Trustworthy Deep Learning for Medical Image Analysis: Exploring Uncertainty and Calibration

A Thesis Submitted

in Partial Fulfilment of the Requirements

for the Degree of

Doctor of Philosophy

by

Abhishek Singh Sambyal

(2019CSZ0001)

Department of Computer Science and Engineering

Indian Institute of Technology Ropar

April, 2025

Dedicated to
My Mahadev, My Bharat and My Family

Declaration of Originality

I hereby declare that the work which is being presented in the thesis entitled **"Trustworthy Deep Learning for Medical Image Analysis: Exploring Uncertainty and Calibration"** has been solely authored by me. It presents the result of my own independent investigation/research conducted during the time period from July 2019 to December 2024 under the supervision of Dr. Deepti R. Bathula, Associate Professor, Department of Computer Science and Engineering, Indian Institute of Technology Ropar (IIT Ropar) and Dr. Narayanan C. Krishnan, Associate Professor, Department of Data Science, Indian Institute of Technology Palakkad (IIT Palakkad). To the best of my knowledge, it is an original work, both in terms of research content and narrative, and has not been submitted or accepted elsewhere, in part or in full, for the award of any degree, diploma, fellowship, associateship, or similar title of any university or institution. Further, due credit has been attributed to the relevant state-of-the-art and collaborations (if any) with appropriate citations and acknowledgments in line with established ethical norms and practices. I also declare that any idea/data/fact/source stated in my thesis has not been fabricated/ falsified/ misrepresented. All the principles of academic honesty and integrity have been followed. I fully understand that if the thesis is found to be unoriginal, fabricated, or plagiarized, the Institute reserves the right to withdraw the thesis from its archive and revoke the associated Degree conferred. Additionally, the Institute also reserves the right to appraise all concerned sections of society of the matter for their information and necessary action (if any). If accepted, I hereby consent for my thesis to be available online in the Institute's Open Access repository, inter-library loan, and the title & abstract to be made available to outside organizations.

Signature of the Student

Name: Abhishek Singh Sambyal
Entry Number: 2019CSZ0001
Program: Ph.D.
Department: Computer Science and Engineering
Indian Institute of Technology Ropar
Rupnagar, Punjab 140001

Date: 11/04/2025

Acknowledgement

I express my sincere gratitude to my supervisors Dr. Deepti R. Bathula and Dr. Narayanan C. Krishnan for taking me as their student and providing continuous guidance at every stage of my Ph.D. degree. I am also thankful to them for giving me complete freedom to pursue my interests and helping me through all the problems I faced during my stay at IIT Ropar.

I would also like to express my gratitude to the members of my Doctoral Committee, Dr. Nitin Auluck, Dr. Shashi Shekhar Jha, Dr. T.V. Kalyan and Dr. Manju Khan for their insightful comments, constructive criticism, and valuable suggestions. Their guidance and expertise have been instrumental in shaping the direction and outcome of my research.

I am also grateful to my parents, brother, friends, and all other family members for their immense support, without which completing my degree would not have been possible.

Certificate

This is to certify that the thesis entitled **"Trustworthy Deep Learning for Medical Image Analysis: Exploring Uncertainty and Calibration"**, submitted by **Abhishek Singh Sambyal (2019CSZ0001)** for the award of the degree of **Doctor of Philosophy** of Indian Institute of Technology Ropar, is a record of bonafide research work carried out under my guidance and supervision. To the best of my knowledge and belief, the work presented in this thesis is original and has not been submitted, either in part or full, for the award of any other degree, diploma, fellowship, associateship, or similar title of any university or institution.
In my opinion, the thesis has reached the standard of fulfilling the requirements of the regulations relating to the Degree.

Signature of the Supervisor

Dr. Deepti R. Bathula
Dept. of Computer Science & Engineering
IIT Ropar
Rupnagar, Punjab 140001

Date:

Signature of the Supervisor

Dr. Narayanan C. Krishnan
Dept. of Data Science
IIT Palakkad
Palakkad, Kerala 678623

Date:

Lay Summary

Modern deep neural networks (DNNs) achieve remarkable performance across various domains such as healthcare, autonomous driving, and weather forecasting. However, relying solely on performance metrics can sometimes lead to unsuitable outcomes because performance alone does not capture the trustworthiness of the model. This highlights the need for trustworthy deep learning, which emphasizes creating reliable, robust, transparent, and fair algorithms to support better decision-making. In this context, this thesis focuses on reducing the uncertainty in model outputs to improve trust in its predictions. Additionally, as the predicted probabilities or confidence scores of DNNs can vary under different conditions, an empirical study was conducted to analyze these variations. Building on these insights, a novel approach was proposed to improve the reliability of the model's confidence scores. By exploring these aspects, this thesis aims to advance the development of trustworthy models that are both robust and reliable for decision-making in healthcare.

Abstract

Performance plays an important role when selecting a DNN for decision-making. However, in healthcare, relying on a DNN's decision without understanding its certainty or calibration can be risky. This thesis focuses on addressing the challenges of analyzing uncertainty and calibration in DNNs, specifically for medical imaging tasks such as segmentation and classification.

Uncertainty in predictions arises from data noise (aleatoric) and flawed model inferences (epistemic). While epistemic uncertainty can be mitigated with more data or larger models, addressing aleatoric uncertainty is more challenging. This work aims to reduce aleatoric uncertainty in a downstream segmentation task through a two-stage approach: (a) a self-supervised task, specifically a reconstruction task, to estimate aleatoric uncertainty with predictions, akin to learning the output distribution, and (b) leveraging additional samples from the learned distribution to reduce aleatoric uncertainty in the segmentation task. Sampling from high-uncertainty regions in the reconstruction highlights areas where the model is less confident, and incorporating these samples improves predictions. The proposed method, evaluated on the benchmark Brain Tumor Segmentation (BraTS) dataset, demonstrated a significant reduction in aleatoric uncertainty for segmentation task while achieving performance comparable to or better than standard augmentation techniques.

To investigate the calibration of DNNs, this thesis focused on two key aspects: first, understanding how confidence calibration is affected under varying conditions, and second, improving the calibration of DNNs so that their probability scores can be reliably used for decision-making. To address the first aspect, a comprehensive empirical study was conducted to evaluate performance and calibration across different scenarios. The experiments involved combinations of three medical imaging datasets, four dataset sizes (ranging from small to large), three architecture sizes (small to large), and three training regimes (fully supervised and self-supervised, with and without pretraining). Additionally, the study examined the factors within DNNs that influence changes in calibration. Key findings include: (a) self-supervised learning improves calibration without compromising performance, (b) dataset characteristics significantly impact both calibration and performance, and (c) employing multiple calibration metrics is crucial for a comprehensive evaluation of calibration error, as relying on a single metric can lead to misleading conclusions.

To improve the calibration of DNNs, various methods have been proposed, ranging from post-hoc adjustments to train-time strategies. However, these approaches often come at the cost of reduced performance. Moreover, many techniques focus on improving calibration for the most confident predicted class rather than addressing calibration across all classes. This thesis aims to improve class-wise calibration without compromising performance. To achieve this, a novel method called Label Smoothing Plus (LS+) was introduced. LS+ incorporates class-wise priors, estimated from validation set accuracies, during training to produce better-calibrated predictions. The proposed approach was evaluated on three

benchmark medical imaging datasets, including one with significant class imbalance, using multiple performance and calibration metrics across two architectures. The results demonstrated a significant reduction in miscalibration, with the predicted confidence scores proving highly suitable for clinical decision-making.

Keywords: Trustworthy deep learning; Uncertainty quantification; Confidence calibration.

List of Publications

(most recent first)

Conferences

- **Abhishek Singh Sambyal**, Usma Niyaz, Saksham Shrivastava, Narayanan C. Krishnan, Deepti R. Bathula, "LS+: Informed Label Smoothing for Improving Calibration in Medical Image Classification", 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2024.

- **Abhishek Singh Sambyal**, Usma Niyaz, Narayanan C. Krishnan, Deepti R. Bathula, "Medical Image Analysis - Does Self-Supervised Learning Improve Calibration?", Bern Interpretable AI Symposium (BIAS), University of Bern, 2023. *(Abstract Submission)*

- **Abhishek Singh Sambyal**, Narayanan C. Krishnan, Deepti R. Bathula, "Towards Reducing Aleatoric Uncertainty for Medical Imaging Task", 19th International Symposium on Biomedical Imaging (ISBI), 2022.

Journal

- **Abhishek Singh Sambyal**, Usma Niyaz, Narayanan C. Krishnan, Deepti R. Bathula, "Understanding Calibration of Deep Neural Networks for Medical Image Classification", Computer Methods and Programs in Biomedicine. 2023.

Other Publications

- Usma Niyaz, **Abhishek Singh Sambyal**, Deepti R. Bathula, "Wavelet-Based Feature Compression for Improved Knowledge Distillation", 21st International Symposium on Biomedical Imaging (ISBI), 2024.

- Arunava Chaudhuri, **Abhishek Singh Sambyal**, Deepti R. Bathula, "Mutually Exclusive Multi-Modal Approach for Parkinson's Disease Classification", 17th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOIMAGING, 2024.

- Usma Niyaz, **Abhishek Singh Sambyal**, Deepti R. Bathula, "Leveraging Different Learning Styles for Improved Knowledge Distillation in Biomedical Imaging", Computer Methods and Programs in Biomedicine. 2024.

- Arunava Chaudhuri, **Abhishek Singh Sambyal**, Deepti R. Bathula, "MEM4S: Mutually Exclusive Multi-Modal Mix-n-Match Strategy For Parkinson's Disease Classification", *(Submitted for peer review)*

Contents

Declaration iv

Acknowledgement v

Certificate vi

Lay Summary vii

Abstract viii

List of Publications x

List of Figures xv

List of Tables xix

List of Symbols and Abbreviations xxi

1 Introduction 1

1.1 Deep Learning in Medical Image Analysis 1

1.2 Trustworthy Deep Learning 1

1.3 Uncertainty and Calibration 5

1.3.1 Uncertainty 5

1.3.2 Calibration 7

1.4 Thesis Organization 9

2 Literature Review 11

2.1 Trustworthy Deep Learning 11

2.2 Uncertainty Quantification 15

2.3 Calibration 19

3 Towards Reducing Aleatoric Uncertainty for Medical Imaging Tasks 27
3.1 Introduction . . . 27
3.2 Motivation . . . 28
3.3 Methodology . . . 29
3.3.1 Modeling Uncertainty . . . 29
3.3.2 Interpreting Aleatoric Uncertainty from Reconstruction Task as a Noise Model . . . 35
3.3.3 Using the Noise Model to Reduce Aleatoric Uncertainty in the Segmentation Task . . . 36
3.4 Results and Discussion . . . 36
3.4.1 Experimental Setting . . . 36
3.4.2 Quantitative Results . . . 37
3.4.3 Qualitative Results . . . 40
3.4.4 Discussion . . . 41
3.5 Conclusion . . . 41

4 Understanding Calibration of Deep Neural Networks for Medical Image Classification 43
4.1 Introduction . . . 43
4.2 Methods . . . 45
4.2.1 Training Regimes . . . 45
4.2.2 Calibration Metrics . . . 46
4.2.3 Experimental Setup . . . 49
4.3 Results . . . 50
4.3.1 Effect of Training Regimes on Calibration . . . 50
4.3.2 Effect of Architecture and Dataset Size . . . 51
4.3.3 Issues with using Single Calibration Metric . . . 54
4.3.4 Factors affecting Performance and Calibration . . . 55
4.3.5 RadImageNet Pretraining . . . 65
4.3.6 Comparison of Fully-Supervised and Reconstruction-Based Self-Supervised Task . . . 66
4.4 Discussion and Conclusion . . . 66

5 LS+: Informed Label Smoothing for Improving Calibration in Medical Image Classification 69
5.1 Introduction . . . 69
5.2 Related Work . . . 70
5.3 Methodology . . . 70
5.3.1 Preliminaries . . . 70
5.3.2 Label Smoothing Plus (LS+) . . . 71
5.4 Experiments and Results . . . 72
5.4.1 Experimental Setting . . . 72

