a task-specific utility function (θ, a) over actions $a \in A$, which extends the Bayesian paradigm. This maximises the expected utility over the posterior to make rational predictions in state θ. The overall process is computed using a 2-step procedure: probabilistic inference and optimal prediction. First, approximate the posterior $p(\theta|D)$ with a $q(\theta|D)$ and then minimise evidence lower bound (ELBO) loss that incorporates the network weights and task-specific utility function under q, where we assume that approximate q measures properties of the posterior. A clearly defined goal of a prediction is necessary as an evaluation criterion in the form of a utility function. Therefore, this should jointly optimise

the approximate posterior with the action that maximises the expected utility with respect to the posterior over the model parameters, which will minimise the posterior risk.

Cobb et al. (Cobb et al., 2018) showed that minimising the KL divergence between an approximate posterior q and a calibrated posterior scaled by the utility function results in the standard evidence lower bound (ELBO) loss for Bayesian neural network inference, as well as an additional task-specific utility function, dependent regularisation term, to stochastic optimisation. This can be implemented as a novel penalty term to the standard neural network.

Following Gal (Gal, 2016), Ghoshal et al. (Ghoshal et al., 2019a) showed that Neural Networks with dropweights applied in the fully connected layer is equivalent to variational Bayesian neural networks. "Dropweights", which randomly drops connections, where weights in the neural networks are set to zeros during both training and inference robust to over-fitting and can be seen as a form of regularisation (Ghoshal et al., 2019a).

We extended the classic technique to 'approximate inference for the loss-calibrated Bayesian framework' (Cobb et al., 2018; Lacoste-Julien et al., 2011) for dropweights based cost-sensitive neural networks, revising backpropagation learning classification procedures that attempt to reduce the cost of misclassified samples, rather than the number of misclassified samples to represent the model error. We encoded asymmetries due to the different types of misclassification errors or probability of occurrence of different classes, i.e. class-imbalanced scenarios in the form of a utility function. We obtained calibrated predictive uncertainty for applications with an asymmetric cost by maximising the utility function (i.e. minimising asymmetric costs) in the backpropagation learning procedure. We have investigated potential issues in commonly used performance metrics, calibration measures, the quality of the estimated uncertainty and propose revised metrics to mitigate them. Furthermore, we show that decisions informed by cost-calibrated uncertainty can improve diagnostic performance to a greater extent than straightforward alternatives.

In experiments, we show the correlation between error in prediction and estimated uncertainty. We propose Maximum Uncertainty Calibration Error (MUCE) as a metric to measure calibrated confidence and its prediction, especially for high-risk applications, where the goal is to minimise the worst-case deviation between error and estimated uncertainty.

Expected Calibration Error (ECE) and Uncertainty Calibration Error (UCE) (Laves et al., 2019a) cannot expose large calibration errors even in the low uncertainty area and are not flexible enough to deal with non-uniform confidence distribution due to the underlying binning strategy. Therefore, we propose Adaptive Expected Calibration Error (AECE) and Adaptive Uncertainty Calibration Error (AUCE) as Measures of uncertainty calibration in deep learning.

In practical decision-making systems, the cost of falsely misdiagnosing a disease when a patient is infected (i.e. a false negative) may be much more significant than incorrectly diagnosing a disease when it is not present (false positive). Therefore, it is important to consider the cost of every type of misclassification error instead of calculating the likelihood of a model trained by minimising the negative log-likelihood (i.e. cross-entropy) loss by considering all types of errors equally.

Coronavirus (COVID-19) represents a new strain of Coronavirus and presumably a mutation of other Coronaviruses (Shan+ et al., 2020). Dealing with it is currently a significant medical challenge around the world. Unfortunately, the existing dataset, which consists of limited image data sources with the expert labelled data set, for detecting COVID-19 positive patients is insufficient, and manual detection is time-consuming. Our goal is to provide a reliable Deep Learning-based solution combined with clinical practices to provide automated detection with estimated bias-reduced well-calibrated uncertainty to aid the screening process.

We evaluated the effectiveness of our approach to detecting Covid-19 from X-Ray images (Ghoshal and Tucker, 2020a,b, 2021). Experimental results show that our method reduces miscalibration considerably without impacting the model's accuracy and improves the reliability of computer-based diagnostics.

6.2 Cost-Sensitive Calibrated Approximate Bayesian Inference Method

6.2.1 Bayesian Neural Network

Given $D = \left\{X^{(i)}, Y^{(i)}\right\}$ where $X \in R^d$ is a d-dimensional input vector and $Y \in \{1 \dots C\}$, given C class label, a set of independent and identically distributed (i.i.d.) training samples size N, a BNN is defined in terms of a prior $p(w)$ on the weights, as well as the likelihood $p(D|w)$. Correspondingly, x_i and y_i refer to single input and class label. Besides, the variables w and Z are respectively network parameters and latent variables of X .

Bayesian neural network (Neal, 1993) places a prior $p(w)$ over the parameters w of a network so that it is able to take into account the uncertainty of a network. After training such a network, a posterior over the weights w should be inferred:

$$p(w|Y,X) = \frac{p(Y|X,w)p(w)}{p(Y|X)} \tag{6.5}$$

where $p(Y|X) = \int_w p(Y|X,w)p(w)dw$.

In particular, the posterior distribution s necessary and sufficient information for making optimal decisions under uncertainty (Berger 1985). Unfortunately, the integration over $p(w)$ in $p(Y|X)$ is intractable due to the non-linear property of $p(Y|X,w)$.

Variational Bayesian methods approximate the true posterior by maximising the evidence lower bound (ELBO) between a variational distribution $q(w|\theta)$ and the true posterior $p(w|D)$ w.r.t. to θ. The corresponding optimisation objective or cost function is

$$L(D,\theta) = E_{q(w|\theta)} \log p(D|w) - \mathrm{KL}(q(w|\theta) \,||\, p(w|Y,X)) \tag{6.6}$$

The first term is the expected value of the likelihood w.r.t. the variational distribution and is called the likelihood cost. The second term is known as Kullback-Leibler (KL) divergence between the variational distribution $q(w|\theta)$ and the prior $p(w)$ and is called the complexity cost. So, a variational distribution $q(w)$ is exploited to approximate the posterior distribution $p(w|Y,X)$ with a KL divergence:

$$KL(q(w) \,||\, p(w|Y,X)) = -\mathbb{E}_{q(w|\theta)} \log \frac{p(Y|X,w)p(w)}{q(w)} dw + const. \tag{6.7}$$

Using Monte Carlo DropWeights approximating distribution $q(w)$, Eq. 6.7 can be written as the standard objective loss of a neural networks with dropout with an additional penalty term:

$$KL(q(w) \,||\, p(w|Y,X)) = -\sum_i \log p(y_i|x_i,\hat{w}_i) + ||w||^2 + const. \tag{6.8}$$

6.2.2 Calibrated Bayesian Neural Network

Bayesian decision theory (Berger 1985) defines a rigorous framework for decision-making under uncertainty in prediction.

When we have access to the true posterior, we can think of probabilistic inference as averaging over the model parameters w to infer a predictive distribution $p(y^*|x^*,Y,X)$ is defined as:

$$p(y^*|x^*,Y,X) = \int_w p(y^*|x^*,w)p(w|Y,X)dw. \tag{6.9}$$

The fundamental concept of cost-sensitive calibration in the case of variational approximation is to include a lower bound to the marginal likelihood $\log p(D)$ with a separate term accounting for the loss. An optimal decision h and utility $u(w,h) >= 0$ defined over the model parameters w. The function $u(w,h)$ is the loss $l(w,h)$ transformed to utility (Berger 1985). This augmented objective is maximised, typically with an alternating algorithm, with

respect to both the parameters α of the approximation and the decisions h. The optimisation is tightly coupled because the decisions influence the approximation and vice versa. However, it turns out that integration over w is intractable. While still retaining a reasonable posterior approximation instead of maximising the approximation accuracy, Lacoste-Julien et al. (Lacoste-Julien et al., 2011) proposed a loss-calibrated approximate inference to maximise the expected utility computed over the approximating distribution.

Variational inference approximate the posterior $p(w|D)$ with $q_\lambda(w)$ parameterised by λ typically by maximising a lower bound $L_{VI}(\lambda)$ for the marginal log-likelihood. So, a variational distribution $q(w)$ is introduced to obtain the lower bound of $\log D(Y,X)$:

$$\log D(Y,X) = \log \int_w q(w) \frac{D(Y,X|w)p(w|Y,X)}{q(w)} dw \tag{6.10}$$

$$\geq \int_w q(w) \log \frac{D(Y,X|w)p(w|Y,X)}{q(w)} dw \tag{6.11}$$

$$= L(q(w)). \tag{6.12}$$

After applying Jensen's inequality, we obtain the calibrated evidence lower bound (ELBO) $L(q(w))$.

So, maximizing the calibrated ELBO $L(q(w))$ is equivalent to maximize the ELBO of traditional BNN (shown in Eq. 6.13) with a new term. Meanwhile, the calibrated evidence lower bound $L(q(w))$ can be further expanded as:

$$L(q(w)) = \underbrace{\int_w q(w) \log D(Y,X|w)}_{\text{new term}} - \underbrace{KL(q(w)||p(w|Y,X))}_{\text{ELBO in BNN (same as Eq. 6.8)}} \tag{6.13}$$

Here, the loss function is divided into two terms. The utility-dependent first term accounts for decision making. It is independent of the observed y and only depends on the current approximation $q_\lambda(w)$, favouring approximations that optimise the utility. The second term is analogous to the standard variational approximation to provide the final bound.

By applying Monte Carlo dropweights approximating distribution $q_\lambda(w)$ and the reparameterization trick (Kingma et al., 2015) the term $\int_w q(w) \log D(Y,X|w)$ in Eq. 6.13 can be approximated by a loss function and optimize it using Adam optimizer.

6.3 Backpropagation algorithm in cost-sensitive learning

Backpropagation algorithm optimizes the gradient of the error function for a given value of input vector using gradient descent by the chain rule. Normally, the backpropagation algorithm minimizes the squared error (or loss) of the network:

$$L = \sum \frac{1}{2} \sum_{i \in output} (y_i - p_i)^2 \tag{6.14}$$

where p_i is the predicted value of the $i-$th neuron and y_i is the actual output.

The misclassification cost is a function of the predicted class and the actual class. This function, cost (actual class, predicted class), is represented as a cost matrix. It is an additional input to the learning procedure and is also used to evaluate the ability of the trained neural network to reduce misclassification costs (Kukar et al., 1998). The cost matrix is defined as follows:

- $Cost[i, j]$ = cost of misclassification an example from "class i" as "class j"
- $Cost[i, i] = 0$ (cost of correct classification)

The cost matrix represents the expected misclassification cost of a sample that belongs to the i-th class:

$$CostMatrix[i] = \frac{1}{1 - P(i)} \sum_{j \neq i} P(j) Cost[i, j] \tag{6.15}$$

$P(i)$ is an estimate of the prior probability that the sample belongs to $i-$th class. In the equal-error cost case we have the uniform cost matrix: $CostMatrix[i] = 1$.