5.4.2 Calibration performance comparison with SOTA 73
5.4.3 Uncertainty-based Retention Curves 75
5.4.4 Clinical Significance of Predicted Confidence Scores 76
5.4.5 Discussion 76
5.5 Ablation Studies 78
5.5.1 Comparison with Temperature Scaling 78
5.5.2 Standard training (HL) followed by fine-tuning with LS approach . . 80
5.5.3 Replacing validation set class-wise accuracy with constant value . . . 80
5.6 Conclusion 82

6 Summary and Future Work **83**
6.1 Thesis Summary 83
6.2 Impact and Applications 85
6.3 Future work 86
6.3.1 Unexplored Questions/Existing Gaps 86

References **89**

List of Figures

1.1 Facets of Trustworthy Deep Learning. 2

1.2 Uncertainties in Deep Neural Network. 5

3.1 1^{st} and 2^{nd} row represent uncertainty in regression an classification tasks, respectively whereas, 1^{st} and 2^{nd} column represent data and model uncertainty. Illustration credits [1]. 28

3.2 **Uncertainties in the reconstruction task.** The input is processed by a heteroscedastic NN with two output heads: **(a)***[top]* reconstruction, and **(b)***[bottom]* uncertainty head. The explicit uncertainty, denoted as $\hat{\sigma}_i$, represents the aleatoric uncertainty associated with the predictions, specifically every pixel of the reconstructed image. The NN is used with MC Dropout and the input image is passed through the model multiple times, resulting in multiple predictions and corresponding uncertainty maps. The variance across these predictions represents the epistemic uncertainty, while the average of the multiple uncertainty maps provides the mean aleatoric uncertainty (Eqn. 3.3). 30

3.3 **Visualization of the reconstruction output of the images (slices) from the test set.** Reconstructed images show high fidelity, corroborated by Structural Similarity Index Measure (SSIM) loss images. Aleatoric uncertainty is prominent where the reconstruction model finds ambiguity. . 34

3.4 **Uncertainties in the segmentation task.** The heteroscedastic regression NN is used to estimate the distribution over the *logit space*. Here, $g_i^{\mathbf{W}}$ represents the raw prediction (real values), and $\hat{\sigma}_i$ denotes the aleatoric uncertainty associated with these predictions (logits). From the learned distribution, a new logit is sampled and subsequently passed through the softmax function to generate the segmentation mask. Here, both the segmentation output and uncertainty map are computed channel-wise. . . . 35

3.5 **Proposed Approach.** `Stage 1`: **(a)** Estimate aleatoric uncertainty in an auxiliary self-supervised task (image reconstruction). **(b)** Aleatoric uncertainty is associated with the prediction of the model. But in this task, prediction is the reconstructed input. So, we interpret the aleatoric uncertainty as noise present in the input data. `Stage 2`: (a) We leverage aleatoric uncertainty (interpreted as noise in the input) estimated from `Stage 1` to generate new images. It can be viewed as a data augmentation process. **(b)** The augmented dataset is used to train the segmentation model which reduces the aleatoric uncertainty in the segmentation task. 37

3.6 **Comparison of prediction and aleatoric uncertainty maps.** *Row 1*: Input and Ground Truth Images; *Row 2-5*: Predicted mask and Uncertainty Map of Baseline, Gaussian, Ours, Full Augmentation methods. 40

4.1 Self-Supervised Learning Framework 46

4.2 **Joint evaluation for performance and calibration across different dataset sizes (x-axis) and architectures for DR dataset.** The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials. The underline shows the statistical significance between FS_p and SSL_p. Black and Pink color signifies $p < 0.05$ and $0.05 < p < 0.1$ level of significance, respectively. ↑: higher is better, ↓: lower is better. 52

4.3 **Joint evaluation for performance and calibration across different dataset sizes (x-axis) and architectures for Histopathology Cancer dataset.** The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials. The underline shows the statistical significance between FS_p and SSL_p. Black and Pink color signifies $p < 0.05$ and $0.05 < p < 0.1$ level of significance, respectively. 53

4.4 **Comparing performance and calibration across different architectures and training regimes for *Covid-19* dataset.** The error bars correspond to $\mu \pm \sigma$, estimated over 3 trials. Relying on a single calibration error metric, such as ECE or ACE, can lead to conflicting conclusions when it comes to model selection. By considering a combination of metrics, we gain a more comprehensive understanding of the model's calibration performance. ↑: higher is better, ↓: lower is better. 54

4.5 **Comparing different aspects of WideResNet learned weights for dataset size 10000 on DR-(a) and Histopathology Cancer-(b) datasets.** (1) and (2) the normalized histogram of weights of three training regimes. (3) Layer-wise comparison of standard deviation (SD) between FS_p and SSL_p. (4) Layer-wise comparison of Frobenius norm between FS_p and SSL_p. 57

4.6 **Comparing calibration metrics (x-axis) vs. standard deviation (SD, y-axis)** of WideResNet architecture for dataset size 10000 on DR and Histopathology cancer datasets. Colors represent training regimes (orange for FS_r, blue for SSL_p, and red for FS_p), and markers are the lowercase initials of each calibration metric; e – ECE, o – OE, a – ACE, m – MCE, b – Brier, n – NLL. Alongside each calibration error cluster, the performance is also reported. Ideally, the metrics should be at the bottom left with comparable performance. **(a)** SSL_p has less calibration error with on-par performance than FS_p training regime, indicating it to be a suitable choice. Calibration error metrics clusters of SSL_p and FS_p are noticeably well separated, correlating with the gap in their SD. **(b)** Here, SSL_p seems to be the best in calibration and performance compared to other training regimes. The noticeable difference we observed here is that the calibration error metrics clusters of SSL_p and FS_p are close (somewhat overlapping) when the SD of their weight distributions are similar. 59

4.7 **Standard Deviation of Weights distribution vs. Calibration scores analysis.** (a), (b), (c), and (d) depict the relationship between the SD of weights distribution and calibration metrics from the smallest dataset size to the largest one (500, 1000, 1000, 10000), respectively of the DR dataset. Additionally, the corresponding weight distribution plots have been overlaid for convenience of reference. Considering the four plots, we can observe the trend that the calibration metrics of different regimes are segregated when there is a difference in the spread of their distributions (as shown in plots c & d) and overlapping when there is no difference in the SD of weights distribution (as shown in plots a & b). Based on the characteristics of SSL_p (shown in blue), it can be remarked that a balance in the spread of weights is necessary to achieve both good performance and calibration. 60

4.8 **CKA plots of trained WideResNet architecture using fully-supervised (pretrained, FS_p) and self-supervised (pretrained, SSL_p) regime for DR dataset.** The plots represents similarity between representations of features. The range of the CKA metric is between 0 and 1, with 0 indicating two completely distinct activations (not similar) and 1 indicating two identical activations (similar). 62

4.9 CKA plots of trained ResNet18 and ResNet50 architectures using FS_r, FS_p, and SSL_p regimes for DR dataset. 62

4.10 CKA plots of trained architectures using different regimes for Histopathology Cancer dataset. 63

4.11 **Joint evaluation for performance and calibration across different dataset sizes (x-axis) of DR dataset using ResNet50 architecture with RadImageNet pretraining.** The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials. ↑: higher is better, ↓: lower is better. 65

4.12 **Comparison of fully supervised (FS_r, random initialization), fully supervised (FS_p, pretraining), and reconstruction-based auxiliary SSL task (SSL_p, pretraining) on DR dataset.** Notably, the calibration of models achieved through the auxiliary task does not precisely align with that of the rotation task. Remarkably, the plots reveal a notable contrast: very low OE (f) but high ECE (c). This discrepancy could hint at potential underconfidence, stemming from substantial regularization induced by the reconstruction-based auxiliary SSL task. However, drawing definitive conclusions is premature, as further experiments, encompassing various architectures and hyperparameter tuning, are necessary. Relying solely on the plots, we abstain from making a judgment regarding the superiority of either FS_p or reconstruction-based SSL_p. 66

5.1 $\mathbf{D_T}$: Training Data, $\mathbf{D_V}$: Validation Data, $\mathbf{g(x)}$: Pre-trained model using Hard Labels, $\mathbf{f(x)}$: Model to be calibrated, $\mathbf{v^k}$: class-specific prior, $\mathbf{u}$: uniform prior. *(Top)* LS uses the same prior for all classes. *(Bottom)* LS+ uses class-specific priors computed from the validation set's class-wise accuracy based on the pre-trained model. 72

5.2 **Retention Curves.** Accuracy as a function of retention fraction along with the Area Under the Retention Curve (R-AUC) values using ResNet-34/50 for all three datasets. HL - *Hard Labels*, LS - *Label Smoothing*, FL - *Focal Loss*, DCA - *Difference between Confidence and Accuracy* and MDCA - *Multi-class Difference in Confidence and Accuracy*. 75

5.3 **Comparison of density plots for correct (green) and incorrect (red) classification confidences for ResNet-34 (top) and ResNet-50 (bottom) on MHIST, Chaoyang and ISIC datasets.** For incorrect predictions, LS, FL and Ours provide low confidence which is desirable. However, methods like HL, DCA and MDCA exhibits higher confidence even when they are wrong making them unreliable. The area under the histogram integrates to 1. We have clipped the y-axis in all the plots to better visualize the trends. 77

6.1 Illustration of Thesis Summary 83

6.2 Representing three scenarios of miscalibration (a) under-confidence (b) over-confidence, and (c) near-calibrated. *Image Credits* [2] 87

6.3 Future methods should perform well in all three aspects: Performance, Calibration and OOD Robustness. 87

List of Tables

3.1 Example of aleatoric uncertainty using a House Price Prediction task. *H2* and *H3* are noisy samples, making it impossible to predict the price of the *Test* sample. 29

3.2 Example of aleatoric uncertainty using a House Price Prediction task. The `City` feature is added to the existing dataset. The new feature removes the noise (aleatoric uncertainty) in the dataset and allows the model to predict the price of the *Test* sample. 29

3.3 **Comparison between different implementations.** i) Baseline: Segmentation model with an uncertainty estimation framework. (ii) Gaussian: Gaussian noise is added to the baseline model. (iii) Full Augmentation: Baseline model with four augmentations - (a) affine image transformations - scale, shear, rotate, vertical, horizontal flipping (b) elastic transformations. ↑: Higher is better, ↓: Lower is better; Best results shown in **bold**. Statistical difference between ours and best/2nd best: ** $p < 0.001$ (highly statistically significant) & $p > 0.05$ (statistically non-significant) shown in *Italics*. 39

4.1 Overview of the models used in this study. 50

4.2 Mean CKA values of different training regimes across varying architectures, datasets and their sizes. 64

5.1 **Quantitative Results.** Performance and Calibration results on the test set of three benchmark datasets. The reported values are the average of 3 runs and given as percentages (%) with SD (σ) as subscript. ↑: Higher is better, ↓: Lower is better. Architectures: R34 *(ResNet-34)*, R50 *(ResNet-50)*; Datasets: D1 *(Chaoyang)*, D2 *(MHIST)* and D3 *(ISIC)*. 74

5.2 **Calibration comparison *(+Temperature Scaling)*** results on the test set of Chaoyang & MHIST datasets. The reported values are the average of 3 runs and given as percentages (%) with SD (σ) as subscript. ↑: Higher is better, ↓: Lower is better. 79

5.3 **HL Trained Model** $\xrightarrow{Fine-tune}$ **LS.** Performance and Calibration results on the test set of two benchmark datasets. The reported values are the average of 3 runs and given as percentages (%) with SD (σ) as subscript. ↑: Higher is better, ↓: Lower is better. Architectures: R34 *(ResNet-34)*, R50 *(ResNet-50)*; Datasets: D1 *(Chaoyang)* and D2 *(MHIST)*. 81

5.4 **LS+ with constant values (0.4/0.6).** Performance and Calibration results on the test set of two benchmark datasets. The reported values are the average of 3 runs and given as percentages (%) with SD (σ) as subscript. ↑: Higher is better, ↓: Lower is better. Architectures: R34 *(ResNet-34)*, R50 *(ResNet-50)*; Datasets: D1 *(Chaoyang)* and D2 *(MHIST)*. 81

List of Symbols and Abbreviations

Towards Reducing Aleatoric Uncertainty for Medical Imaging Tasks

$\hat{\psi}$	Sample parameters from an approximated posterior distribution
$\hat{\sigma}_{ij}$	Aleatoric uncertainty of pixel j of image $\mathbf{x}_i$
$\hat{y}_{ij}$	Regressed value of pixel j of image $\mathbf{x}_i$
$\mathbf{W}$	Weights of the segmentation model
$\mathcal{D}$	Dataset
ψ	Parameters of the reconstruction model
$\sigma_i^{\mathbf{W}}$	Uncertainty map associated with output of the segmentation model on image i
$g_i^{\mathbf{W}}$	Real value output using the weight $\mathbf{W}$ of the segmentation model on image i
$q(\psi)$	Approximated posterior distribution
BNN	Bayesian Neural Network
MCD	Monte Carlo Dropout
N	Number of training examples

Understanding Calibration of Deep Neural Networks for MedIA

FS_p	Fully-Supervised with pretraining
FS_r	Fully-Supervised with random initialization
SSL_p	Rotation-based Self-Supervision with pretraining
acc	Accuracy
ACE	Adaptive Calibration Error

CKA Centered Kernel Alignment

conf Confidence

DR Diabetic Retinopathy

ECE Expected Calibration Error

MCE Maximum Calibration Error

NLL Negative Log Likelihood

OE Overconfidence Error

RMSCE Root Mean Square Calibration Error

SCE Static Calibration Error

SD Standard Deviation

Informed Label Smoothing for Improving Calibration in Medical Image Classification

$\mathcal{D}_T$ Training data

$\mathcal{D}_V$ Validation data

$\mathcal{V}_k^{acc}$ Validation set accuracy for class k

$\mathbf{u}$ Uniform prior

$\mathbf{v}^k$ Class specific prior

D_{KL} Kullback-Leibler divergence

$H(\cdot)$ Entropy function

CE Cross entropy

DCA Difference between confidence and accuracy

ECP Explicit confidence penalty

FL Focal loss

HL Hard label

ISIC International Skin Imaging Collaboration dataset

LS Label smoothing

LS+ Label smoothing plus

MDCA Multi-class difference in confidence and accuracy

MHIST Minimalist Histopathology Image Analysis dataset

R-AUC Area Under the Retention Curve

R34 ResNet-34

R50 ResNet-50

TS Temperature scaling

1 Introduction

1.1 Deep Learning in Medical Image Analysis

Deep neural networks (DNNs) have gained immense popularity in recent years due to their remarkable ability to learn complex patterns and make accurate predictions [3, 4]. The widespread adoption of deep learning is largely attributed to its successful applications in various domains, including computer vision, natural language processing, agriculture, finance, weather forecasting and healthcare. While DNNs aid in decision making, training them requires large amount of labeled data. Advances in transfer learning and fine-tuning, self-supervised learning, synthetic data generation, semi-supervised learning and active learning have made DNNs more accessible and practical for a broader range of applications when the data is scarce.

Medical image analysis (MedIA) plays a pivotal role in modern healthcare by facilitating accurate diagnosis, treatment planning, and disease monitoring. In recent years, DNNs have transformed this field by offering unprecedented accuracy and efficiency in analyzing complex medical images. However, despite these advancements, the adoption of DNNs in real world clinical settings is limited due to concerns about their trustworthiness [5, 6, 7]. Given the criticality of medical decision-making, ensuring that DNNs are transparent, robust, fair, and secure is of paramount importance. As systems become increasingly autonomous, it is crucial not only to focus on making accurate predictions but also to understand the associated risks. Assessing these risks helps determining trust in the predicted outcomes [8]. In addition, DNNs face numerous limitations in healthcare applications such as data scarcity, privacy, transparency, generalizability, robustness, and biases and fairness. To mitigate these concerns, it is essential to develop trustworthy DNNs that clinicians and patients can rely on with confidence [9, 10].

1.2 Trustworthy Deep Learning

Trustworthy deep learning is an emerging paradigm focused on evaluating and enhancing the reliability of DNNs in safety-critical applications [10, 11]. This paradigm encompasses

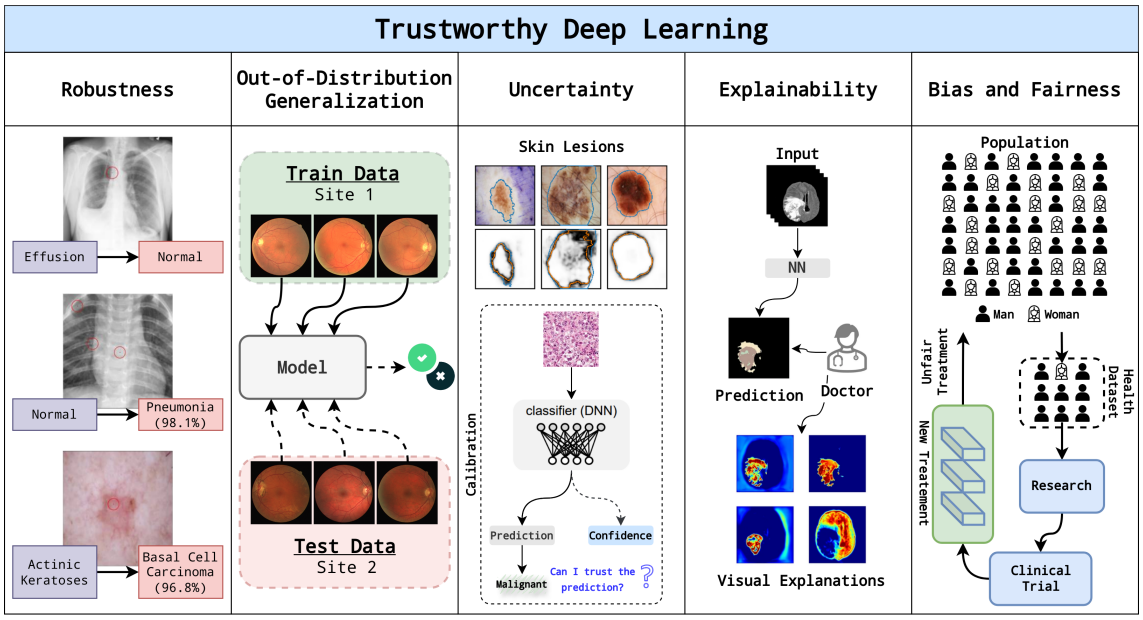


Figure 1.1: Facets of Trustworthy Deep Learning.

several subfields, a few of them are illustrated in Figure 1.1, each representing a distinct area of research that has gained considerable attention recently.

***Robustness*:** DNNs are sensitive to small changes in input data, which can lead to significant alterations in their predictions. *In Figure 1.1 (1st column) [12], examples of pixel attacks on chest X-ray and skin images are illustrated. In the first example, a single-pixel attack on the chest X-ray image altered the classification from effusion to normal. Similarly, in the second example, a multi-pixel attack involving three pixels changed the label from normal to pneumonia with 98.1% confidence. Additionally, for the skin image, adding just a single pixel to the input was enough to shift the classification from one class to another.* Ensuring the robustness is crucial for maintaining reliability and accuracy, particularly in sensitive fields like healthcare, where even minor inaccuracies can have serious consequences. Additionally, DNNs are vulnerable to adversarial attacks [13], where subtle modifications to input can result in incorrect predictions if the model is not adequately resilient. Several methods have been proposed to improve robustness when dealing with adversarial attacks [14, 15, 16], pixel attacks [17, 18] and perturbations [19]. In the context of medical imaging, inconsistencies in dataset collection can exacerbate these vulnerabilities, potentially leading to dangerous outcomes. Therefore, it is essential to develop robust models that are resistant to such attacks.

***Out-of-Distribution (OOD) Generalization*:** When models encounter data that differ significantly from their training dataset, they often face challenges in maintaining accuracy and reliability. This issue, known as distribution shift, occurs when the characteristics of the test data is not well-represented in the training data, which can result in a marked decline in performance. In MedIA, several works are dedicated to improve OOD generalization

by learning robust features across training datasets or harmonizing between domains [20, 21, 22]. *As illustrated in Figure 1.1 (2nd column), the training dataset comprises retinal images collected from Site 1, while the test dataset is drawn from a different location, Site 2. This figure demonstrates a scenario where the model is trained on data from one site but is required to make predictions on data from another site, highlighting the challenge of out-of-distribution (OOD) generalization. The disparities between the training and testing datasets involve shifts in data characteristics due to the differing environments. These variations might involve demographic factors such as the race of the patients, as well as differences in the equipment and methods used for data acquisition – ranging from the types of scanners and imaging protocols to the software used for processing the images [23]. The figure underscores the difficulty the model faces in adapting to and performing well on the test data from Site 2, given that it has been trained exclusively on data from Site 1. As demonstrated in [24], the model achieved high performance when trained and tested on the same domain (site), such as ORIGA (D1) or Drishti-GS (D2) fundus dataset, with dice scores of 0.95 and 0.87, respectively. However, when the model was trained on D1 and tested on D2, its performance dropped significantly, yielding a dice score of 0.62.* Therefore, to enhance the robustness of DNNs, it is essential to improve their generalizability when encountering dataset shifts by employing strategies such as ensembling. [25], confidence scores [26], distance based methods [27], likelihood ratios [28] and multi-task learning [29] that allow the model to adapt to varying data distributions.

***Uncertainty*:** Beyond simply generating predictions, it is vital for models to assess and communicate the uncertainty in the outputs. This capability allows practitioners to trust the DNN predictions, offering an additional layer of insight that can be critical in healthcare applications. *In Figure 1.1 (3rd column), the task is to detect the boundary of a skin lesion, where the DNN predicts for each pixel whether it is part of the boundary or not; however, it lacks an indication of the certainty or confidence of these predictions. By introducing an additional variable, uncertainty, alongside the predictions, we can better assess the trustworthiness of the results. As illustrated in the figure (second column, bottom figure of skin lesions), the model exhibits more uncertainty (darker shade) in the top right region of the predicted lesion boundary stating the model is unsure about its predicted output. However, in the other regions model exhibit less uncertainty [30]. This added layer of information allows practitioners to exercise caution when interpreting the results, focusing on these uncertain regions for further investigation and guide clinicians to take a closer look or seek additional diagnostic tools before making critical decisions. It helps reduce the risk of errors and improves the overall trustworthiness of the model.* Several works have been proposed to estimate and reduce uncertainties in DNNs [31, 32, 33, 34] and to improve calibration [35, 36, 37, 38].

***Explainability*:** The lack of transparency in DNNs decision making process is a widely recognized challenge, often leading to their characterization as "black boxes". This opacity

can significantly impede their adoption in clinical settings, where trust and accountability are paramount. *For example in Figure 1.1 (4th column), a doctor examines a brain MRI alongside two key outputs from an DNN: the segmented mask and saliency maps (visual explanation). The segmented mask highlights areas identified as potentially abnormal, such as tumors, while the saliency maps visually indicate which parts of the input the DNN focused on during its analysis. This combination allows the doctor to assess both the DNN's prediction and its reasoning, fostering trust and enabling informed decision-making regarding the patient's diagnosis and treatment.* Regulatory bodies have responded to this concern by requiring that AI systems, particularly those classified as high-risk, provide transparent and understandable explanations of their outputs [39]. This is crucial for ensuring that clinicians can confidently interpret and validate the results. Several studies have examined how to enhance the explainability of DNNs in medical imaging systems by using methods such as saliency maps [40], perturbations [41], counterfactual explanations [42], and quality improvement [43]. Demystifying the reasoning behind a model's predictions, clinicians are not only more likely to trust the outcomes but also better equipped to identify potential errors or biases.