The total misclassification cost should be minimised in cost-sensitive learning, given the cost matrix. The backpropagation algorithm updates the weights w_{ji} with the delta rule by computing the local gradients δ's and proceeds backwards, starting with the output layer, layer by layer. Each neuron from the output layer of the network represents one of the possible classes. The normalised output can be viewed as an estimate of the probability $P(i)$ that the sample belongs to $i-$th class.

$$P(i) = \frac{p_i}{\sum_{i \in output} p_j} \tag{6.16}$$

Instead of minimising the squared error, the cost-sensitive modifications of the backpropagation algorithm minimise the misclassification costs by changing the error function (Kukar et al., 1998). The loss function is corrected by introducing the fact $U[i, j]$, i = expected class, j = actual class:

$$L = \sum_{p \in examples} \frac{1}{2} \sum_{i \in output} (y_i - p_i)^2 . U[class(p), i]^2 \tag{6.17}$$

The factor $U[i, j]$ is defined as:

- $U[i, j] = CostMatrix[i], i = j$
- $U[i, j] = Cost[i, j], i \neq j$

The behaviour of the backpropagation algorithm in the equal-error or uniform case remains the same.

The $U[i, j]$ is a constant factor in the partial derivatives of the error function in the derivation of the backpropagation algorithm. So the delta rule that takes in account the misclassification cost can be written as follows (c is the expected class of the current training sample):

- $\delta = (y_j - p_j).p_j(1 - p_j).U^2[c, j]$ for output neurons
- $\delta = p_j(1 - p_j)\sum_k \delta_k w_{kj}$ for hidden neurons

The δ factor for output neurons is normalized to ensure the convergence of the modified backpropagation algorithm as:

$$\delta^l = \frac{\delta}{max_{i,j} U[i, j]^2} \tag{6.18}$$

6.4 Measure of Uncertainty Calibration in Deep Learning

We describe the most prevalent methods to measure the miscalibration of estimated uncertainty associated with the classification. Our cost-sensitive calibrated BNN model incorporates asymmetric misclassification costs as a utility function to enable rejection of uncertain predictions and so, in turn, minimises the misclassification, resulting in improved performance, incorporating practical considerations.

6.4.1 Uncertainty Calibration Error (UCE)

We leverage the following modified notion of Eq. (6.3) for bias-reduced Uncertainty Calibration Error (UCE) to measure the degree of calibration to quantify the miscalibration of the difference in expectation between model error and estimated uncertainty (Laves et al., 2019a).

The estimated bias-reduced uncertainty (Ghoshal et al., 2020) of a neural network is partitioned into M equally-spaced bins (each of size $1/M$), and a weighted average of the bin error and uncertainty difference. Mathematically, approximated Uncertainty Calibration Error (UCE) is defined as (Laves et al., 2019a):

$$\text{UCE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{error}(B_m) - \text{uncertainty}(B_m)| \tag{6.19}$$

, where n is the number of samples and B_m is the set of indices of samples whose uncertainty falls into the interval.

We propose the following modified notion of Eq. (6.4) to quantify the maximum uncertainty calibration error (MUCE):

$$\text{MUCE} = \max_{m \in \{1,\ldots,M\}} |\text{error}(B_m) - \text{uncertainty}(B_m)| \tag{6.20}$$

As a measure, MUCE is most appropriate for high-risk applications, where the goal is to minimise the worst-case deviations between error and estimated uncertainty. Therefore, MUCE calculates the maximum calibration uncertainty for the bins.

In critical applications, it might be necessary to enforce a low MUCE in order to reduce the risk of overconfidence in prediction. A concise way to visualise the degree of calibration of a model is called Uncertainty-Reliability diagrams.

6.4.2 Sharpness

Sharpness with log score refers to the negative entropy of the predictive distributions. The accuracy of estimated uncertainty in predictive distributions should be evaluated by maximising the sharpness of the subject during calibration. In practice, estimated uncertainty is not sufficient to calculate the useful probability of making a prediction. For example, a perfectly calibrated binary classifier on a balanced classification problem will always return an uncertainty of 50%, as this is the probability of making a false prediction.

Therefore, the sharpness (Kuleshov et al., 2018) of a model is a good measure of how close the confidence estimates are between 0 and 1. We propose measuring sharpness using the variance as:

$$\text{sharpness}(f) = Var[|\text{error}(B_m) - \text{uncertainty}(B_m)|]. \tag{6.21}$$

6.5 Cost-sensitive Medical Image Dataset:

The novel Coronavirus 2019 (COVID-2019), which results in pneumonia at varying severity, has rapidly become a pandemic. We have selected 68 Posterior-Anterior (PA) X-ray images of lungs with COVID-19 cases from Dr Joseph Cohen's Github repository (Cohen et al., 2020). This repository is constantly updated with images shared by researchers. In addition, we augmented the dataset with normal and pneumonia images from Kaggle's Chest X-Ray Images. This has produced a total of 5,941 PA chest radiography images across four classes (Normal: 1583, Bacterial Pneumonia: 2786, non-COVID-19 Viral Pneumonia: 1504, and COVID-19: 68). Finally, we standardised and resized all images to 224 x 224 pixels.

6.6 Experiment

We used a pre-trained ResNet50V2 model (He et al., 2016b) and acquired data only to fine-tune the original model. Next, we introduced Dropweights, followed by a softmax activated layer, which was then applied in the fully connected layer on top of the ResNet50V2 convolutional base to estimate a meaningful model uncertainty.

We split the whole dataset into 80% and 20% between training and testing sets, respectively. Real-time data augmentation was also applied, leveraging Keras ImageDataGenerator during training to prevent overfitting and enhance the learning capability of the model. The Adam optimiser was used with a learning rate of 1e-5 and a decay factor of 0.2. All our experiments were run for 25 epochs, and the batch size was set to 8. Dropweights with rates of 0.3 were added to a fully connected layer. After every epoch, we monitored the validation accuracy 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 25 through the Bayesian Deep Residual Neural Networks.

6.7 Utility Function

In this study, the utility function in table 6.1 prescribes fewer false negatives for Covid-19, Normal, Viral Pneumonia and Bacterial cases, relative to the other categories from the costs of incorrect diagnoses to a task-specific utility function. Maximum utility (2.1) is for correct prediction and the lowest utility (1.2) is given to errors in predicting the Normal and Covid. In safety critical applications, the utility function values to be assigned according to the functional requirements.

	Normal	Bacterial Pneumonia	Viral Pneumonia	Covid
Normal	2.1	1.2	1.2	1.2
Bacterial Pneumonia	1.4	2.1	1.4	1.4
Viral Pneumonia	1.4	1.4	2.1	1.4
Covid	1.2	1.2	1.2	2.1

Table 6.1 Illustration of a utility matrix for a Covid-19 detection example

6.8 Results and Discussions

6.8.1 Model Performance

The normalised confusion matrices in Figure 6.1 demonstrate how the different models compare when making predictions. Each confusion matrix displays the resulting classification when averaging the utility function with respect to the dropweights samples of each network. We highlight that our Loss-Calibrated (i.e. cost-sensitive) BNN model captures our preferences by avoiding false negatives of the 'Covid-19' diseased class. There is also a clear performance gain from the cost-sensitive model. This compares favourably to the standard model, where there is a common failure mode of predicting a patient as being 'Normal' when they are 'Covid-19' infected.

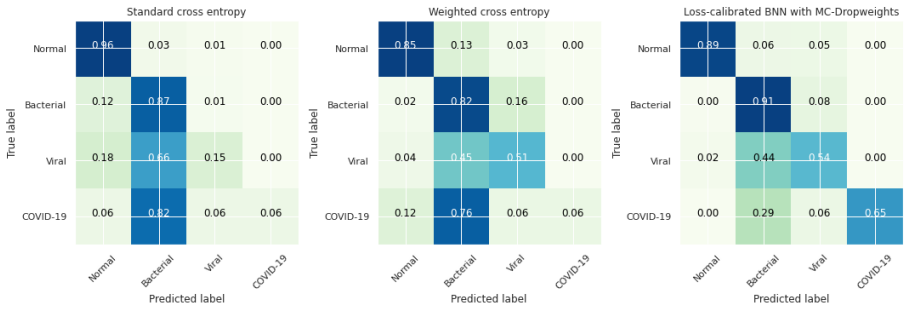


Fig. 6.1 Confusion Matrix. Left: Standard NN model with cross-entry loss. Middle: Standard NN model with weighted cross-entry loss. Right: Loss-Calibrated BNN model.

6.8.2 The Relation between Cost as Expected Loss and Predictive Accuracy

In this automatic disease detection in X-Ray Images example, our goal is to reduce false negatives whilst being concerned about false positives. Table 6.2 below demonstrates a strong correlation between prediction accuracy and the loss as costs of incorrect misdiagnoses to a task-specific utility function; for example, the lowest utility and, therefore, the highest cost is

assigned for a patient, who is misdiagnosed as being healthy when their condition is severe. The result evident from the table 6.2 is that calibrated BNN with MC-Dropweights provides significant improvements in calibrated confidence over measured uncertainty from standard NNs.

	Standard BNN	Weighted Cross Entropy	Loss-Calibrated BNN
Accuracy (%)	70.12	74.09	80.88
Expected loss	0.20	0.16	0.12
ECE	13.27	4.71	7.01
UCE	16.05	28.79	10.86
MUCE	3.77	5.42	1.91
Sharpness	0.009	0.026	0.027

Table 6.2 Calibration error results for different models.

6.8.3 Reliability Diagrams

Reliability diagrams is a visual representation of model calibration (DeGroot and Fienberg, 1983; Guo et al., 2017). Figure 6.2 diagrams plot the expected accuracy obtained for each bin (fraction of positives) against the binned predicted confidences. A perfectly calibrated model would result in a 45-degree line. Any deviation from this perfect diagonal represents miscalibration, where a lower ECE (close to zero) indicates a better calibration.

Model Uncertainty-Reliability diagrams in Figure 6.3 represent the deviation of the perfect calibration by plotting the binned measured model uncertainties against the error obtained for each bin (fraction of negatives). The UCE is defined as the absolute error of these bins (i.e., the gap between uncertainty and accuracy) weighted by the number of samples in the bins, where a higher UCE indicates a better calibration. Loss-calibrated (i.e. cost-sensitive) model updates the posterior approximation, so that estimated uncertainty in Bayesian inference for optimal decisions are better in terms of a user-defined asymmetric cost and error in prediction to avoid overconfidence.

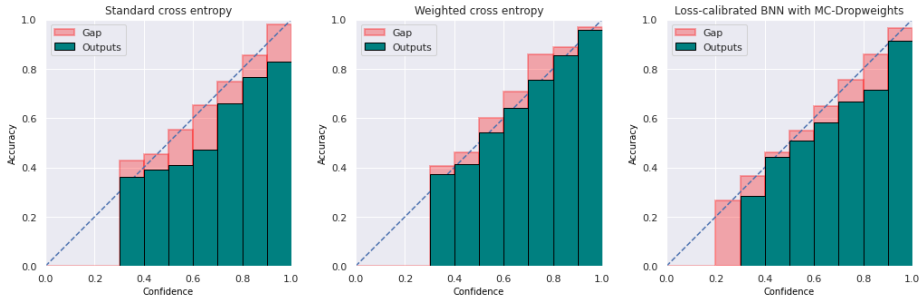


Fig. 6.2 Confidence-Reliability diagrams showing three classifiers (Left: Standard NN model with cross-entry loss. Middle: Standard NN model with weighted cross-entry loss. Right: Loss-Calibrated (i.e. cost-sensitive) BNN model.) and its confidence histogram (M = 10 bins) on Covid-19 image dataset.

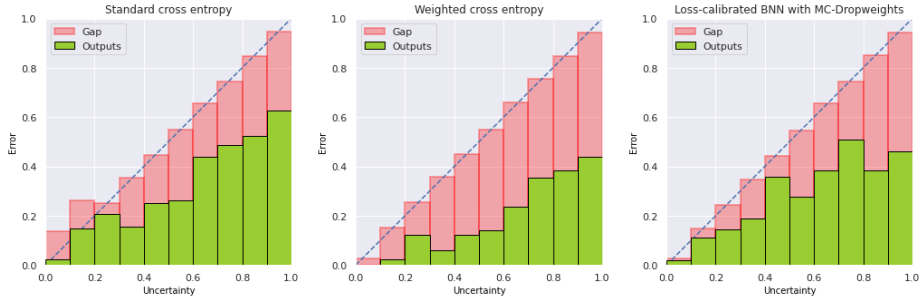


Fig. 6.3 Uncertainty-Reliability diagrams showing three classifiers (Left: Standard NN model with cross-entry loss. Middle: Standard NN model with weighted cross-entry loss. Right: Loss-Calibrated (i.e. cost-sensitive) BNN model.) and its Uncertainty histogram (M = 10 bins) for Covid-19 image dataset.

6.9 Uncertainty Calibration Evaluation Measures in Deep Learning

Our cost-sensitive BNN model incorporates asymmetric misclassification costs as a utility function to enable the rejection of uncertain predictions and minimise the misclassification, resulting in improved performance that incorporates practical considerations.

6.9.1 Adaptive Expected Calibration Error (AECE)

The reliability of a deep learning model's confidence in its predictions, i.e., calibration, is critical to trust a medical diagnosis prediction (Nixon et al., 2019; Raghu et al., 2019). The computation of calibration for multi-class settings depends on the binning strategy. Equal-size binning was proposed to address the known issues of the common fixed equal-size binning (Nixon et al., 2019; Vaicenavicius et al., 2019). However, it is vulnerable to the

undetectable accuracy gap, internal compensation, and inaccurate accuracy estimation (Ding et al., 2020). Ding showed that the Area Under Receiver Operator Characteristics (AUROC) and the Area Under Precision Recall (AUPR) metrics are meaningless or even misleading for minimal changes in prediction models, and so cannot be compared for confidence estimation performance scores unless evaluated on the exact same classifier. Instead, they proposed to use the Area Under Risk Coverage (AURC) curve measure. On the risk-coverage curve, each point coordinate corresponds to a certain confidence threshold which is calculated as:

$$Coverage = \frac{TN+FN}{TN+TP+FP+FN} \; Risk = \frac{FN}{TN+FN} \tag{6.22}$$

To address the known issues of the fixed equal-size binning in multi-class settings, we explored an adaptive binning strategy (Hendrycks and Dietterich, 2019; Nguyen and O'Connor, 2015), where the number of samples in a bin is adaptive to the distribution of the samples in the confidence range. Unlike the Brier score (Lakshminarayanan et al., 2017), our metric is scalar-valued to measure calibration.

$$\text{AECE} = \frac{1}{MC} \sum_{c=1}^{C} | \sum_{m=1}^{M} \text{accuracy}(m,c) - \sum_{m=1}^{M} \text{confidence}(m,c)| \tag{6.23}$$

, where, adaptive calibration range M for class label C, respectively; and N is the total number of data points. Calibration range m defined by the $\lfloor \frac{N}{M} \rfloor$ index of the sorted and thresholded predictions to use different binning strategy such that these bins are not uniformly distributed.

6.9.2 Adaptive Uncertainty Calibration Error (AUCE)

Calibrated uncertainty in machine learning is critical in safety-critical applications such as the exploration phase of many algorithms, autonomous vehicles, medical applications.

We propose the following modified notion of equation 6.23 for bias-reduced Adaptive Uncertainty Calibration Error (AUCE) as a measure of the degree of calibration to quantify the miscalibration of the difference in expectation between model error and estimated uncertainty.

$$\text{AUCE} = \frac{1}{MC} \sum_{c=1}^{C} | \sum_{m=1}^{M} \text{error}(m,c) - \sum_{m=1}^{M} \text{uncertainty}(m,c)| \tag{6.24}$$

, where n is the number of samples and B_m is the set of indices of samples whose uncertainty falls into the interval.

We propose the following definition to quantify the maximum adaptive uncertainty calibration error (MAUCE):

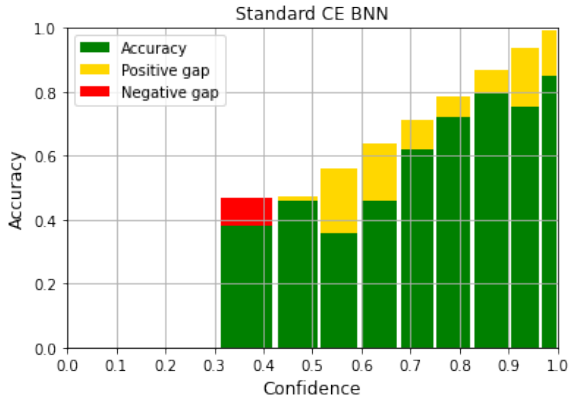
$$\text{MAUCE} = \max_{m\in\{1,\ldots,M\}} \frac{1}{C}\sum_{c=1}^{C} \left|\frac{1}{M}\sum_{m=1}^{M} \text{error}(m,c) - \frac{1}{M}\sum_{m=1}^{M} \text{uncertainty}(m,c)\right| \tag{6.25}$$

As a measure, MAUCE is most appropriate for high-risk applications, where the goal is to minimise the worst-case deviations between error and estimated uncertainty. In critical applications, it might be necessary to enforce a high MAUCE in order to reduce the risk of overconfidence in prediction.

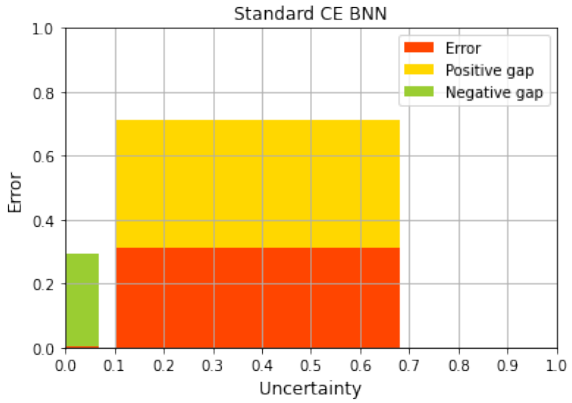
Table 6.3 and 6.4 shows that the cost sensitive neural networks (i.e. loss-calibrated) exhibits considerably better performance for Covid-19 X-Ray and CT image dataset, which demonstrates the importance of considering the asymmetric cost involved in misclassification in medical image analysis while estimating uncertainty in deep learning model. A concise way to visualise the degree of calibration of a model is called Uncertainty-Reliability diagrams as shown in figure 6.4.

%Metrics	Standard BNN	Weighted BNN	Calibrated BNN
Accuracy(↑)	0.7039	0.7355	0.8102
Loss(↓)	0.1986	0.1631	0.1153
Adaptive Expected Calibration Error (AECE) (↓)	0.1289	0.0442	0.0716
Adaptive Maximum Calibration Error (AMCE) (↓)	0.2001	0.1231	0.1608
Adaptive Uncertainty Calibration Error (AUCE) (↑)	0.3293	0.3338	0.5365
Maximum Adaptive Uncertainty Calibration Error (MAUCE) (↑)	0.3984	0.3849	0.6115
Area under the Risk-Coverage curve (AURC)(↓)	0.1728	0.1399	0.0819
Adaptive-Area under the Risk-Coverage curve (AURC)(↓)	0.1223	0.1034	0.0626

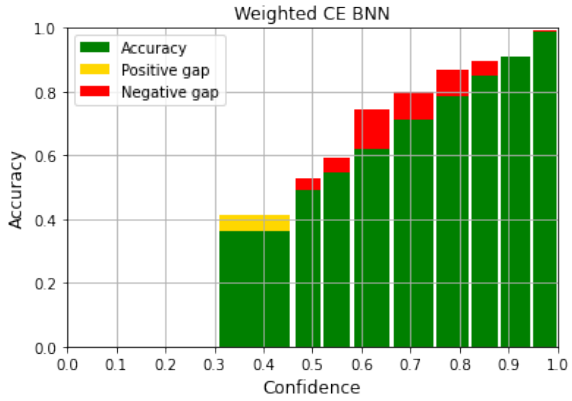
Table 6.3 Model Performance Metrics - Covid-19 X-Rays Image



(a) Reliability Diagrams

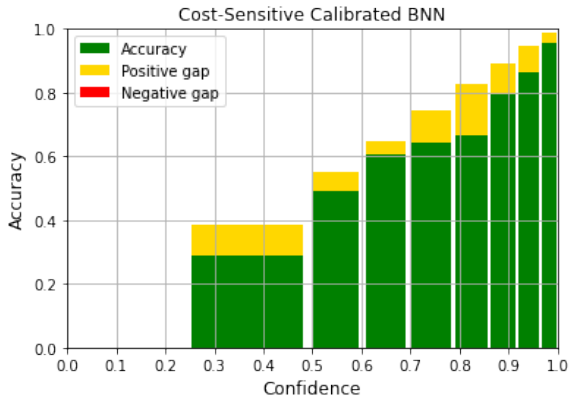


(b) Uncertainty Reliability Diagrams

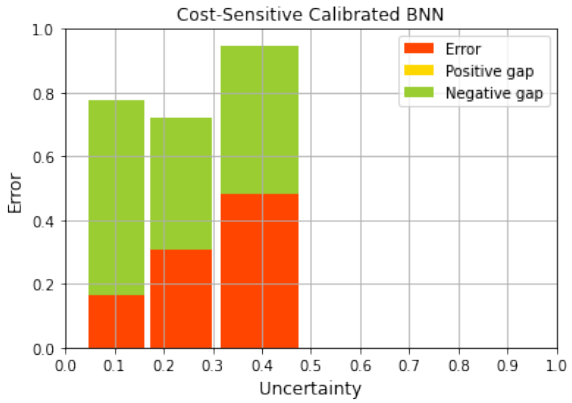


(c) Reliability Diagrams

(d) Uncertainty Reliability Diagrams



(e) Reliability Diagrams



(f) Uncertainty Reliability Diagrams

Fig. 6.4 Undetectable error in Reliability Diagrams and Uncertainty Reliability Diagrams. In all Reliability Diagrams, positive error means confidence is larger than accuracy.