***Bias and Fairness*:** Bias can permeate into a DNN at multiple stages, including data curation, model training, and evaluation. Addressing these biases is crucial to prevent AI systems from perpetuating or even exacerbating existing disparities, thereby ensuring that the outcomes are fair and equitable for all patient groups. *As illustrated in Figure 1.1 (5th column), suppose there is a need for COVID-19 vaccination, and a sample dataset is taken from the population, where a 90:10 male-to-female ratio is selected by chance. After conducting research and clinical trials, a new treatment is formulated. This treatment may not be fair when administered to the population, as the sampled dataset is biased toward males, which could lead to reduced effectiveness and safety for females.* When the training data does not adequately represent the diversity within the patient population, the model is likely to inherit and propagate these biases, resulting in skewed predictions that disproportionately affect certain groups. This can lead to unjust healthcare outcomes, where some patient demographics receive sub-optimal care due to the model's biased behavior. Mitigating bias involves careful consideration of data sources, rigorous validation against diverse datasets, and ongoing monitoring to ensure that the AI system delivers consistent and unbiased results. In light of that, several studies have been conducted on sources [44], reduction [45, 46, 47] and assessment of bias [48, 49], which is essential for building AI tools that not only perform well but also contribute positively to reducing health disparities and promoting fairness in medical practice.

While these aspects of *trustworthy deep learning* have been well studied, this thesis narrows its focus to the critical areas of uncertainty and calibration within medical imaging applications. Specifically, it aims to explore methods for reducing uncertainty, understanding the calibration of deep neural networks (DNNs), and mitigating the miscalibration in the

predicted confidences of these models.

1.3 Uncertainty and Calibration

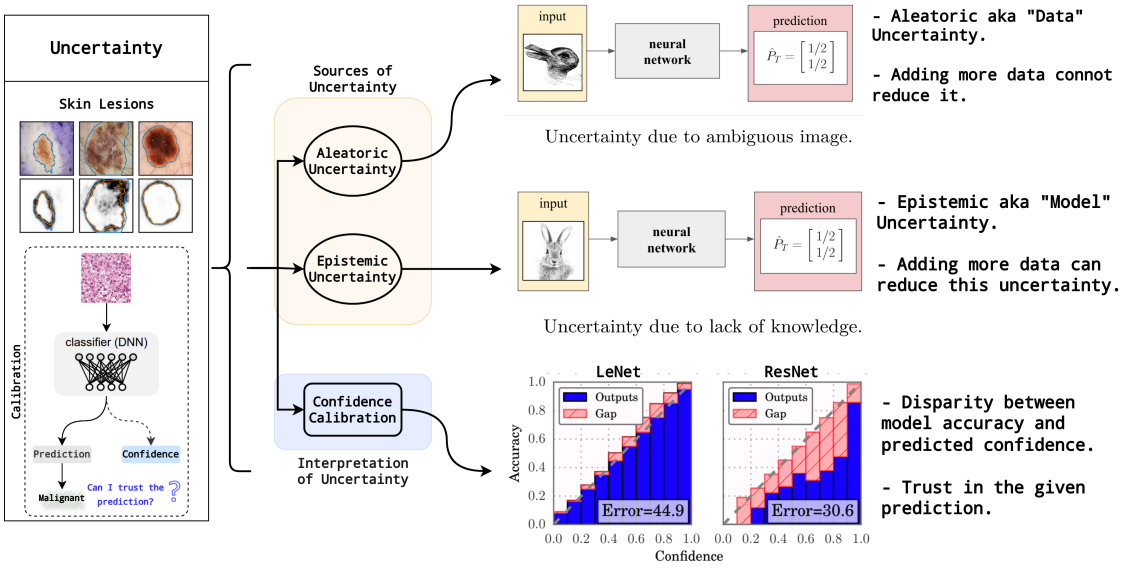


Figure 1.2: Uncertainties in Deep Neural Network.

1.3.1 Uncertainty

Uncertainty in DNNs is one of the facets of trustworthy deep learning [32]. While accuracy is a crucial metric, the reliability of DNNs is also a major concern, particularly in critical areas such as medical imaging. This has driven the development of methods to estimate and reduce prediction uncertainty [50, 51].

There are two primary sources of uncertainties: aleatoric and epistemic uncertainty [52, 53, 34]. *Aleatoric uncertainty* arises from the data itself, including inherent noise and lack of discriminative features. *For example in Figure 1.2 (top), the input is an ambiguous image, it is not clear whether it is a duck or rabbit. When presented to the neural network, it predicts both the classes with a 50% probability indicating uncertainty in the prediction due the input: as aleatoric or data uncertainty. Even if we add more samples which has the same type of ambiguity the aleatoric uncertainty will still exist.* This type of uncertainty reflects the variations in model outputs due to changes in input and is generally irreducible for a given dataset. However, improving the quality of existing features by making then more distinctive or learning the data distribution can help reduce aleatoric uncertainty [54]. *Epistemic uncertainty*, on the other hand, is related to the limitations of the model and its parameters i.e. lack of knowledge. *In Figure 1.2 (middle), the input is a clear image of a rabbit. The DNN outputs a 50% probability for both classes. This uncertainty in the prediction arises from the lack of understanding within the neural network: epistemic or*

model uncertainty. It can be caused by the limited number of rabbit instances seen during training or limited the capacity of the neural network; increasing them can reduce this uncertainty.

Aleatoric uncertainty presents a unique challenge, particularly in medical imaging. This type of uncertainty is intrinsic to the data generation process and persists regardless of the amount of data collected. In the context of medical imaging, several factors contribute to this uncertainty: *patient movement* – even slight movements during imaging can introduce blur or artifacts; *equipment variability* – different scanners or imaging protocols may produce slightly different results; *partial volume effect* – when a single voxel contains multiple tissue types, leading to ambiguous intensity values. Developing robust models that can account for and quantify this inherent uncertainty is crucial.

In recent years, significant efforts have been made for estimating *aleatoric uncertainty* by developing probabilistic models that explicitly model noise in the data [31] and incorporating uncertainty estimates into the decision-making process [34, 33, 55, 56], and *epistemic uncertainty* through strategies such as data augmentation (artificially increasing the size and diversity of training datasets) [55], bayesian inference (incorporating probabilistic methods to estimate model parameters) [56, 57, 33], and ensembling techniques (combining predictions from multiple models to improve overall accuracy) [36, 25]. Effectively estimating these uncertainties is vital to enhancing the overall quality and dependability of the results [7, 8]. Despite significant research efforts in estimating uncertainties, reducing aleatoric uncertainty has not been explored yet due to the equivocal understanding of its "irreducibility."

Objective 1: To Reduce Aleatoric Uncertainty in DNN for Medical Image Segmentation Task.

Aleatoric uncertainty arises from the inherent randomness or noise in the data generation process, simply acquiring more data cannot mitigate it. However, the thesis hypothesises that aleatoric uncertainty can be reduced if the data noise is factored in during model training. Thus, two problems need to be solved to test this hypothesis: 1) How to estimate the data noise? 2) How to incorporate the estimated noise during model training to reduce aleatoric uncertainty?

This work proposes a novel approach that interprets data uncertainty estimated from a self-supervised task as noise inherent to the data and utilizes it to reduce aleatoric uncertainty in another task related to the same dataset via data augmentation. It is validated on a medical image segmentation dataset, demonstrating a significant reduction in aleatoric uncertainty while achieving comparable or superior performance to standard augmentation techniques.

1.3.2 Calibration

DNNs often output probabilities that are used to convey confidence or certainty in their predictions. When a model predicts a particular class with a high probability, it suggests a strong conviction in that prediction. For instance, in the context of medical diagnosis, a high probability linked to a specific disease implies a greater chance of its presence based on the given input data. However, it is crucial to recognize that the trustworthiness of such interpretations hinges on the model's calibration [58, 37, 59]. *For example, when the task is to distinguish between benign and malignant samples, if a model predicts a benign label with 0.9 probability, it implies that the model is expected to be accurate 90% of the time.* Calibration ensures that the probabilities correspond to the actual likelihood of events, which is essential for accurate/trustworthy interpretation.

Accurate probability estimates from well-calibrated models improve interpretability methods like saliency maps [43], which help in understanding model decisions and associated uncertainties. Miscalibration, where models may be overconfident or underconfident, undermines these efforts, prompting a focus on improving calibration through techniques such as post-hoc adjustments [60, 37], data augmentation [61], and ensembling. While research often centers on generic datasets and metrics like Expected Calibration Error (ECE), which has its limitations like scale-dependent interpretation, lack of normalized range, arbitrary choice of number of bins, etc. [62], ongoing advancements aim to develop better methods for estimating predictive uncertainty and improving trustworthiness in medical imaging contexts [63, 64].

Understanding Calibration in Medical Image Classification

Training DNNs for medical image analysis is particularly challenging due to the limited availability of labeled datasets, which is exacerbated by the complexity and expertise required for data collection and annotation [65, 2]. To address this issue, transfer learning (TL) using pre-trained models from ImageNet is commonly employed to improve DNN performance [66, 67]. While TL has demonstrated enhanced performance and robustness in traditional computer vision tasks, similar improvements are less evident in medical imaging applications [68]. Similarly, fully-supervised and self-supervised training methods have proven effective in enhancing reliability, robustness, and uncertainty [69, 70, 71, 72], their impact on model calibration for medical imaging tasks remains under-explored.

Existing literature on DNNs primarily focuses on natural images and performance metrics, often neglecting the crucial aspect of calibration. However, it is important to note that natural image datasets typically have an abundance of data, unlike medical images, which require expert-level annotation, making them more scarce. Assuming that DNNs behave similarly across different dataset sizes is flawed, as model performance is directly tied to dataset size [68]. Additionally, the nature of the data differs significantly; in medical

imaging, even small perturbations in the input image can drastically alter the underlying issue. Consequently, the varying sizes of datasets in medical imaging make model capacity a critical factor in ensuring proper generalization. This highlights the motivation to investigate model performance and calibration across different training regimes and dataset sizes.

Objective 2: To Explore Model Calibration in Medical Imaging - Training Regimes, Dataset Sizes, Architecture Impact and Calibration Metrics.

The thesis hypothesizes that model calibration in MedIA is affected when models are pretrained on natural images. Additionally, factors such as dataset size, architecture, and training strategies (both fully-supervised and self-supervised) influence model calibration. This thesis proposes a dedicated experimental setup to investigate these hypotheses.

The study aims to bridge the existing gap by investigating the calibration of medical image classification systems under various scenarios. Specifically, it examines the effects of different training regimes, including fully-supervised learning with random initialization and pretrained models (ImageNet/RadImageNet), as well as self-supervised learning approaches such as rotation-based SSL and autoencoder-based SSL. Additionally, the study assesses the impact of dataset and architecture scales on model calibration. By exploring these factors, this research seeks to provide insights into how different training methodologies and dataset characteristics influence both performance and calibration, ultimately contributing to a better understanding of model reliability in medical imaging.

Improving Calibration for Medical Image Classification

As understanding calibration for medical imaging is crucial, improving calibration is also essential when using predicted confidences (probability) for decision-making [35]. Several methods have been proposed to improve calibration, including *post-hoc* – temperature scaling (TS) [37], weight scaling [73] and *train-time* – label smoothing [74, 75], and entropy-based regularization [76, 38], difference between Confidence and Accuracy [77, 78]. Approaches like label smoothing [75], focal loss [79], margin-based label smoothing [80] modifying label distribution to achieve calibration. While these methods aim to improve calibration, they can sometimes cause performance degradation if not carefully managed, as the regularization involved in such approaches may compromise the model's overall effectiveness.

Label smoothing (LS) is a method that involves modifying hard, one-hot encoded labels into soft labels. It is widely adopted for training DNNs due to its simplicity and performance benefits. LS not only enhances model performance but also improves calibration, making it a popular choice. This can be achieved by using a weighted sum of the cross-entropy loss between label distribution and predicted distribution, and the KL divergence loss between a uniform prior and the predicted distribution. By encouraging the predicted distribution to approach a uniform prior, the KL term helps regularize the loss, which improves calibration.

However, this traditional label smoothing approach has notable limitations: *(a) it does not consider the current calibration level of the DNN, (b) all classes are treating uniformly regardless of their calibration quality i.e. uniform prior is used.*

Objective 3: To Improve Calibration using Informed Label Smoothing

Motivated by the shortcomings of conventional label smoothing and other calibration approaches, this thesis hypothesizes that substituting the uniform prior with a class-wise informed prior will improve model calibration. To test this hypothesis, a few key questions need to be addressed: 1) How to estimate better prior compared to uniform prior considering the current calibration of the DNN? 2) How to ensure the estimated prior considers class-wise calibration? 3) How to make prior estimates unbiased i.e. should not be dependent on training data?