%Metrics	Standard BNN	Weighted BNN	Calibrated BNN
Accuracy(↑)	0.8672	0.8974	0.9376
Loss(↓)	0.1328	0.1026	0.0624
Adaptive Expected Calibration Error (AECE) (↓)	0.0197	0.0210	0.0111
Adaptive Maximum Calibration Error (AMCE) (↓)	0.0541	0.0847	0.0466
Adaptive Uncertainty Calibration Error (AUCE) (↑)	0.0420	0.8797	0.6943
Maximum Adaptive Uncertainty Calibration Error (MAUCE) (↑)	0.0420	0.8797	0.6943
Area under the Risk-Coverage curve (AURC)(↓)	0.0822	0.0695	0.0216
Adaptive-Area under the Risk-Coverage curve (AURC)(↓)	0.0729	0.0641	0.0197

Table 6.4 Model Performance Metrics - Covid-19 CT Image

6.10 Uncertainty Quality Metrics:

We have evaluated the quality of uncertainty estimate using two statistical matrices: Predictive Log-Likelihood (PLL) and Brier score (BS).

Predictive Log-Likelihood (PLL) is a widely accepted metric for uncertainty quality, used as the main uncertainty quality metric (Gal and Ghahramani, 2016; Hernández-Lobato et al., 2016). A fundamental property is that PLL makes no assumptions about the form of the distribution. While PLL is an elegant measure, it has been criticised for allowing outliers to affect the score negatively.

%Metrics	Standard BNN	Weighted BNN	Calibrated BNN
Predictive Log Likelihood (PLL)(↓)	-10.18	-10.18	-7.368
Brier Score (↓)	0.8845	0.8730	0.0077

Table 6.5 Estimated Uncertainty Quality Metrics - Covid-19 X-Rays Image

6.11 Conclusion

Bayesian decision-theoretic cost-sensitive calibrated uncertainty estimation the technique of MC dropweights achieves encouraging performance when learning an approximate distribution over the weight parameters, incorporating uncertainty and user-defined asymmetric utility functions. Critical decision-making for medical imaging applications requires high accuracy and reliable estimation of predictive uncertainty to interpret the model. The significance of our experiment demonstrates the usefulness of the model to large networks with real-world medical imaging Covid-19 diseases detection in improving the reliability in clinical decision making. We discussed the issues with the existing performance measures and leveraged risk coverage curve and adaptive calibration methods to mitigate these issues to simulate the interventional Human-in-the-loop (HITL) decision making task in safety-critical applications with an asymmetric cost or non-trivial losses.

Chapter 7

Conclusion and future research

This chapter summarises the main results from Chapter 3 to Chapter 6 and discusses the directions for future work.

7.1 Conclusion

Digital pathology combined with medical imaging and deep learning-based diagnosis has the potential to revolutionise patient care and support clinicians by making the diagnosis and monitoring of disease much more efficient. Deep learning has achieved a remarkable performance in medical image analysis. Deep learning models focus exclusively on improving the accuracy of point predictions without assessing the quality of their outputs and are often considered black boxes in nature. However, improved interpretability, explicability and transparency are needed to translate automated decision-making for computer-based medical applications, which have critical safety impacts. Furthermore, knowing how much confidence there is in a prediction is essential for gaining clinicians' trust in the technology. These concerns are reinforced by the core problem addressed in this thesis: the overconfidence of deep models when making false predictions. How do we know when the machine does not know?

Neural networks trained with Dropout (Gal, 2016) struggle to fully capture the posterior distribution of the weights. Lakshminarayanan et al. (Lakshminarayanan et al., 2017) showed that MC-Dropout could produce overconfident wrong predictions and, by simply averaging the prediction over multiple models, one achieves a better performance and confidence scores. On the other hand, Beluch et al. (Beluch et al., 2018) demonstrated that an ensemble is better than a single MC-Dropout, but going beyond 3 networks in their deterministic ensemble method does not significantly improve performance. This leads to over-confident predictions, particularly in a deep learning-based medical image analysis.

The main goal of this thesis is to quantify uncertainty to make medical imaging with deep learning more robust and accurate, which leads us to our research questions. In this subsection, we discuss the results obtained in this thesis.

1. How to measure model uncertainty in deep learning?

   We presented a possible answer to the questions above using Bayesian modelling and variational inference that ensures posterior distributions remain meaningful through appropriate calibration to real-world applications. We have shown that Neural Networks with DropWeights can be interpreted as a Bayesian approximation that improves the quality of estimated uncertainty in deep learning. We applied proper scoring rules, such as the Brier Score and Predictive Log-Likelihood (PLL), along with more intuitive heuristics, such as the adaptive uncertainty calibration error (AUCE), to understand how different neural networks dealt with model uncertainty without compromising accuracy.

2. Is approximate Bayesian neural networks a good principle for deep learning?

   Bayes' theorem is one of the most important formulae in the field of statistics and probability theory. Since the posterior predictive distribution in neural networks is computationally intractable, all Bayesian inference procedures in deep learning are approximate. The Bayesian inference provides a natural and principled way of modelling uncertainty with a prior distribution on its weights, which is robust to overfitting (i.e. regularisation) in deep learning. The key property of a Bayesian approach is marginalisation instead of optimisation. Deep neural networks can represent many different architectures but high-performing models corresponding to different settings of hyper-parameters. In such cases, marginalisation makes the difference for both calibration and accuracy while retaining scalability.

3. Are there any alternatives of quantifying uncertainty that align better with our goal?

   Deep ensembles have been considered as an alternative approach to approximate Bayesian methods. Deep ensembles directly average model predictions from different networks, while an approximate Bayesian neural network computes a weighted average using the posterior of the network weights. Therefore, Deep Ensembles Bayesian Neural Networks with DropWeights (section 3.3.5) brings encouragement and additional insights to Bayesian neural networks.

4. Can we develop practical inference algorithms to measure model uncertainty in cost-calibrated situations?

Applications like pathology, ophthalmology, radiology, and dermatology have already benefited from using Bayesian neural networks. Therefore, the key challenge in applying Bayesian neural networks to real-world applications is balancing the need for meaningful predictive distributions with the ability to scale to complex neural networks for high-dimensional image data, where the cost of misclassification is asymmetric. Overcoming this challenge requires the involvement of domain experts and an understanding of the functional aspects of their requirements and how to apply loss-calibration to solve issues of asymmetric cost in BNNs. Chapter 6 is an example of this kind of problem leveraging Covid-19 chest x-ray image data and shows how the challenges in asymmetric misclassification cost using cost/utility matrix can drive improvements in the deep learning techniques.

This thesis explored these ideas using concepts from the calibration of confidence, uncertainty in deep learning, and cost-sensitive neural networks in decision analysis for medical imaging. Specifically, we showed how to improve calibrated predictive uncertainty by leveraging Deep Neural Networks with DropWeights, while minimising asymmetric cost to achieve a better performance and uncertainty estimates compared to independent evaluations, ensembles method or MC-dropout. First, we show how to improve predictive uncertainty by Dropweights based Bayesian neural networks learning an approximate distribution over its weights in medical image segmentation and its application in active learning. Second, we use the Jackknife resampling technique to correct bias in quantified uncertainty in image classification and propose metrics to measure uncertainty performance. The third part of the thesis is motivated by the discrepancy between the model predictive error and the objective in quantified uncertainty when costs for misclassification errors or imbalanced datasets are asymmetric. Fourth, we develop cost-sensitive modifications of the Bayesian neural networks in disease detection and propose metrics to measure the quality of quantified uncertainty. Finally, we leveraged adaptive binning strategy in order to measure Adaptive Uncertainty Calibration Error (AUCE) and Adaptive Maximum Uncertainty Calibration Error (AMUCE), which directly corresponds to measure uncertainty calibration error that directly corresponds to estimated uncertainty performance and address problematic evaluation methods.

We evaluated the effectiveness of the tools on nuclei images segmentation, multi-class Brain MRI image classification, multi-level cell type-specific protein expression prediction in ImmunoHistoChemistry (IHC) images and cost-sensitive classification for Covid-19 detection from X-Rays and CT image dataset.

Our approach is thoroughly validated by measuring the quality of uncertainty using two metrics: Predictive Log-Likelihood (PLL) and Brier Score (BS), which produced an equally

good or better result and paved the way for future advances that address important practical problems at the intersection of deep learning and Bayesian decision theory.

Based on estimated predictive uncertainty with the ability model to say "I don't know", deep learning models will be able to flag clinicians for a second opinion or further examine patients for highly uncertain model predictions. In conclusion, estimated uncertainty with point prediction leads to improved prediction quality and, hence, a more informed decision. Moreover, cost-sensitive neural networks with DropWeights allows quantifying better calibrated predictive uncertainty.

7.2 Future Research

This thesis work can open up many opportunities for the use of the deep ensembles technique in deep learning towards the overall goal of achieving efficient medical image analysis, which is an open research problem. In our work, we considered the use of DropWeights as a Bayesian approximation to represent the model's uncertainty. Since the single DropWeights model collapses around a subspace of the posterior distribution, we assume that each ensemble model member will capture the behaviour around a different local mode. This will require a more detailed theoretical analysis for future research.

The model was tested on an independent dataset of IHC images corresponding to clinical samples from another laboratory, which showed lower overall accuracy. Independent datasets that are generated by different laboratories can be considered the most challenging approach for assessing if a model is fully generalisable despite acquiring all images that were digitally available by the other laboratory. Furthermore, these images differed significantly in cell morphology, image quality, colour settings during acquisition, as well as the overall brightness and contrast. It is therefore not surprising that the results will differ significantly and lead to a higher discard rate, unless a universally accepted standardisation of IHC staining workflows and digitisation of images is introduced. To achieve such a standard is undoubtedly a difficult task, as even stainings generated by the same equipment and protocols may differ between laboratories due to the exact batch or brand of the reagents (Mengel et al., 2002). Additionally, there are several steps in the workflow that can never be controlled such as preprocessing and fixation of already existing archived tissue material, making standardisation almost impossible. Another possibility for future projects utilising the proposed workflow is to include images generated by multiple laboratories in the initial training of the model, which would likely improve the overall generalisability.

Bayesian neural networks address the inherent uncertainty in model predictions. However, factors that contribute to uncertainty have not been extensively investigated, for example,

how uncertainty changes due to input pixels or how to measure the change in uncertainty when a given input feature is known compared to when it is unknown.

One useful future research direction will be to extend other methods for representing uncertainty to interpret uncertainty estimates differently over multi-label classification tasks in medical imaging. The properties in the dataset, such as label correlations and label cardinality, can strongly affect the uncertainty quantification in predictive probability performance of a deep learning algorithm in multi-label settings. Unfortunately, there is no systematic study on how and why the performance varies over different data properties; any such study would be useful in deciding multi-label algorithms and active learning in medical imaging.

With the Bayesian interpretation of modern deep learning, it should hopefully include a better-calibrated uncertainty estimate to improve the performance, the effect on the quality of uncertainty across different data modalities and network architecture, and data efficiency to solve a very tedious, time-consuming and challenging manual medical image analysis.

I hope that the deep learning framework presented in this thesis will lay the foundations of a new and exciting field of study, combining modern deep learning, deep ensembles and approximate Bayesian techniques in a more principled and practical way.

References

Achrack, O., Kellerman, R., and Barzilay, O. (2020). Multi-loss sub-ensembles for accurate classification with uncertainty estimation. *arXiv preprint arXiv:2010.01917*.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515.

Aggarwal, C., Kong, X., Gu, Q., Han, J., and Yu, P. (2014). Active learning: A survey. *In: Aggarwal, C.C. (ed.) Data Classification: Algorithms and Applications*, pages 571–606 CRC Press.

Altaf, F., Islam, S. M., Akhtar, N., and Janjua, N. K. (2019). Going deep in medical image analysis: Concepts, methods, challenges, and future directions. *IEEE Access*, 7:99540–99572.

Aprupe, L., Litjens, G., Brinker, T. J., van der Laak, J., and Grabe, N. (2019). Robust and accurate quantification of biomarkers of immune cells in lung cancer micro-environment using deep convolutional neural networks. *PeerJ*, 7:e6335.

Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., Mendonça, A. M., and Campilho, A. (2020). Dr| graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, 63:101715.

Ashukha, A., Lyzhov, A., Molchanov, D., and Vetrov, D. (2020). Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*.

Awate, S. P., Garg, S., and Jena, R. (2019). Estimating uncertainty in mrf-based image segmentation: A perfect-mcmc approach. *Medical image analysis*, 55:181–196.

Ayhan, M. S. and Berens, P. (2018). Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks.

Ayhan, M. S., Kühlewein, L., Aliyeva, G., Inhoffen, W., Ziemssen, F., and Berens, P. (2020). Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical Image Analysis*, 64:101724.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495.

Balasooriya, N. M. and Nawarathna, R. D. (2017). A sophisticated convolutional neural network model for brain tumor classification. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–5. IEEE.

Barber, D. and Bishop, C. M. (1998). Ensemble learning in bayesian neural networks. *Nato ASI Series F Computer and Systems Sciences*, 168:215–238.

Basharin, G. P. (1959). On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability & Its Applications*, 4(3):333–336.

Baumgartner, C. F., Tezcan, K. C., Chaitanya, K., Hötker, A. M., Muehlematter, U. J., Schawkat, K., Becker, A. S., Donati, O., and Konukoglu, E. (2019). Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210.

Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. (2018). The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155.

Benítez, J. M., Castro, J. L., and Requena, I. (1997). Are artificial neural networks black boxes? *IEEE Transactions on neural networks*, 8(5):1156–1164.

Bian, C., Yuan, C., Wang, J., Li, M., Yang, X., Yu, S., Ma, K., Yuan, J., and Zheng, Y. (2020). Uncertainty-aware domain alignment for anatomical structure segmentation. *Medical Image Analysis*, 64:101732.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Blom, S., Erickson, A., Östman, A., Rannikko, A., Mirtti, T., Kallioniemi, O., and Pellinen, T. (2019). Fibroblast as a critical stromal cell type determining prognosis in prostate cancer. *The Prostate*, 79(13):1505–1513.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. *In Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research, pages 1613–1622*.

Bogunovic, H., Seaman, J., Margaron, P., Seeböck, P., Gerendas, B. S. S., Lorand, D., Normand, G., and Schmidt-Erfurth, U. (2020). Detection of retinal fluids in oct scans by an automated deep learning algorithm compared to human expert grading in the hawk & harrier trials. *Investigative Ophthalmology & Visual Science*, 61(7):5187–5187.

Bragman, F. J., Tanno, R., Eaton-Rosen, Z., Li, W., Hawkes, D. J., Ourselin, S., Alexander, D. C., McClelland, J. R., and Cardoso, M. J. (2018). Quality control in radiotherapy-treatment planning using multi-task learning and uncertainty estimation. *International Conference on Medical Imaging with Deep Learning, 2018*.

Bulten, W., Bándi, P., Hoven, J., van de Loo, R., Lotz, J., Weiss, N., van der Laak, J., van Ginneken, B., Hulsbergen-van de Kaa, C., and Litjens, G. (2019). Epithelium segmentation using deep learning in h&e-stained prostate specimens with immunohistochemistry as reference standard. *Scientific reports*, 9(1):1–10.

Caicedo, J. C., Goodman, A., Karhohs, K. W., Cimini, B. A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., et al. (2019). Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253.

Cao, X., Chen, H., Li, Y., Peng, Y., Wang, S., and Cheng, L. (2020a). Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation. *IEEE Transactions on Medical Imaging*, 40(1):431–443.

Cao, Y., Geddes, T. A., Yang, J. Y. H., and Yang, P. (2020b). Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*, 2(9):500–508.

Carneiro, G., Pu, L. Z. C. T., Singh, R., and Burt, A. (2020). Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. *Medical image analysis*, 62:101653.

Caruana, R., Lawrence, S., and Giles, C. L. (2001). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in neural information processing systems*, pages 402–408.

Chen, T. and Chefd'Hotel, C. (2014). Deep learning based automatic immune cell detection for immunohistochemistry images. In *International workshop on machine learning in medical imaging*, pages 17–24. Springer.

Chen, T., Fox, E., and Guestrin, C. (2014). Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR.

Chiou, E., Giganti, F., Punwani, S., Kokkinos, I., and Panagiotaki, E. (2020). Harnessing uncertainty in domain adaptation for mri prostate lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 510–520. Springer.

Chollet, F. et al. (2015). Keras documentation. *keras. io*.

Chu, W.-T. and Guo, H.-J. (2017). Movie genre classification based on poster images with deep neural networks. In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*, pages 39–45.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al. (2013). The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057.

Cobb, A. D., Roberts, S. J., and Gal, Y. (2018). Loss-calibrated approximate inference in bayesian neural networks. *arXiv preprint arXiv:1805.03901*.

Cohen, J. P., Morrison, P., and Dao, L. (2020). Covid-19 image data collection. *arXiv 2003.11597*.

Combalia, M., Hueto, F., Puig, S., Malvehy, J., and Vilaplana, V. (2020). Uncertainty estimation in deep neural networks for dermoscopic image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 744–745.

Dahal, L., Kafle, A., and Khanal, B. (2020). Uncertainty estimation in deep 2d echocardiography segmentation. *arXiv preprint arXiv:2005.09349*.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. Springer Science & Business Media.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., and Udluft, S. (2017). Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. *arXiv preprint arXiv:1710.07283*.

Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

Di Scandalea, M. L., Perone, C. S., Boudreau, M., and Cohen-Adad, J. (2019). Deep active learning for axon-myelin segmentation on histology data. *arXiv preprint arXiv:1907.05143*.

Dickie, D. A., Job, D. E., Rodriguez, D., Robson, A., Danso, S., Pernet, C., Bastin, M. E., Deary, I. J., Shenkin, S. D., Wardlaw, J., et al. (2016). Brain imaging of normal subjects (brains) age-specific mri atlases from young adults to the very elderly (v1. 0).

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., and Sun, J. (2021). Repvgg: Making vgg-style convnets great again. *arXiv preprint arXiv:2101.03697*.

Ding, Y., Liu, J., Xiong, J., and Shi, Y. (2020). Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4–5.

Djureinovic, D., Fagerberg, L., Hallström, B., Danielsson, A., Lindskog, C., Uhlén, M., and Pontén, F. (2014). The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Molecular human reproduction*, 20(6):476–488.

Do, H. P., Guo, Y., Yoon, A. J., and Nayak, K. S. (2020). Accuracy, uncertainty, and adaptability of automatic myocardial asl segmentation using deep cnn. *Magnetic resonance in medicine*, 83(5):1863–1874.

Donnat, C. and Holmes, S. (2020). Modeling the heterogeneity in covid-19's reproductive number and its impact on predictive scenarios. *arXiv preprint arXiv:2004.05272*.

Draper, D. (1994). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B*, 56.

Duan, W., Xu, C., Liu, Q., Xu, J., Weng, Z., Zhang, X., Basnet, T. B., Dahal, M., and Gu, A. (2020). Levels of a mixture of heavy metals in blood and urine and all-cause, cardiovascular disease and cancer mortality: A population-based cohort study. *Environmental Pollution*, 263:114630.

Duch, W. (2003). Coloring black boxes: visualization of neural network decisions. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 3, pages 1735–1740. IEEE.

Dusenberry, M., Jerfel, G., Wen, Y., Ma, Y., Snoek, J., Heller, K., Lakshminarayanan, B., and Tran, D. (2020). Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pages 2782–2792. PMLR.

Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., and Cardoso, M. J. (2018). Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 691–699. Springer.

Eaton-Rosen, Z., Varsavsky, T., Ourselin, S., and Cardoso, M. J. (2019). As easy as 1, 2... 4? uncertainty in counting tasks for medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 356–364. Springer.

Eggenreich, S., Payer, C., Urschler, M., and Štern, D. (2020). Variational inference and bayesian cnns for uncertainty estimation in multi-factorial bone age prediction. *arXiv preprint arXiv:2002.10819*.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd.

Elsken, T., Metzen, J. H., Hutter, F., et al. (2019). Neural architecture search: A survey. *J. Mach. Learn. Res.*, 20(55):1–21.

Epstein, L. G. and Wang, T. (1995). Uncertainty, risk-neutral measures and security price booms and crashes. *Journal of Economic Theory*, 67(1):40–82.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118.

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., et al. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & cellular proteomics*, 13(2):397–406.

Feng, M., Deng, Y., Yang, L., Jing, Q., Zhang, Z., Xu, L., Wei, X., Zhou, Y., Wu, D., Xiang, F., et al. (2020). Automated quantitative analysis of ki-67 staining and he images recognition and registration based on whole tissue sections in breast carcinoma. *Diagnostic Pathology*, 15:1–12.

Filos, A., Farquhar, S., Gomez, A. N., Rudner, T. G., Kenton, Z., Smith, L., Alizadeh, M., de Kroon, A., and Gal, Y. (2019). A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*.

Foong, A. Y., Li, Y., Hernández-Lobato, J. M., and Turner, R. E. (2019). 'in-between'uncertainty in bayesian neural networks. *arXiv preprint arXiv:1906.11537*.

Fortuin, V. (2021). Priors in bayesian deep learning: A review. *arXiv preprint arXiv:2105.06868*.

Gal, Y. (2016). *Uncertainty in deep learning*. PhD thesis, University of Cambridge.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning, ICML 2016*, volume 3, pages 1651–1660.

Gal, Y., Hron, J., and Kendall, A. (2017a). Concrete dropout. *arXiv preprint arXiv:1705.07832*.

Gal, Y., Islam, R., and Ghahramani, Z. (2017b). Deep bayesian active learning with image data. *arXiv preprint arXiv:1703.02910*.

Ganaie, M., Hu, M., et al. (2021). Ensemble deep learning: A review. *arXiv preprint arXiv:2104.02395*.

Gantenbein, M., Erdil, E., and Konukoglu, E. (2020). Revphiseg: A memory-efficient neural network for uncertainty quantification in medical image segmentation. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pages 13–22. Springer.

Geread, R. S., Morreale, P., Dony, R. D., Brouwer, E., Wood, G. A., Androutsos, D., and Khademi, A. (2019). Ihc color histograms for unsupervised ki67 proliferation index calculation. *Frontiers in bioengineering and biotechnology*, 7:226.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459.

Ghesu, F. C., Georgescu, B., Gibson, E., Guendel, S., Kalra, M. K., Singh, R., Digumarthy, S. R., Grbic, S., and Comaniciu, D. (2019). Quantifying and leveraging classification uncertainty for chest radiograph assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 676–684. Springer.