This work introduces a novel strategy called Label Smoothing Plus (LS+), which utilizes class-specific priors estimated from validation set accuracies. The estimated priors are used during DNN training replacing the uniform prior which benefits calibration. Its effectiveness is validated across multiple medical imaging datasets and DNN architectures, focusing on various performance and calibration metrics for classification tasks. Results showed notable reduction in calibration error with nominal improvement in performance compared to state-of-the-art approaches.

1.4 Thesis Organization

- Chapter 1: This chapter provides an overview of the trustworthy deep learning paradigm and its subfields in medical image analysis, emphasizing the need for reliable AI systems in healthcare. It outlines the central challenges of uncertainty and calibration, along with the thesis's goals and contributions.

- Chapter 2: A detailed review of the relevant literature on uncertainty quantification, calibration techniques, and methods to improve model trustworthiness in medical imaging.

- Chapter 3: This chapter investigates strategies for mitigating aleatoric uncertainty in medical imaging, presenting method to reduce the impact of data variability and enhance model robustness across diverse medical imaging tasks.

- Chapter 4: This chapter explores the various factors influencing calibration in DNNs used for medical image classification.

- Chapter 5: This chapter introduces an informed label smoothing technique aimed at improving confidence calibration by leveraging validation dataset accuracies to estimate prior distributions for guiding the training of DNN-based medical imaging systems.

- Chapter 6: The final chapter summarizes the thesis's contributions, highlighting the findings and their implications for trustworthy deep learning in medical imaging. It also outlines future research directions to further improve model trustworthiness and calibration in healthcare applications.

2 Literature Review

2.1 Trustworthy Deep Learning

Many AI systems today are susceptible to subtle attacks, exhibit bias against underrepresented groups, and fail to protect user privacy, undermining user trust and experience. A systematic approach, involving collaboration across disciplines and throughout the AI system lifecycle, is needed to improve trustworthiness. In this section, we review studies on trustworthy deep learning that address critical factors such as robustness, OOD generalization, explainability, transparency and fairness. The robustness of DNNs refers to their sensitivity to variations in inputs or model parameters. Several approaches have been proposed to enhance DNN robustness, such as making convolutional neural networks (CNNs) more resilient to adversarial attacks by reducing noise in medical images using a simple approach which involves applying Gaussian filtering to input data during preprocessing [14]. Noisy input images can lead to noisy feature representations, which are further distorted by adversarial perturbations, significantly degrading model accuracy. In contrast, preprocessing reduces noise in the feature representations, mitigating the impact of adversarial perturbations and improving accuracy. Moreover, this work introduces a robust CNN architecture that integrates sparsity denoising operators into each layer, effectively reducing noise in the features learned during training. These operators address both inter-sample noise (i.e., noise within a batch) and intra-sample noise (i.e., noise within individual inputs). Training this framework alongside adversarial training reduced noise in the learned feature representations while enhancing robustness and accuracy against both white-box and black-box attacks. White-box attacks involve scenarios where the attacker has complete access to the model, including its architecture, parameters, and training data, enabling them to create highly targeted adversarial examples [81, 82]. On the other hand, black-box attacks are more challenging as they are conducted without any internal knowledge of the model, forcing the attacker to rely solely on input-output interactions to generate adversarial inputs [83, 84]. For malignant lung nodule prediction, a simple defense strategy was proposed using ensembles of DNNs with adversarial images incorporated during training [15]. This approach improved both classification accuracy and resilience against Fast Gradient Sign Method (FGSM) [81] and 1-Pixel [17, 18] adversarial attacks.

In MedIA, the widespread use of DNNs across various tasks poses significant security risks, as these models are particularly vulnerable to adversarial attacks. A study was proposed to investigate the susceptibility of medical imaging systems to such attacks [16]. It explored several scenarios across three medical domains (ophthalmology, radiology, and pathology) including: (a) the effect of different DNN weight initializations on the transferability of adversarial attacks from a surrogate model to a target model, (b) the transferability of adversarial examples when the surrogate model is trained on a dataset of a different size compared to the target model, and (c) the transferability of adversarial examples between different architectures. The results showed that pretraining (e.g., on ImageNet) increases the transferability of adversarial attacks, making models more vulnerable to attacks by pretrained surrogate models. Additionally, differences in data size and architecture reduce the success of these attacks. Based on these findings, to minimize the transferability of attacks in MedIA, it is recommended to avoid using standard architectures and publicly available datasets. Furthermore, the disclosure of customized architectures and other model specifications should be restricted.

The traditional machine learning paradigm assumes that training and testing datasets are Independent and Identically Distributed (i.i.d.) samples from the same underlying distribution. However, this assumption often breaks down in real-world applications, particularly in healthcare, where data can come from diverse sources or involve unseen samples during testing [23]. To address this issue, the Out-of-Distribution (OOD) Generalization paradigm has emerged, tackling dataset drift by exploring various research areas such as Domain Adaptation, Domain Generalization, Federated Learning, and OOD Detection [85]. To tackle the challenge of generalizing DNNs to unseen (target) samples, a standard transfer learning approach can be employed by fine-tuning the model on a small annotated dataset from the target domain. While this method can be effective, acquiring additional annotated samples from the target domain is often difficult and costly. To address this issue, unsupervised domain adaptation methods [86, 87] have been proposed, which do not require extra annotated samples. A novel *unsupervised domain adaptation framework* was proposed that uses adversarial learning for segmentation tasks [86]. It learned the mapping from source domain inputs to labels using a ConvNet Segmenter architecture consisting of a CNN with dilated residual blocks. To generalize well on the target domain, unlabelled data was leveraged, as annotating could be time-consuming and expensive. The Domain Adaptation Module (DAM) mapped target domain (e.g., CT) inputs to the source domain (e.g., MRI) feature space, with the lower layers of the ConvNet Segmenter replaced by DAM during training. At inference, DAM layers could substitute the ConvNet's lower layers, and the feature output could be mapped to label space using the established high-level layers. Additionally, a Domain Critic Model (DCM) aligned multiple feature map levels between domains by jointly optimizing DAM and DCM via adversarial loss, which improved cross-modality domain adaptation performance. However, this method still necessitate some information about the target domain, such as unlabeled data or domain-specific knowledge, to establish a robust mapping from the source to the target

domain. Therefore, approaches that don't require any information from the unseen domain is ideal. In MedIA, test images could come from various sources like different hospitals, scanners or race. This discrepancy in the test distribution can affect the DNN to overfit on the training dataset reducing the generalizability. The *Domain-oriented Feature Embedding (DoFE)* [88] framework improves generalization on unseen test data in segmentation tasks by leveraging multi-source domain data without requiring target samples. It extracts more discriminative semantic features from these diverse domains and uses a domain knowledge pool containing the feature embeddings of each source. By combining the semantic features and domain knowledge, an aggregate feature representation is obtained. Additionally, the original image features are then augmented with these domain-oriented aggregated features, where aggregation is based on the similarity between the input image and the multi-source domain images. The proposed approach showed significant improvement in the dice score on OOD fundus datasets.

Explainability is important for building trust in DNNs, which often work like black boxes. The purpose of explainable AI (XAI) is to show how these models make decisions, making them more understandable and trustworthy. Several categorizations are provided for explainable AI techniques [89, 90] and this section distinguish XAI techniques based on three criteria: model-based versus post hoc, model-specific versus model agnostic, and global versus local adapted from [89]. These techniques generate saliency maps that indicate which areas of an image contribute to a given prediction. [91] is a model specific post-hoc technique that produces a map by calculating how the output changes with respect to the input pixels, using partial derivatives of the output relative to the input image. Furthermore, [92] employs a deconvnet to interpret feature activity in the intermediate layers of DNNs. *Class Activation Maps (CAM)* [93] visualize which regions of an image contribute most to predictions made by convolutional neural networks (CNNs). CAM achieves this by applying global average pooling to the feature maps from the last convolutional layer, generating a weighted sum based on the model's learned class-specific weights, and producing a heatmap that highlights the most relevant areas for a specific class. Subsequently, *Grad-CAM* [94] and *Grad-CAM++* [95] were introduced as an improvement over CAM and do not require modification to the architecture or retraining. Some perturbation based techniques have been proposed to observe the importance of certain areas of the images by perturbing the input images like locally interpretable explanations and occlusion sensitivity. *Local Interpretable Model-Agnostic Explanations (LIME)* [96, 97] improves the interpretability of deep learning models by providing localized, human-understandable explanations. For non-linear decision boundaries, it is challenging for humans to comprehend how a model generates predictions, as summarizing the entire decision boundary into a single explanation is often impractical. LIME addresses this by focusing on the local area around an individual prediction, where it is possible to generate a simple, interpretable explanation specific to that region. This approach bypasses the need to understand the model's entire decision boundary and instead provides insights into the reasoning behind a specific prediction. Within the local region, interpretable

models like linear regression or decision trees can be applied to understand the prediction process. The *occlusion sensitivity* approach [92] systematically occludes parts of an image and observes their effect on the model's output, revealing how the classification process depends on specific visual features. However, such perturbations, like replacing a region of the input image with a constant value, may not be ideal since medical images are inherently noisy and blurry. To address this, [41] proposed using a Variational Autoencoder (VAE) to replace pathological regions with healthy tissue. This approach demonstrated improved localization of pathological areas compared to earlier methods. Building on a similar concept, an image in-painting approach [98] was proposed, demonstrating notable improvements over earlier methods in identifying abnormalities in breast mammography and tuberculosis in chest X-rays. Additional XAI methods have been developed that incorporate textual explanations through image captioning, as well as image captioning combined with visual explanations, across various modalities such as histology [99, 100], X-ray [101, 102], and CT [103].

Bias in deep learning refers to a model's tendency to produce skewed or unfair outcomes due to prejudiced data or flawed assumptions during training. It has been reported in various medical imaging domain like skin lesions [104], chest x-ray [105, 106] and brain imaging [107]. Fairness in AI involves methods for evaluating and mitigating such biases [108]. Several approaches have been proposed to mitigate bias at different stages of DNN development: (a) during the preparation of training data (preprocessing), (b) during model training (inprocessing), and (c) in the model's predictions (postprocessing) [45]. In the preprocessing stage, data imbalance is a major contributor to bias, which can be mitigated by adding more data from underperforming classes (or groups) [109], or by reweighting/increasing the importance of underrepresented classes [45, 46]. *Representation Neutralization for Fairness (RNF)* [110] is proposed to improve fairness even when input representations are biased by debiasing the classification head of the DNN. In this method, a biased teacher DNN is first trained using two inputs from the same class, producing corresponding learned representations and predictions. The encoder is then frozen, and only the weights of the classification head are updated by minimizing the mean square error between the output of the classification head on the averaged representations and the averaged predictions, a process referred to as representation neutralization. This reduces undesirable correlations between fairness-sensitive information in the representation and the class labels. Furthermore, the *fair meta-learning segmentation* strategy [45] is proposed to improve segmentation fairness through multi-task learning. Here, the DNNs are trained simultaneously for the segmentation task in cardiac MR images and the classification of protected attribute(s), with both networks optimized jointly. The additional classification network ensures that learning from a dominant group does not negatively impact the learning of other groups.

2.2 Uncertainty Quantification

The literature on uncertainty estimation in DNNs presents a broad range of techniques aimed at improving model reliability and decision-making. Researchers have developed both Bayesian and non-Bayesian methods to address and mitigate aleatoric and epistemic uncertainty across various domains. This section reviews key contributions from these approaches, tracing the evolution of strategies designed to estimate and reduce uncertainties, ultimately enhancing model performance and trustworthiness.