Ghesu, F. C., Georgescu, B., Mansoor, A., Yoo, Y., Gibson, E., Vishwanath, R., Balachandran, A., Balter, J. M., Cao, Y., Singh, R., et al. (2021). Quantifying and leveraging predictive uncertainty for medical image assessment. *Medical Image Analysis*, 68:101855.

Ghoshal, B., Hikmet, F., Pineau, C., Tucker, A., and Lindskog, C. (2021a). Deephistoclass: A novel strategy for confident classification of immunohistochemistry images using deep learning. *Molecular & Cellular Proteomics*, page 100140.

Ghoshal, B., Lindskog, C., and Tucker, A. (2020). Estimating uncertainty in deep learning for reporting confidence: An application on cell type prediction in testes based on proteomics. In *International Symposium on Intelligent Data Analysis*, pages 223–234. Springer.

Ghoshal, B., Swift, S., and Tucker, A. (2021b). Bayesian deep active learning for medical image analysis. In *International Conference on Artificial Intelligence in Medicine*, pages 36–42. Springer.

Ghoshal, B. and Tucker, A. (2020a). Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. *arXiv preprint arXiv:2003.10769*.

Ghoshal, B. and Tucker, A. (2020b). On calibrated model uncertainty in deep learning. *The European Conference on Machine Learning (ECML PKDD 2020) Workshop on Uncertainty in Machine Learning*.

Ghoshal, B. and Tucker, A. (2021). 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)*, pages 503–509. IEEE.

Ghoshal, B., Tucker, A., Sanghera, B., and Lup Wong, W. (May, 2019a). Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. *Computational Intelligence for Social Media Data Mining and Knowledge Discovery*, 37:701–734.

Ghoshal, B., Tucker, A., Sanghera, B., and Wong, W. (June 2019b). Estimating uncertainty in deep learning for reporting confidence to clinicians when segmenting nuclei image data. *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, 1:318–324.

González-López, J., Ventura, S., and Cano, A. (2019). Distributed selection of continuous features in multilabel classification using mutual information. *IEEE transactions on neural networks and learning systems*, 31(7):2280–2293.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Goodman, M., Ward, K. C., Osunkoya, A. O., Datta, M. W., Luthringer, D., Young, A. N., Marks, K., Cohen, V., Kennedy, J. C., Haber, M. J., et al. (2012). Frequency and determinants of disagreement and error in gleason scores: a population-based study of prostate cancer. *The Prostate*, 72(13):1389–1398.

Graves, A. (2011). Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org.

Hamilton, B. A. (2018). Kaggle. 2018 data science bowl: Find the nuclei in divergent images to advance medical discovery.

Hansen, L. K. and Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001.

Harris, B. (1975). The statistical estimation of entropy in the non-parametric case. Technical report, WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER.

Hassibi, B. and Stork, D. G. (1993). Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in neural information processing systems*, pages 164–171.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer.

Hebb, D. O. (1949). The organization of behavior; a neuropsycholocigal theory. *A Wiley Book in Clinical Psychology*, 62:78.

Heek, J. (2018). Well-calibrated bayesian neural networks. *University of Cambridge*.

Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. (2020). Augmix: A simple method to improve robustness and uncertainty under data shift. In *International Conference on Learning Representations*, volume 2.

Hernández-Lobato, D., Hernandez-Lobato, J., Shah, A., and Adams, R. (2016). Predictive entropy search for multi-objective bayesian optimization. In *International Conference on Machine Learning*, pages 1492–1501.

Hernández-Lobato, J. M. and Adams, R. (2015). Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pages 1861–1869. PMLR.

Hikmet, F., Méar, L., Edvinsson, Å., Micke, P., Uhlén, M., and Lindskog, C. (2020). The protein expression profile of ace2 in human tissues. *Molecular systems biology*, 16(7):e9610.

Hinton, G. E. and Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hoebel, K., Andrearczyk, V., Beers, A., Patel, J., Chang, K., Depeursinge, A., Müller, H., and Kalpathy-Cramer, J. (2020). An exploration of uncertainty information for segmentation quality assessment. In *Medical Imaging 2020: Image Processing*, volume 11313, page 113131K. International Society for Optics and Photonics.

Houlsby, N. (2014). *Efficient Bayesian active learning and matrix modelling*. PhD thesis, University of Cambridge.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *stat*, 1050:24.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, S., Pezzotti, N., Mavroeidis, D., and Welling, M. (2020). Simple and accurate uncertainty quantification from bias-variance decomposition. *arXiv preprint arXiv:2002.05582*.

Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., and Welling, M. (2019). Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–145. Springer.

Huang, K.-H. and Lin, H.-T. (2017). Cost-sensitive label embedding for multi-label classification. *Machine Learning*, 106(9):1725–1746.

Huang, L., He, X., Fang, L., Rabbani, H., and Chen, X. (2019). Automatic classification of retinal optical coherence tomography images with layer guided convolutional neural network. *IEEE Signal Processing Letters*, 26(7):1026–1030.

Hüllermeier, E. and Waegeman, W. (2019). Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction.

Irshad, H., Veillard, A., Roux, L., and Racoceanu, D. (2013). Methods for nuclei detection, segmentation, and classification in digital histopathology: a review - current status and future potential. *IEEE reviews in biomedical engineering*, 7:97–114.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.

Jackson, C. R., Sriharan, A., and Vaickus, L. J. (2020). A machine learning algorithm for simulating immunohistochemistry: development of sox10 virtual ihc and evaluation on primarily melanocytic neoplasms. *Modern Pathology*, 33(9):1638–1648.

Jena, R. and Awate, S. P. (2019). A bayesian neural net to segment images with uncertainty estimates and good calibration. In *International Conference on Information Processing in Medical Imaging*, pages 3–15. Springer.

Jensen, M. H., Jørgensen, D. R., Jalaboi, R., Hansen, M. E., and Olsen, M. A. (2019). Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–548. Springer.

Joseph, J., Roudier, M. P., Narayanan, P. L., Augulis, R., Ros, V. R., Pritchard, A., Gerrard, J., Laurinavicius, A., Harrington, E. A., Barrett, J. C., et al. (2019). Proliferation tumour marker network (ptm-net) for the identification of tumour region in ki67 stained breast cancer whole slide images. *Scientific reports*, 9(1):1–12.

Joskowicz, L., Cohen, D., Caplan, N., and Sosna, J. (2019). Inter-observer variability of manual contour delineation of structures in ct. *European radiology*, 29(3):1391–1399.

Jumeau, F., Com, E., Lane, L., Duek, P., Lagarrigue, M., Lavigne, R., Guillot, L., Rondel, K., Gateau, A., Melaine, N., et al. (2015). Human spermatozoa as a model for detecting missing proteins in the context of the chromosome-centric human proteome project. *Journal of proteome research*, 14(9):3606–3620.

Jungo, A., Balsiger, F., and Reyes, M. (2020). Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14:282.

Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., and Reyes, M. (2018a). On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 682–690. Springer.

Jungo, A., Meier, R., Ermis, E., Herrmann, E., and Reyes, M. (2018b). Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation. *arXiv preprint arXiv:1806.03106*.

Jungo, A. and Reyes, M. (2019). Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer.

Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., Sjöstedt, E., Butler, L., Odeberg, J., Dusart, P., et al. (2021). A single–cell type transcriptomics map of human tissues. *Science Advances*, 7(31):eabh2169.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584.

Khairnar, P., Thiagarajan, P., and Ghosh, S. (2020). A modified bayesian convolutional neural network for breast histopathology image classification and uncertainty quantification. *arXiv preprint arXiv:2010.12575*.

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516.

Kingma, D. P., Salimans, T., and Welling, M. (2015). Variational dropout and the local reparameterization trick. *stat*, 1050:8.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Krzywinski, M. and Altman, N. (2013). Importance of being uncertain.

Kukar, M., Kononenko, I., et al. (1998). Cost-sensitive learning with neural networks. In *ECAI*, volume 98, pages 445–449.

Kuleshov, V., Fenner, N., and Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*.

Kull, M., Nieto, M. P., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, pages 12316–12326.

Kumar, A., Rao, A., Bhavani, S., Newberg, J. Y., and Murphy, R. F. (2014). Automated analysis of immunohistochemistry images identifies candidate location biomarkers for cancers. *Proceedings of the National Academy of Sciences*, 111(51):18249–18254.

Kumar, A., Sarawagi, S., and Jain, U. (2018). Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814. PMLR.

Kwon, Y., Won, J.-H., Joon Kim, B., and Paik, M. (2018). Uncertainty quantification using bayesian neural networks in classification: Application to ischemic stroke lesion segmentation. In *Medical Imaging with Deep Learning*.

LaBonte, T., Martinez, C., and Roberts, S. A. (2019). We know where we don't know: 3d bayesian cnns for credible geometric uncertainty. *arXiv preprint arXiv:1910.10793*.

Lacoste-Julien, S., Huszár, F., and Ghahramani, Z. (2011). Approximate inference for the loss-calibrated bayesian. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 416–424.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413.

Laves, M.-H., Ihler, S., Kortmann, K.-P., and Ortmaier, T. (2019a). Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv preprint arXiv:1909.13550*.

Laves, M.-H., Ihler, S., Ortmaier, T., and Kahrs, L. A. (2019b). Quantifying the uncertainty of deep learning-based computer-aided diagnosis for patient safety. *Current Directions in Biomedical Engineering*, 5(1):223–226.

Lê, M., Unkelbach, J., Ayache, N., and Delingette, H. (2016). Sampling image segmentations for uncertainty quantification. *Medical image analysis*, 34:42–51.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, J. and AlRegib, G. (2020). Gradients as a measure of uncertainty in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2416–2420. IEEE.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14.

Lewandowski, A. (2017). Batch normalized deep kernel learning for weight uncertainty. *NIPS*.

Li, B. and Alstrøm, T. S. (2020). On uncertainty estimation in active learning for image segmentation. *arXiv preprint arXiv:2007.06364*.

Li, Y., Chen, J., Xie, X., Ma, K., and Zheng, Y. (2020). Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 614–623. Springer.

Li, Z., Wang, Y., and Yu, J. (2017). Brain tumor segmentation using an adversarial network. In *International MICCAI brainlesion workshop*, pages 123–132. Springer.

Li, Z., Zhang, T., Cheng, S., Zhu, J., and Li, J. (2019). Stochastic gradient hamiltonian monte carlo with variance reduction for bayesian inference. *Machine Learning*, 108(8):1701–1727.

Liao, J., Liu, D., Su, G., and Liu, L. (2021). Recognizing diseases with multivariate physiological signals by a deepcnn-lstm network. *Applied Intelligence*, pages 1–13.

Lindqvist, J., Olmin, A., Lindsten, F., and Svensson, L. (2020). A general framework for ensemble distribution distillation. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Liu, J. Z., Paisley, J., Kioumourtzoglou, M.-A., and Coull, B. (2019a). Accurate uncertainty estimation and decomposition in ensemble learning. *arXiv preprint arXiv:1911.04061*.

Liu, L., Yang, Q., Zhang, M., Wu, Z., and Xue, P. (2019b). Fluorescence lifetime imaging microscopy and its applications in skin cancer diagnosis. *Journal of Innovative Optical Health Sciences*, 12(05):1930004.

Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151.

Long, W., Yang, Y., and Shen, H.-B. (2020). Imploc: a multi-instance deep learning model for the prediction of protein subcellular localization based on immunohistochemistry images. *Bioinformatics*, 36(7):2244–2250.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

Luo, F., Das, A., Chen, J., Wu, P., Li, X., and Fang, Z. (2019). Metformin in patients with and without diabetes: a paradigm shift in cardiovascular disease management. *Cardiovascular diabetology*, 18(1):1–9.

Lyzhov, A., Molchanova, Y., Ashukha, A., Molchanov, D., and Vetrov, D. (2020). Greedy policy search: A simple baseline for learnable test-time augmentation. In *Conference on Uncertainty in Artificial Intelligence*, pages 1308–1317. PMLR.

Ma, J., Lin, F., Wesarg, S., and Erdt, M. (2018). A novel bayesian model incorporating deep neural network and statistical shape model for pancreas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 480–487. Springer.

Ma, J., Wei, Z., Zhang, Y., Wang, Y., Lv, R., Zhu, C., Chen, G., Liu, J., Peng, C., Wang, L., Wang, Y., and Chen, J. (2020). How distance transform maps boost segmentation cnns: An empirical study. In Arbel, T., Ayed, I. B., de Bruijne, M., Descoteaux, M., Lombaert, H., and Pal, C., editors, *Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 479–492. PMLR.

Ma, Y.-A., Chen, T., and Fox, E. B. (2015). A complete recipe for stochastic gradient mcmc. *arXiv preprint arXiv:1506.04696*.

MacKay, D. J. (1992a). The evidence framework applied to classification networks. *Neural computation*, 4(5):720–736.

MacKay, D. J. (1992b). A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472.

Macke, J., Murray, I., and Latham, P. (2013). Estimation bias in maximum entropy models. *Entropy*, 15(8):3109–3129.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32:13153–13164.

Malinin, A. and Gales, M. (2018). Predictive uncertainty estimation via prior networks. page 7047–7058.

Malinin, A., Mlodozeniec, B., and Gales, M. (2020). Ensemble distribution distillation. *8th International Conference on Learning Representations*.

Malone, I. B., Cash, D., Ridgway, G. R., MacManus, D. G., Ourselin, S., Fox, N. C., and Schott, J. M. (2013). Miriad-public release of a multiple time point alzheimer's mr imaging dataset. *NeuroImage*, 70:33–36.

Martel, A. L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M. A., Zhou, S. K., Racoceanu, D., and Joskowicz, L. (2020). *Medical Image Computing and Computer Assisted Intervention-MICCAI 2020: 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I*, volume 12261. Springer Nature.

Martens, J. and Grosse, R. (2015). Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR.

McClure, P., Rho, N., Lee, J. A., Kaczmarzyk, J. R., Zheng, C. Y., Ghosh, S. S., Nielson, D. M., Thomas, A. G., Bandettini, P., and Pereira, F. (2019). Knowing what you know in brain segmentation using bayesian deep neural networks. *Frontiers in neuroinformatics*, 13:67.

Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., and Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878.

Mengel, M., von Wasielewski, R., Wiese, B., Rüdiger, T., Müller-Hermelink, H. K., and Kreipe, H. (2002). Inter-laboratory and inter-observer reproducibility of immunohistochemical assessment of the ki-67 labelling index in a large multi-centre trial. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 198(3):292–299.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mobiny, A., Nguyen, H. V., Moulik, S., Garg, N., and Wu, C. C. (2019). Dropconnect is effective in modeling uncertainty of bayesian deep networks. *arXiv preprint arXiv:1906.04569*.

Moccia, S., Wirkert, S. J., Kenngott, H., Vemuri, A. S., Apitz, M., Mayer, B., De Momi, E., Mattos, L. S., and Maier-Hein, L. (2018). Uncertainty-aware organ classification for surgical data science applications in laparoscopy. *IEEE Transactions on Biomedical Engineering*, 65(11):2649–2659.

Moeskops, P., Viergever, M. A., Mendrik, A. M., De Vries, L. S., Benders, M. J., and Išgum, I. (2016). Automatic segmentation of mr brain images with a convolutional neural network. *IEEE transactions on medical imaging*, 35(5):1252–1261.

Mohammed, B. A. and Al-Ani, M. S. (2020). Review research of medical image analysis using deep learning. *UHD Journal of Science and Technology*, 4(2):75–90.

Morriss, N. J., Conley, G. M., Ospina, S. M., Meehan III, W. P., Qiu, J., and Mannix, R. (2020). Automated quantification of immunohistochemical staining of large animal brain tissue using qupath software. *Neuroscience*, 429:235–244.

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Nagpal, K., Foote, D., Liu, Y., Chen, P.-H. C., Wulczyn, E., Tan, F., Olson, N., Smith, J. L., Mohtashamian, A., Wren, J. H., et al. (2019). Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine*, 2(1):1–10.

Nair, T., Precup, D., Arnold, D. L., and Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557.

Natekar, P., Kori, A., and Krishnamurthi, G. (2020). Demystifying brain tumor segmentation networks: interpretability and uncertainty analysis. *Frontiers in computational neuroscience*, 14:6.

Neal, R. M. (1993). Bayesian learning via stochastic dynamics. In *Advances in neural information processing systems*, pages 475–482.

Neal, R. M. (1996). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Nemeth, C. and Fearnhead, P. (2021). Stochastic gradient markov chain monte carlo. *Journal of the American Statistical Association*, 116(533):433–450.

Newberg, J. and Murphy, R. F. (2008). A framework for the automated analysis of subcellular patterns in human protein atlas images. *Journal of proteome research*, 7(6):2300–2308.

Ng, M., Guo, F., Biswas, L., and Wright, G. A. (2018). Estimating uncertainty in neuralnetworks for segmentation quality control. Technical report, Technical report.

Nguyen, K. and O'Connor, B. (2015). Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

Niu, K., Guo, J., Pan, Y., Gao, X., Peng, X., Li, N., and Li, H. (2020). Multichannel deep attention neural networks for the classification of autism spectrum disorder using neuroimaging and personal characteristic data. *Complexity*, 2020.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. (2019). Measuring calibration in deep learning. In *CVPR Workshops*, pages 38–41.

Oberdiek, P., Rottmann, M., and Gottschalk, H. (2018). Classification uncertainty of deep neural networks based on gradient information. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 113–125. Springer.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Orlando, J. I., Seeböck, P., Bogunović, H., Klimscha, S., Grechenig, C., Waldstein, S., Gerendas, B. S., and Schmidt-Erfurth, U. (2019). U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1441–1445. IEEE.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pages 4026–4034.

Ouyang, W., Winsnes, C. F., Hjelmare, M., Cesnik, A. J., Åkesson, L., Xu, H., Sullivan, D. P., Dai, S., Lan, J., Jinmo, P., et al. (2019). Analysis of the human protein atlas image classification competition. *Nature methods*, 16(12):1254–1261.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*.

Ozdemir, O., Woodward, B., and Berlin, A. A. (2017). Propagating uncertainty in multi-stage bayesian convolutional neural networks with application to pulmonary nodule detection. *arXiv preprint arXiv:1712.00497*.

Papadopoulos, H., Vovk, V., and Gammerman, A. (2007). Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 388–395. IEEE.

Papamarkou, T., Hinkle, J., Young, M. T., and Womble, D. (2019). Challenges in markov chain monte carlo for bayesian neural networks. *arXiv preprint arXiv:1910.06539*.

Patel, B. N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., Rajpurkar, P., Amrhein, T., Gupta, R., Halabi, S., et al. (2019). Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine*, 2(1):1–10.

Perrone, M. P. and Cooper, L. N. (1992). When networks disagree: Ensemble methods for hybrid neural networks. Technical report, BROWN UNIV PROVIDENCE RI INST FOR BRAIN AND NEURAL SYSTEMS.

Pineau, C., Hikmet, F., Zhang, C., Oksvold, P., Chen, S., Fagerberg, L., Uhlén, M., and Lindskog, C. (2019). Cell type-specific expression of testis elevated genes based on transcriptomics and antibody-based proteomics. *Journal of proteome research*, 18(12):4215–4230.

Prassni, J.-S., Ropinski, T., and Hinrichs, K. (2010). Uncertainty-aware guided volume segmentation. *IEEE transactions on visualization and computer graphics*, 16(6):1358–1365.

Qiu, X., Meyerson, E., and Miikkulainen, R. (2019). Quantifying point-prediction uncertainty in neural networks via residual estimation with an i/o kernel. *arXiv preprint arXiv:1906.00588*.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43(3/4):353–360.

Rączkowski, Ł., Możejko, M., Zambonelli, J., and Szczurek, E. (2019). Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning. *Scientific reports*, 9(1):1–12.

Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Kleinberg, B., Mullainathan, S., and Kleinberg, J. (2019). Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, pages 5281–5290.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Ramalho, T. and Miranda, M. (2020). Density estimation in representation space to predict model uncertainty. In *International Workshop on Engineering Dependable and Secure Machine Learning Systems*, pages 84–96. Springer.

Ravanbakhsh, M., Klein, T., Batmanghelich, K., and Nabi, M. (2019). Uncertainty-driven semantic segmentation through human-machine collaborative learning. *arXiv preprint arXiv:1909.00626*.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). Science forum: the human cell atlas. *elife*, 6:e27041.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Chen, X., and Wang, X. (2020). A survey of deep active learning. *arXiv preprint arXiv:2009.00236*.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

Ritelli, M., Dordoni, C., Venturini, M., Chiarelli, N., Quinzani, S., Traversa, M., Zoppi, N., Vascellaro, A., Wischmeijer, A., Manfredini, E., et al. (2013). Clinical and molecular characterization of 40 patients with classic ehlers–danlos syndrome: identification of 18 col5a1 and 2 col5a2 novel mutations. *Orphanet journal of rare diseases*, 8(1):1–19.

Ritter, H., Botev, A., and Barber, D. (2018). A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, volume 6. International Conference on Representation Learning.

Ronneberger, O., Fischer, P., and Brox, T. (Springer, 2015). U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical image computing and computer-assisted intervention*, pages 234–241.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 20(1):23–38.

Roy, A. G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A. D. N., et al. (2019). Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22.

Ruhe, D., Cina, G., Tonutti, M., de Bruin, D., and Elbers, P. (2019). Bayesian modelling in practice: Using uncertainty to improve trustworthiness in medical applications. *arXiv preprint arXiv:1906.08619*.

Saad, A., Möller, T., and Hamarneh, G. (2010). Probexplorer: Uncertainty-guided exploration and editing of probabilistic medical image segmentation. In *Computer Graphics Forum*, volume 29, pages 1113–1122. Wiley Online Library.

Saha, M., Chakraborty, C., Arun, I., Ahmed, R., and Chatterjee, S. (2017). An advanced deep learning approach for ki-67 stained hotspot detection and proliferation rate scoring for prognostic evaluation of breast cancer. *Scientific reports*, 7(1):1–14.

Salimans, T. and Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29:901–909.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Sensoy, M., Kaplan, L., and Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. page 3179–3189.

Settles, B. (2009). Active learning literature survey. *Computer Sciences Technical Report 1648, University of Wisconsin, Department of Computer Science*.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Shan+, F., Gao+, Y., Wang, J., Shi, W., Shi, N., Han, M., Xue, Z., Shen, D., and Shi, Y. (2020). Lung infection quantification of covid-19 in ct images with deep learning. *arXiv preprint arXiv:2003.04655*.