A substantial amount of research on uncertainty estimation in neural networks is based on the *Bayesian principle* [111]. Bayesian Neural Networks (BNNs) are a type of neural network that incorporate Bayesian inference to model uncertainty in predictions by treating the network's weights as probability distributions rather than fixed values. Instead of learning a single set of weights, BNNs learn a posterior distribution over the weights, combining prior beliefs with observed data. This provides a natural framework to estimate uncertainty in predictions, offering more robust and interpretable models. As exact Bayesian inference is intractable for DNNs due to the large number of parameters, various approximation methods have been developed, such as Laplacian approximation [112], Markov Chain Monte Carlo (MCMC) methods [113], and variational Bayesian methods [114, 56]. The quality of predictive uncertainty estimated by BNNs depends on the accuracy of the approximation, which is influenced by computational constraints and the choice of the prior distribution. An incorrectly chosen prior can lead to unreliable predictive uncertainties. One of the popular Bayesian approaches, *Bayes by Backprop* [56], offers a practical solution for training BNNs to overcome intractability and estimate predictive uncertainties. It employs variational inference, which approximates the true posterior distribution of the weights with a simpler, tractable distribution. This is achieved through the reparameterization trick, allowing gradients to be back-propagated through stochastic nodes in the network. As a result, this approach demonstrated performance comparable to other methods and provided better uncertainty estimates when test examples came from unseen data distribution.

Although Bayesian methods offer a theoretically sound framework for estimating uncertainty, they carry a high computational cost due to the need to estimate the full posterior distribution over the model's parameters. Monte Carlo Dropout (MCD) [57] proposed an alternative to the Bayesian approach where variational inference is approximated with dropout regularization. In practice, dropout is applied during both training and testing [115]. The test samples are passed through the neural network multiple times with dropout enabled, generating different predictions each time, as different sets of neurons are activated in each iteration. The variance across these predictions reflects the uncertainty arising from the different neural network configurations created by dropout. When assessing dropout-based uncertainty, it showed significant improvements in predictive log-likelihood compared to popular variational inference techniques, as well as better uncertainty estimates for classification tasks when the model predictions were incorrect.

Bayesian deep learning approaches define a probability distribution over model weights and model outputs for modeling epistemic and aleatoric uncertainty respectively [116].

One study explored the understanding of different types of uncertainties, namely aleatoric and epistemic, both individually and in combination [34]. It proposed a unified Bayesian deep learning framework that enables learning mappings from input data to aleatoric uncertainty, along with estimating epistemic uncertainty. Aleatoric uncertainty can be categorized as either homoscedastic or heteroscedastic. Homoscedastic uncertainty refers to a constant level of noise that remains uniform across the entire dataset, while heteroscedastic uncertainty represents instance-specific noise that varies across individual data points. Heteroscedastic uncertainty can be modeled using a probabilistic neural network known as a heteroscedastic neural network (HNN) [117], which predicts the parameters of the output distribution. For example, in a regression task where the output is modeled as a Gaussian distribution, the HNN predicts the distribution parameters such as the mean and variance. To capture epistemic uncertainty, the heteroscedastic neural network is used with Monte Carlo (MC) dropout [57]. It allows the network to generate multiple predictions, with the variance among them representing uncertainty. For classification tasks, heteroscedastic uncertainty was modeled by modifying a standard classification model to marginalize over the intermediate heteroscedastic regression uncertainty in the logit space. It was observed that modeling both uncertainties together improved performance in tasks such as semantic segmentation and monocular depth regression across various datasets. The proposed method also provided better-quality uncertainty estimates in scenarios like monocular depth regression, where high aleatoric uncertainty was noted for distant objects and occlusion boundaries in images. Additionally, the study showed that aleatoric uncertainty, being inherent to the data, cannot be reduced by adding more data, whereas epistemic uncertainty, related to model knowledge, can be minimized with additional data. Notably, epistemic uncertainty increased for out-of-distribution samples (situations differing from the training set) while aleatoric uncertainty remained unchanged in such cases. Modeling both uncertainties together was observed to enhance performance in tasks such as semantic segmentation and monocular depth regression across various datasets. This approach also produced higher-quality uncertainty estimates, particularly in cases like monocular depth regression, where high aleatoric uncertainty was noted around distant objects and occlusion boundaries.

The *ensemble* method [25] is another widely used non-Bayesian approach for estimating predictive uncertainty, valued for its simplicity and scalability. It improves upon BNNs, which require extensive modifications to the training process and are computationally expensive. The method proposes to train an ensemble of heteroscedastic neural networks. Epistemic uncertainty is captured by the variance between predictions in the ensemble, while aleatoric (heteroscedastic) uncertainty is represented by the average of the variances from each network in the ensemble. Experiments on regression and classification tasks have shown that ensembles provide better uncertainty estimates than BNNs and is more robust to domain shift, particularly in detecting out-of-distribution samples. A deeper understanding of the advantages of ensembles over BNNs becomes evident when analyzing the function space [118]. Ensemble methods have been shown to often outperform popular

variational Bayesian approaches by exploring diverse modes within the function space, rather than being limited to a single mode. By averaging the predictions of ensembles, these methods navigate multiple minima and frequently converge on flatter, more stable solutions (optimized functions). This characteristic enhances generalization and yields more reliable uncertainty estimates. In contrast, while BNNs have a strong theoretical foundation, they face challenges in effectively exploring the function space. This is primarily due to the difficulties associated with accurately approximating the posterior distribution of the weights, which can lead to sharper minima and less dependable uncertainty assessments.

The *Bayesian decision-theoretic DNN framework* [33] was introduced in MedIA, offering a principled approach to uncertainty estimation along with improved calibration for image segmentation tasks. In this framework, the label-probability distribution at each voxel is modeled using a Dirichlet distribution, with its concentration parameters estimated during training. The learned distribution is then used to generate new label-probability vectors, which are subsequently converted into discrete labels to produce segmentation predictions for each voxel. Per-voxel uncertainty is quantified by computing the square root of the trace of the covariance matrix of the Dirichlet distribution. These uncertainty estimates are analytically derived at test time, eliminating the need for approximate or computationally expensive algorithms. The Bayesian DNN framework demonstrated superior segmentation performance across brain, chest, and cell datasets, while significantly enhancing various calibration metrics. Another study showed that incorporating uncertainty into the DNN training improved anomaly detection in chest X-ray images [119]. It employs a probabilistic autoencoder that generates both the reconstructed image and a pixel-wise uncertainty mask. During training, the mean squared error is normalized by the predicted uncertainty map, which is referred to as the abnormality score. The underlying idea is that in a normal image, areas or pixels exhibiting higher reconstruction errors are typically associated with greater uncertainties, which generally results in a lower overall abnormality score. Conversely, when an image contains an anomaly, the presence of relatively high reconstruction errors in regions of low uncertainty can contribute to a higher abnormality score. This distinction helps in identifying anomalies effectively based on the characteristics of reconstruction errors and associated uncertainties.

Furthermore, the *DR/GRADUATE* [120] approach was proposed to simultaneously provide explanations for predictions (DR severity levels) along with the associated uncertainty. It uses a custom DNN that models uncertainty alongside predictions, similar to a probabilistic neural network. Additionally, it generates pixel-wise explainability maps for each DR grade, where every lesion in the predicted lesion map is translated into the explainability map. This approach delivers state-of-the-art performance while providing uncertainty estimates, predictions and explanations, all at the same time. The explainability maps highlight key regions in the images, enhancing trust and helping identify cases that would benefit from a second evaluation. Another work proposed to leverage predicted uncertainty to reduce false positives in liver lesion detection [121]. In the first phase, a custom DNN inspired by U-Net [122] and Bayesian SegNet [123] is used to generate predictions along with an uncertainty

map. In the second phase, 3D patches are extracted from the uncertainty map at detected lesion sites for each patient. The analysis revealed that false positive predictions tend to have a smaller volume compared to true positives. Based on this observation, the maximum diameter of the detected lesion is used as a feature, alongside aggregated uncertainty from the patches. Additionally, radiomics features [124] are also incorporated and fed into an SVM classifier to differentiate between true and false positive lesions. Furthermore, the model predicts higher uncertainty for incorrect predictions, validating the reliability of its uncertainty estimates.

For brain lesion detection, quantile regression loss was introduced to estimate uncertainty in variational autoencoders (VAE) [125]. Traditional VAEs often suffer from shrinkage or underestimation of variance because, when predictions are nearly perfect (i.e., with zero reconstruction error), maximizing the log-likelihood pushes the estimated variance toward zero. This can lead to overfitting, where zero variance fails to reflect the true model performance on test data. To address this, *Quantile-Regression VAE (QR-VAE)* was proposed, predicting multiple quantiles of the output distribution at each pixel, rather than estimating only the mean and variance. The reconstruction loss leverages these predicted quantiles and minimizes the pinball loss for each quantile. This approach is straightforward to implement and outperforms other VAE methods in brain lesion detection tasks. While previous work has focused on estimating and incorporating uncertainty in segmentation and classification tasks, bounding-box-based detection is under-explored. A *single-scale multi-level pyramid CNN* [126] was introduced that incorporates bounding-box level uncertainties using both Monte Carlo (MC) samples and predictive variance. Instance-level uncertainty is measured by the variance of the Monte Carlo samples for a bounding box, while predictive variance is estimated similar to a heteroscedastic neural network. Integrating these uncertainties into DNN training improved performance and reduced false positives in lung nodule detection tasks (LUNA16 dataset).

Test-time Augmentation (TTA) [55, 127] is an alternate, non-probabilistic approach for estimating aleatoric uncertainty that is less expensive compared to Bayesian methods. During testing, multiple variations of a single test instance are generated using augmentations such as geometric and color transformations, which include random cropping and resizing, adjustments to brightness, hue, saturation, and contrast, as well as random horizontal and vertical flips and rotations. For each augmented input, the model generates a corresponding output, and the variance among these predictions serves as a measure of aleatoric uncertainty. This process helps assess how much the model's output varies with the test input. By exploring different perspectives of the test data, this approach effectively captures aleatoric uncertainty. Furthermore, this method is easy to implement, requiring no modifications to the model and no additional training data. Results from segmentation tasks on 2D and 3D Magnetic Resonance Imaging (MRI) of fetal brains and brain tumors indicate that TTA provides superior uncertainty estimates compared to Monte Carlo dropout [57] and reduces overconfidence in incorrect predictions. Additionally, TTA demonstrated improvement in segmentation accuracy over both single-prediction and dropout-based multiple prediction

methods.

Other methods that offer a distinct perspective on uncertainty quantification include Conformal Prediction (CP) and Evidential Deep Learning (EDL). CP [128] provides statistically rigorous guarantees for estimating prediction intervals in regression tasks and prediction sets in classification tasks. Both the prediction intervals and sets are designed to encompass the true value with a specified confidence level. For a classification task, instead of a single class label as the output, the model generates a set of possible predictions for a given input. The size of this prediction set reflects the model's uncertainty; a larger set indicates greater difficulty in classifying the input sample, while a smaller set suggests more confidence in the classification. CP guarantees that the true label will be included in the predicted set of classes, providing a coverage assurance that is particularly valuable when conveying the model's confidence alongside its predictions. Recently, it has gained popularity in the machine learning community for its ability to deliver valid and interpretable uncertainty estimates without relying on specific assumptions about the underlying data distribution or predictive model. Consequently, this framework is advantageous across various fields, including medical diagnosis, anomaly detection, financial risk assessment, and any domain where reliable uncertainty quantification is critical. Moreover, CP has shown promise when integrated into human-in-the-loop systems, enhancing decision-making processes by providing clearer insights into model predictions.

The *Deep Evidential Regression* [129] introduces an innovative framework that integrates evidential reasoning into DNNs to estimate continuous targets while quantifying uncertainty. The method allows the neural network to predict hyperparameters of a Normal Inverse-Gamma (NIG) distribution, which captures both the mean and variance of the target variable. During training, the model learns to align its predicted evidence with the true outputs through a specially designed loss function that penalizes misalignment, promoting well-calibrated uncertainty estimates. At inference, the network outputs parameters of the evidential distribution, enabling it to generate predictions along with uncertainty estimates efficiently, without the need for sampling. This approach has demonstrated robustness across various tasks like depth estimation, OOD detection and adversarial attacks.

2.3 Calibration

Calibration in DNNs can be categorized into two main types: *post-hoc* and *train-time methods*. *Post-hoc* calibration refers to adjusting a model's predicted probabilities to better align them with actual outcomes after the model has been trained. These methods are computationally efficient since they do not require retraining the model. On the other hand, *train-time* calibration involves incorporating strategies during the training phase, such as modifying the loss function to encourage better-calibrated outputs directly during model training. This approach can produce inherently better-calibrated models but may come with additional training complexity.

Post-hoc methods: *Platt Scaling* [60] is a post-hoc calibration method (a parametric

model) that utilizes a separate calibration or validation set to learn parameters that best adjust the model's output probabilities. For a binary classification task, Platt scaling applies logistic regression to map the outputs to calibrated probabilities by minimizing the negative log-loss on the calibration set. Alternatively, *Isotonic Regression* [130] is a statistical technique that fits a non-decreasing (or non-increasing) function to a sequence of observations. It is a variant of linear regression that allows for a piece-wise linear fit, breaking the problem into linear segments and performing linear interpolation between them. It provides a flexible, non-parametric approach to modeling monotonic relationships in data. It is often preferred over Platt scaling when dealing with larger datasets, non-sigmoid calibration curves, and situations where the relationship between predicted scores and probabilities is complex. Another popular method, *Temperature Scaling (TS)* [37] is an extension of Platt scaling, which is simple and effective for improving the calibration of the model. Modern DNNs often suffer from miscalibration, particularly overconfidence, which can stem from factors like large model capacity, batch normalization, and use of less weight decay. TS adjusts the model's output logits using a single scalar parameter, called as temperature which is learned by minimizing the negative log-likelihood on a calibration dataset. This adjustment helps refine the model's probability estimates by flattening or sharpening the predicted probability distribution, leading to more calibrated confidence scores compared to Platt scaling and Isotonic regression.

When using a confidence-interval-based binning method, where each validation sample is assigned to a bin based on its predicted confidence score, larger confidence intervals tend to accumulate more samples, leaving other bins with very few. As a result, using a single temperature or separate temperatures for low-confidence bins can be problematic, as these are often derived from a very limited number of validation samples, making them less reliable. *Bin-wise Temperature Scaling (BTS)* [131] extends TS by dividing the validation samples into multiple bins and computing separate temperatures for each bin using a sample-based binning method, which ensures an equal number of validation samples in each bin, except for the high-confidence bins. While this approach resolves the issues of confidence-interval-based binning, it expands the range within each bin. This can lead to large differences between the minimum and maximum confidence values within the same bin. To further improve BTS, Augmentation-based Bin-wise Temperature Scaling (ABTS) was introduced, applying data augmentation to bins with fewer validation samples. This augmentation helps identify more stable temperatures for low-confidence bins, enhancing the overall performance of BTS.

Varying the temperature in TS induces a continuous path from the original class distribution to a uniform distribution. *Confidence-based Weights Scaling (CWS)* [73] is a weight scaling calibration method that computes a convex combination of the network's output class distribution and a uniform distribution. It uses the Adaptive Expected Calibration Error (adaECE) to learn the optimal calibration procedure through weight scaling. Both TS and CWS maintain the predicted class order and do not affect accuracy. Additionally, CWS preserves consistency by maintaining the order of clinical decision-making across different

patients based on the predicted confidence i.e., a more confident patient will remain more confident after applying CWS, which is not always the case with TS. This method also demonstrated improved calibration compared to state-of-the-art techniques.

Train-time methods: *Label Smoothing (LS)* [75] is a train-time regularization technique to improve generalization and prevent overconfidence in DNN predictions. Instead of assigning hard labels (0 or 1) for classification tasks, LS modifies the labels to create a softer target distribution. For instance, it replaces the hard label of 1 for the correct class with a value slightly less than 1, while distributing the remaining probability mass among the incorrect classes. It is initially proposed as a technique to improve generalization, but various studies have demonstrated its benefits in improving the calibration of the model [74, 132, 133, 134]. The study on understanding why label smoothing improves generalization and calibration made a few key observations [74]: (a) representations learned with label smoothing differ from those learned without it, (b) label smoothing implicitly calibrates the model, making predicted confidences more aligned with actual accuracy, (c) in Knowledge Distillation framework, using predictions from the teacher trained using LS to train the student network, do not provide enough information thereby reducing its performance. Subsequently, *Explicit Confidence Penalty (ECP)* [76] was proposed to regularize models when their predictive output distributions exhibit low entropy. Overconfident models tend to produce sharp distributions, placing nearly all the probability on a single class, which often leads to overfitting. To counter this, the method adds an entropy-based regularization to the cross-entropy loss, penalizing overly peaked distributions. It has been observed that this regularization improves model generalization across various tasks, including image classification, language modeling, machine translation, and speech recognition. It is important to highlight that calibration methods like TS and ECP improve calibration by reducing overconfidence, but this reduction can compromise valid high-confidence predictions. To address this, *Maximum Mean Calibration Error (MMCE)* [135] was introduced as a more principled approach, minimizing calibration error during training. MMCE uses an auxiliary loss term computed in a reproducing kernel in a Hilbert space (RKHS) [136], which can be efficiently optimized alongside the negative log-likelihood loss without requiring extensive hyperparameter tuning. This loss serves as a surrogate for calibration error, reaching zero only when the model is perfectly calibrated.

Focal Loss (FL) [137], on the other hand, was introduced to improve model performance in class-imbalanced scenarios. It modifies the standard cross-entropy loss by down-weighting the loss contribution of correctly classified samples, thereby focusing more on hard-to-classify instances. Additional study [79] showed that FL can be beneficial for improving calibration. For a classification task, cross-entropy aims to minimize the the Kullback-Leibler divergence (D_{KL}) between the predicted softmax distribution and the target distribution, however focal loss minimizes a regularized version of D_{KL}. It is shown that the miscalibration and negative log-likelihood overfitting are linked to each other and are affected by the peaky distribution of misclassified samples. Furthermore, this study provides both theoretical and empirical evidence to support the improvement in calibration achieved through FL.

During standard training, cross-entropy loss acts as a surrogate for minimizing the D_{KL} between the target and predicted distributions, effectively providing an upper bound on D_{KL}. In contrast, FL can be seen as an upper bound on a regularized KL divergence, where a negative entropy term is added to D_{KL}. This regularization term is controlled by a hyperparameter that encourages higher entropy in the predictive distribution, reducing overconfidence in DNN outputs. Moreover, this study also provides a principled approach for selecting the hyperparameter. The results demonstrate that models trained with FL not only improve calibration under i.i.d. assumptions but also in out-of-distribution scenarios.

Previous state-of-the-art calibration losses, such as LS and FL/ECP, consist of two components: a classification loss term and a second term, either D_{KL} or the entropy of the predicted distribution. It has been shown that these loss functions are closely related. Each of these losses attempts to minimize the distance between class logits, driving the predictions toward a uniform distribution. *Margin-based Label Smoothing (MbLS)* [80, 132] introduces a constrained-optimization approach that unifies calibration losses proposed earlier and provides a more informed way of regularizing the model. It penalizes logits where the distance between the logits of classes exceeds a specified margin, guiding the model toward a more informative objective (non-uniform distribution regularization). When the margin is set to zero, this approach is equivalent to LS (which uses uniform distribution to add regularization). Extensive evaluations have demonstrated that MbLS improves calibration compared to other state-of-the-art methods without compromising discriminative performance across various medical image segmentation tasks and natural image datasets.

Calibration metrics like Expected Calibration Error (ECE) [138] and Static Calibration Error (SCE) [63] are used to assess the miscalibration of DNN predictions, and leveraging this information during training can provide significant benefits. In a similar vein, the *Difference between Confidence and Accuracy (DCA)* [77] has been introduced as an auxiliary loss function aimed at reducing overconfidence in DNNs, particularly for medical image classification tasks. Inspired by the ECE, DCA penalizes the model when the cross-entropy loss decreases without a corresponding increase in accuracy. It measures the model's calibration by calculating the difference between the average predicted confidence and the actual accuracy. A larger difference results in a greater penalty, and a smaller difference incurs less. This approach is simple and can be easily integrated into any classification task. Evaluation on various medical image classification datasets using multiple DNN architectures showed significant improvement in calibration while maintaining the overall classification accuracy. Despite its benefits in improving calibration, this approach has limitations, such as its inability to capture class-wise calibration and its focus solely on the top predicted class. Consequently, it fails to address miscalibration in other classes. To address this issue, the *Multi-class Difference in Confidence and Accuracy (MDCA)* method has been proposed [78]. MDCA is a class-wise calibration technique inspired by SCE, which quantifies miscalibration across different classes. Similar to DCA, it serves as an auxiliary loss function that measures the misalignment between the confidence scores and the frequency of each class in the training dataset. The primary objective of

this approach is to reduce this misalignment, thereby enhancing the model's calibration. MDCA has demonstrated improvements in calibration across various image classification and segmentation tasks, and out-of-distribution scenarios. However, while it enhances calibration, it may lead to a decrease in overall performance. Although these approaches use the ECE/SCE calibration metrics during training to capture the current calibration of the model, ECE is not differentiable. *Differentiable Expected Calibration Error (DECE)* [139] addresses the non-differentiability of ECE computation by proposing differentiable approximations for accuracy and binning. While ECE requires accuracy, confidence, and binning for computation, only confidence is inherently differentiable. DECE overcomes this limitation by introducing soft binning to avoid the non-differentiability associated with hard binning and by exploiting meta-learning techniques to optimize the loss. However, the reliance on approximations and meta-learning in DECE adds complexity and may limit its broader applicability compared to direct optimization of the loss function.

For long-tailed datasets, DNN tends to produce overconfident predictions for high-frequency classes. Calibration methods like TS and LS use a single scalar and a smoothing factor for all classes, respectively. However, this may not be optimal for improving calibration, as different classes contribute unequally to DNN training due to variations in sample size. *Class-Distribution-Aware Calibration* [133] introduced Class-Distribution-Aware TS (CDA-TS) and Class-Distribution-Aware LS (CDA-LS), where the vectors are computed based on the frequency of samples in each class. In CDA-TS, the corresponding temperature is used to scale the logits for each class to compensate for the over-confidence. CDA-LS selects the appropriate smoothing factor value for each class and flattens the hard labels according to their corresponding class distribution. Experiments demonstrated that incorporating class distribution knowledge enhanced both calibration and performance in class-imbalanced scenarios. Beyond improving calibration, it is also important to understand when and how model calibration can enhance the reliability of DNNs. This is investigated by focusing on the following points under various degrees of class imbalance [140]: (a) selecting calibration methods for enhanced performance (b) determining an optimal "calibration-guided" threshold for different levels of data imbalance and (c) evaluating performance improvements when using thresholds derived from calibrated probabilities compared to the default 0.5 threshold.

While prior works like LS, ECP, and FL penalize models for predicting low-entropy distributions on every instance, *Maximum Entropy on Erroneous Predictions (MEEP)* [38] introduced a more selective approach. In segmentation tasks, MEEP penalizes only the misclassified, or harder, pixels, while high-confidence correctly classified pixels remain unaffected (not considered during training). The overall loss consists of two terms: (a) segmentation loss, which ensures high-quality predictions, and (b) the entropy of misclassified pixels, which helps to identify uncertain regions.

In the literature, ensembles have been recognized as a state-of-the-art method for enhancing calibration and uncertainty estimation due to the model diversity that it introduces. However, this technique often incurs significant computational costs. *Multi-Head Multi-Loss*

Model Calibration [141] presents a streamlined ensembling strategy that improves model calibration without the need to train multiple DNNs. Instead of relying on a single linear classifier, it proposes a multi-head architecture where each head is trained using distinct weighted cross-entropy loss function. This multi-head setup induces ensemble diversity, while the varying loss weights prevent the heads from making similar predictions. By averaging the outputs from these multiple heads, the approach achieves better calibration while maintaining accuracy. The method demonstrates the lowest ECE compared to deep ensembles and other calibration techniques, while also achieving comparable improvements in accuracy.