Shannon, C. E. (July–October 1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*.

Shu, Y., Cao, Z., Long, M., and Wang, J. (2019). Transferable curriculum for weakly-supervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4951–4958.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C., Oksvold, P., Edfors, F., Limiszewska, A., Hikmet, F., et al. (2020). An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science*, 367(6482).

Smith, L. and Gal, Y. (2018). Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*.

Soberanis-Mukul, R. D., Navab, N., and Albarqouni, S. (2020). Uncertainty-based graph convolutional networks for organ segmentation refinement. In *Medical Imaging with Deep Learning*, pages 755–769. PMLR.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Staal, J., Abramoff, M. D., Niemeijer, M., Viergever, M. A., and Van Ginneken, B. (2004). Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509.

Stålhammar, G., Martinez, N. F., Lippert, M., Tobin, N. P., Mølholm, I., Kis, L., Rosin, G., Rantalainen, M., Pedersen, L., Bergh, J., et al. (2016). Digital image analysis outperforms manual biomarker assessment in breast cancer. *Modern Pathology*, 29(4):318–329.

Stenman, S., Bychkov, D., Kücükel, H., Linder, N., Haglund, C., Arola, J., and Lundin, J. (2020). Antibody supervised training of a deep learning based algorithm for leukocyte segmentation in papillary thyroid carcinoma. *IEEE Journal of Biomedical and Health Informatics*, 25(2):422–428.

Stiles, J. and Jernigan, T. L. (2010). The basics of brain development. *Neuropsychology review*, 20(4):327–348.

Sullivan, D. P., Winsnes, C. F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., Campbell, L., Leifsson, H., Rhodes, S., Nordgren, A., et al. (2018). Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature biotechnology*, 36(9):820–828.

Sundermeyer, M., Schlüter, R., and Ney, H. (2012). Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Swiderska-Chadaj, Z., Pinckaers, H., van Rijthoven, M., Balkenhol, M., Melnikova, M., Geessink, O., Manson, Q., Sherman, M., Polonia, A., Parry, J., et al. (2019). Learning to detect lymphocytes in immunohistochemistry with deep learning. *Medical image analysis*, 58:101547.

Tanno, B., Leonardi, S., Babini, G., Giardullo, P., De Stefano, I., Pasquali, E., Saran, A., and Mancuso, M. (2017a). Nanog-driven cell-reprogramming and self-renewal maintenance in ptch1+/- granule cell precursors after radiation injury. *Scientific reports*, 7(1):1–11.

Tanno, R., Worrall, D., Kaden, E., Ghosh, A., Grussu, F., Bizzi, A., Sotiropoulos, S. N., Criminisi, A., and Alexander, D. C. (2019). Uncertainty quantification in deep learning for safer neuroimage enhancement. *arXiv preprint arXiv:1907.13418*.

Tanno, R., Worrall, D. E., Ghosh, A., Kaden, E., Sotiropoulos, S. N., Criminisi, A., and Alexander, D. C. (2017b). Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–619. Springer.

Tellez, D., Balkenhol, M., Otte-Höller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., et al. (2018). Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136.

Tewary, S., Arun, I., Ahmed, R., Chatterjee, S., and Mukhopadhyay, S. (2021). Autoihc-analyzer: computer-assisted microscopy for automated membrane extraction/scoring in her2 molecular markers. *Journal of Microscopy*, 281(1):87–96.

Teye, M., Azizpour, H., and Smith, K. (2018). Bayesian uncertainty estimation for batch normalized deep networks. *arXiv preprint arXiv:1802.06455*.

Thommen, D. S., Koelzer, V. H., Herzig, P., Roller, A., Trefny, M., Dimeloe, S., Kiialainen, A., Hanhart, J., Schill, C., Hess, C., et al. (2018). A transcriptionally and functionally distinct pd-1+ cd8+ t cell pool with predictive potential in non-small-cell lung cancer treated with pd-1 blockade. *Nature medicine*, 24(7):994–1004.

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Blal, H. A., Alm, T., Asplund, A., Björk, L., Breckels, L. M., et al. (2017). A subcellular map of the human proteome. *Science*, 356(6340).

Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., and Michalak, S. (2019). On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *arXiv preprint arXiv:1905.11001*.

Tousignant, A., Lemaître, P., Precup, D., Arnold, D. L., and Arbel, T. (2019). Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data. In *International Conference on Medical Imaging with Deep Learning*, pages 483–492. PMLR.

Turney, P. D. (1994). Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of artificial intelligence research*, 2:369–409.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. *Science*, 347(6220).

Uhlen, M., Karlsson, M. J., Zhong, W., Tebani, A., Pou, C., Mikes, J., Lakshmikanth, T., Forsström, B., Edfors, F., Odeberg, J., et al. (2019). A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science*, 366(6472).

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al. (2017). A pathology atlas of the human cancer transcriptome. *Science*, 357(6352).

Ulmer, A., Beutel, J., Süsskind, D., Hilgers, R.-D., Ziemssen, F., Lüke, M., Röcken, M., Rohrbach, M., Fierlbeck, G., Bartz-Schmidt, K.-U., et al. (2008). Visualization of circulating melanoma cells in peripheral blood of patients with primary uveal melanoma. *Clinical Cancer Research*, 14(14):4469–4474.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. B. (2019). Evaluating model calibration in classification. *arXiv preprint arXiv:1902.06977*.

Valen, D. A. V., Kudo, T., Lane, K. M., Macklin, D. N., Quach, N. T., DeFelice, M. M., Maayan, I., Tanouchi, Y., Ashley, E. A., and Covert, M. W. (nov 2016). Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLOS Computational Biology*, 12(11):e1005177.

Valkonen, M., Isola, J., Ylinen, O., Muhonen, V., Saxlin, A., Tolonen, T., Nykter, M., and Ruusuvuori, P. (2019). Cytokeratin-supervised deep learning for automatic recognition of epithelial cells in breast cancers stained for er, pr, and ki-67. *IEEE Transactions on Medical Imaging*, 39(2):534–542.

Van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. (2020). Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR.

Van Eycke, Y.-R., Balsat, C., Verset, L., Debeir, O., Salmon, I., and Decaestecker, C. (2018). Segmentation of glandular epithelium in colorectal tumours to automatically compartmentalise ihc biomarker quantification: A deep learning approach. *Medical image analysis*, 49:35–45.

Vandenbrouck, Y., Lane, L., Carapito, C., Duek, P., Rondel, K., Bruley, C., Macron, C., Gonzalez de Peredo, A., Coute, Y., Chaoui, K., et al. (2016). Looking for missing proteins in the proteome of human spermatozoa: an update. *Journal of Proteome Research*, 15(11):3998–4019.

Vandenbrouck, Y., Pineau, C., and Lane, L. (2020). The functionally unannotated proteome of human male tissues: A shared resource to uncover new protein functions associated with reproductive biology. *Journal of Proteome Research*, 19(12):4782–4794.

Varadarajan, A. V., Bavishi, P., Ruamviboonsuk, P., Chotcomwongse, P., Venugopalan, S., Narayanaswamy, A., Cuadros, J., Kanai, K., Bresnick, G., Tadarati, M., et al. (2020). Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. *Nature communications*, 11(1):1–8.

Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., and Fergus, R. (2013). Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45.

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI brainlesion workshop*, pages 178–190. Springer.

Wang, H. and Yeung, D.-Y. (2016). Towards bayesian deep learning: A survey. *arXiv preprint arXiv:1604.01662*.

Wang, J., Bao, Y., Wen, Y., Lu, H., Luo, H., Xiang, Y., Li, X., Liu, C., and Qian, D. (2020a). Prior-attention residual learning for more discriminative covid-19 screening in ct images. *IEEE Transactions on Medical Imaging*, 39(8):2572–2583.

Wang, K., Zhang, D., Li, Y., Zhang, R., and Lin, L. (2016). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., and He, Z. (2020b). Double-uncertainty weighted method for semi-supervised learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 542–551. Springer.

Warfield, S. K., Zou, K. H., and Wells, W. M. (2002). Validation of image segmentation and expert quality with an expectation-maximization algorithm. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 298–306. Springer.

Wei, J. N., Duvenaud, D., and Aspuru-Guzik, A. (2016). Neural networks for the prediction of organic chemistry reactions. *ACS central science*, 2(10):725–732.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688.

Wen, Y., Tran, D., and Ba, J. (2020). Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *8 th International Conference on Learning Representations*.

Wenger, J., Kjellström, H., and Triebel, R. (2019). Non-parametric calibration for classification. *arXiv preprint arXiv:1906.04933*.

Wickstrøm, K., Kampffmeyer, M., and Jenssen, R. (2020). Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical image analysis*, 60:101619.

Wilson, R. and Spann, M. (1988). *Image segmentation and uncertainty*. John Wiley & Sons, Inc.

Wu, X.-Z. and Zhou, Z.-H. (2017). A unified view of multi-label performance measures. In *International Conference on Machine Learning*, pages 3780–3788. PMLR.

Xia, Y., Liu, F., Yang, D., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., and Roth, H. (2020). 3d semi-supervised learning with uncertainty-aware multi-view co-training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3646–3655.

Xu, J., Luo, X., Wang, G., Gilmore, H., and Madabhushi, A. (2016). A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing*, 191:214–223.

Xu, Y.-Y., Yang, F., Zhang, Y., and Shen, H.-B. (2013). An image-based multi-label human protein subcellular localization predictor (i locator) reveals protein mislocalizations in cancer tissues. *Bioinformatics*, 29(16):2032–2040.

Yang, H., Shan, C., Kolen, A. F., et al. (2020). Deep q-network-driven catheter segmentation in 3d us by hybrid constrained semi-supervised learning and dual-unet. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 646–655. Springer.

Ye, C., Li, Y., and Zeng, X. (2020). An improved deep network for tissue microstructure estimation with uncertainty quantification. *Medical image analysis*, 61:101650.

Yu, S., Xiao, D., Frost, S., and Kanagasingam, Y. (2019). Robust optic disc and cup segmentation with deep learning for glaucoma detection. *Computerized Medical Imaging and Graphics*, 74:61–71.

Yuan, Y. and Bar-Joseph, Z. (2019). Deep learning for inferring gene relationships from single-cell expression data. *Proceedings of the National Academy of Sciences*, 116(52):27151–27158.

Zhang, L., Chang, M., Beck, C. A., Schwarz, E. M., and Boyce, B. F. (2016). Analysis of new bone, cartilage, and fibrosis tissue in healing murine allografts using whole slide imaging and a new automated histomorphometric algorithm. *Bone research*, 4(1):1–9.

Zhang, L. and Ji, Q. (2011). A bayesian network model for automatic and interactive image segmentation. *IEEE Transactions on Image Processing*, 20(9):2582–2593.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., and Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*.

Zhou, S. K., Greenspan, H., and Shen, D. (2017). *Deep learning for medical image analysis*. Academic Press.

Zhou, X., Li, C., Rahaman, M. M., Yao, Y., Ai, S., Sun, C., Wang, Q., Zhang, Y., Li, M., Li, X., et al. (2020). A comprehensive review for breast histopathology image analysis using classical and deep neural networks. *IEEE Access*, 8:90931–90956.

Zhou, Z.-H. and Liu, X.-Y. (2010). On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257.