Calibration metrics: A calibration metric is a quantitative measure used to evaluate how well a model's predicted probabilities align with the actual outcomes, identifying miscalibration. In the literature, numerous calibration metrics have been proposed, identifying different aspects of calibration, such as overall calibration error, upper bound calibration error, class-wise calibration error, and many others. This section discusses several of these metrics and their combined use for a comprehensive evaluation of DNN calibration. *Expected Calibration Error (ECE)* [138] is one of the most widely used metrics for measuring model calibration. It quantifies the difference between accuracy and average predicted confidence by dividing predictions into fixed-size intervals (bins), computing accuracy and average confidence for each bin, and then averaging the absolute differences between the two. Despite its popularity, ECE has several limitations: (a) it was originally designed for binary classification tasks, (b) it only considers the maximum predicted probability, (c) it uses fixed-size binning, and (d) the number of bins (hyperparameter) can affect the calibration error estimate. To address these issues, several alternative metrics have been proposed. Adaptive Calibration Error (ACE) [142, 63] extends ECE by using adaptive binning instead of fixed-size bins. As DNN predictions tend to be overconfident, with high probabilities concentrated in a few intervals, adaptive binning focuses on these densely predicted intervals and applies binning to ensure an equal number of predictions in each bin. Furthermore, *Static Calibration Error (SCE)* or *Class-wise ECE* [63] extends the concept of ECE for multi-class setting by assessing calibration error for each class separately. This approach provides a more nuanced understanding of how well a model's predicted probabilities align with its true performance across different classes, helping to identify potential miscalibrations that ECE may overlook, particularly in imbalanced class distributions or when predictions beyond the top class are significant. Overall, SCE serves as a valuable tool for improving model reliability in applications where accurate probability estimates are crucial. Additionally, *Maximum Calibration Error (MCE)* [138, 37] measures the upper bound or worst-case error and is often analyzed alongside ECE or *Root Mean Square Calibration Error (RMSCE)* [143, 144] for a comprehensive understanding of calibration. These metrics assess miscalibration, which can manifest as either overconfidence or underconfidence. Without specific information about the type of miscalibration, it is challenging to develop targeted methodologies. Given that overconfidence is more prevalent in the literature (modern DNNs characteristic), *Overconfidence Error (OE)* [145] explicitly

measures the overconfidence in predicted outputs.

These metrics are primarily diagnostic tools that assess calibration but provide no insight into overall model performance. A well-calibrated model does not necessarily indicate strong performance, nor does high performance guarantee good calibration. This disconnect between calibration and performance makes these metrics less reliable when used in isolation. An alternative approach to measuring calibration is through Proper Scoring Rules (PSRs) [146], which evaluate both the calibration and performance of a model. The *Brier score (BS)* [147, 148] is a widely used PSR that calculates the mean squared difference between predicted probabilities and actual outcomes for each prediction, with lower scores indicating better calibration. In theory, the Brier score can be decomposed into distinct terms for discrimination and calibration assessment, but in practice, separating these components is challenging. Similarly, Negative Log-Likelihood (NLL) is another metric used to assess calibration. When a model assigns a high probability to the correct class, the NLL is low, whereas confident incorrect predictions result in a high NLL. This metric can be sensitive to overconfidence, especially in complex models, as small deviations from perfect predictions can lead to significant increases in the score.

In practice, there are three categories of calibration metrics based on variations in calibration conditions [149, 150]: (a) Canonical calibration, (b) Class-wise calibration, and (c) Top-label calibration. As the name suggests, Top-label calibration considers only the maximum predicted class scores, ignoring all other values. It is the weakest form of calibration among the three because, while the predicted class may be calibrated, miscalibration of the other classes goes unnoticed, potentially resulting in a perfect calibration error (CE) despite inaccuracies elsewhere. Class-wise calibration, which is stronger than Top-label, compares predicted scores for each class individually, requiring the marginal distribution for each class to align with the true distribution rather than just the joint distribution. Canonical calibration is the strongest form, as it requires the entire probability distribution to align, rather than focusing solely on the top label or marginal probabilities. For example, canonical CE is often quantified using expected $\mathcal{L}1$ or $\mathcal{L}2$ errors, which can be further generalized to $\mathcal{L}_p$ CE (*Expected Calibration Error Kernel Density Estimate* [151]), serving as a distance measure between the target distribution and the conditional distribution of the target given the input.

3

Towards Reducing Aleatoric Uncertainty for Medical Imaging Tasks

3.1 Introduction

Deep neural networks have achieved state-of-the-art performance in a variety of machine learning tasks. While prediction accuracy is an important measure characterizing the model's goodness, another critical factor contributing to the model's trust in many safety critical applications including medical imaging, is the prediction's uncertainty [33, 50, 51]. This has motivated the development of many methods to estimate the prediction uncertainty. Broadly, the uncertainty in the model's output stems from two sources as showing in Figure 3.1 - the inherent limitation of the data such as the absence of rich features, presence of noise, termed as aleatoric uncertainty, and the limitations in the learned model referred to as epistemic uncertainty. Aleatoric uncertainty, also known as *Data uncertainty* is the uncertainty caused due to errors in the measurement, i.e., uncertainty arising from the intrinsic variability in the data. It measures the variation in the output of a model due to changes in the input. The aleatoric uncertainty is supposed to be irreducible for a specific dataset; however, incorporating additional features or improving the quality of the existing features can assist in its reduction. Epistemic uncertainty, also known as *Model uncertainty* captures uncertainty in the model parameters. This uncertainty can be reduced by increasing the training data size or model capacity.

Recently, many efforts have been directed towards reducing model uncertainty using data augmentation, bayesian inference, and ensembling [152, 34, 25, 123]. However, aleatoric uncertainty has not received its due attention. As it originates from the data generation process, it cannot be explained by acquiring more data. Several factors contribute to this randomness/noise in medical images, including patient movement, different scanners, and partial volume effect. Although challenging, it is crucial to address this uncertainty in risk-sensitive applications like medical imaging to improve the robustness of the prediction against data noise when making critical decisions.

We propose a novel approach to reducing aleatoric uncertainty in one task (like segmentation) by leveraging the uncertainty estimates from another task (like reconstruction) performed on the same dataset. As a self-supervised task, image reconstruction provides a unique

opportunity to judiciously estimate aleatoric uncertainty and interpret it as noise associated with the data.

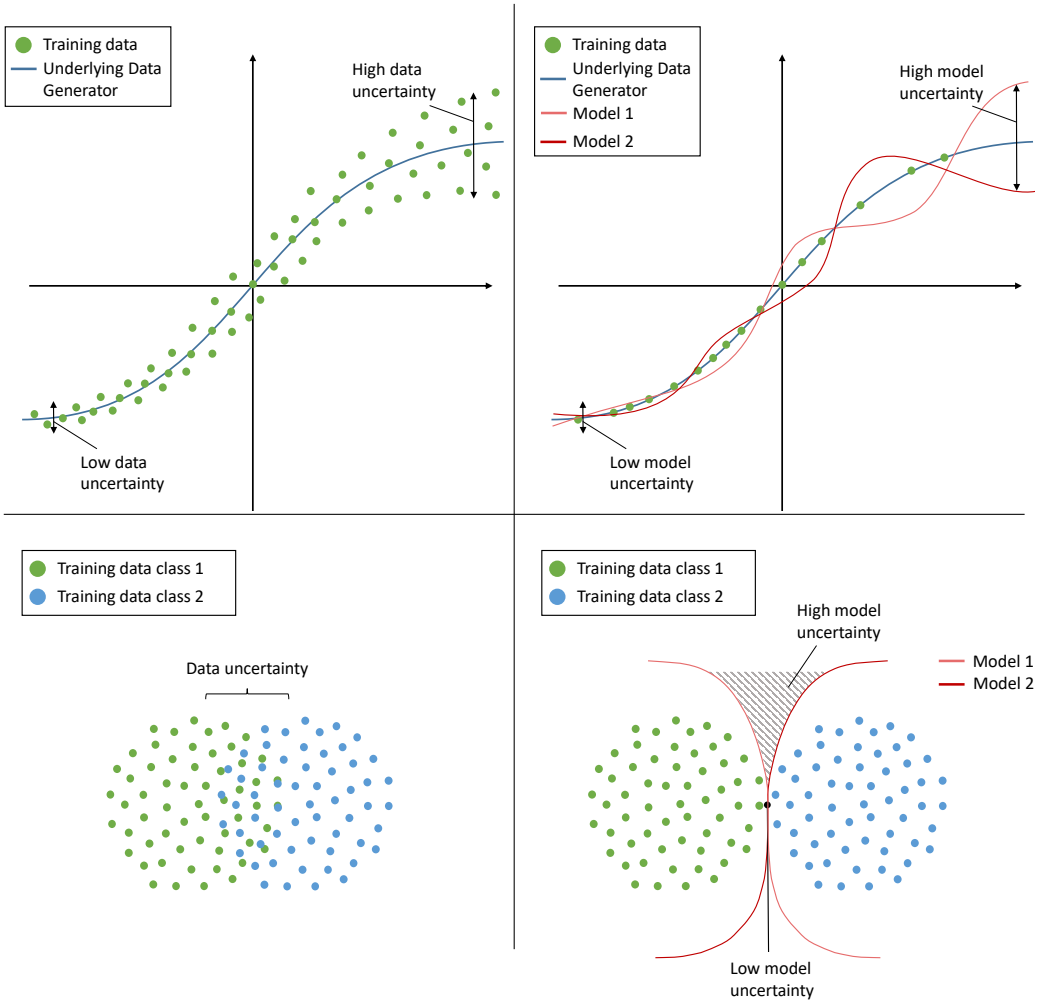


Figure 3.1: 1^{st} and 2^{nd} row represent uncertainty in regression an classification tasks, respectively whereas, 1^{st} and 2^{nd} column represent data and model uncertainty. Illustration credits [1].

3.2 Motivation

Table 3.1 provides an example of aleatoric uncertainty. Consider a house price prediction task where features such as house name, number of rooms, square footage, and garden presence are given as inputs, and the goal is to predict the house price based on these features. As shown in the table, houses *H2* and *H3* (in *red*) have identical input features but different prices. Hence, it is impossible to predict the price of the test sample. This variation represents noise in the data (i.e., the observed labels might be noisy) and is referred to as aleatoric uncertainty. Since this type of uncertainty is inherent to the dataset and cannot be removed, it is often considered irreducible in the literature. Consequently, there has been limited focus on exploring methods to reduce aleatoric uncertainty.

The key question to address here is whether aleatoric uncertainty in the given data can be

House	Rooms	Sq. Feet	Garden	`Price(Rs.)`
H1	3	2000	No	`35 lakhs`
H2	2	2500	Yes	`50 lakhs`
H3	2	2500	Yes	`30 lakhs`
H4	1	1500	No	`15 lakhs`
Test*	2	2500	Yes	`? lakhs`

Table 3.1: Example of aleatoric uncertainty using a House Price Prediction task. *H2* and *H3* are noisy samples, making it impossible to predict the price of the *Test* sample.

reduced or eliminated. If we carefully introduce a new feature that differentiates the *H2* and *H3* samples, we can use the augmented dataset to predict the price of a test sample more accurately. In Table 3.2, adding the `City` feature separates the noisy samples (*H2* and *H3*) into distinct categories, making it possible to predict the price of the test sample with greater confidence.

House	Rooms	Sq. Feet	Garden	**City**	`Price(Rs.)`
H1	3	2000	No	**Chandigarh**	`35 lakhs`
H2	2	2500	Yes	**Chandigarh**	`50 lakhs`
H3	2	2500	Yes	**Ropar**	`30 lakhs`
H4	1	1500	No	**Delhi**	`15 lakhs`
Test*	2	2500	Yes	**Ropar**	`30 lakhs`

Table 3.2: Example of aleatoric uncertainty using a House Price Prediction task. The `City` feature is added to the existing dataset. The new feature removes the noise (aleatoric uncertainty) in the dataset and allows the model to predict the price of the *Test* sample.

3.3 Methodology

Given a dataset $\mathcal{D}$, of N paired training examples, $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$, the goal is to learn a function g (a segmentation task) to predict $\mathbf{y}_i$ given $\mathbf{x}_i$ along with reducing the aleatoric uncertainty in the prediction. We achieve this by first defining an auxiliary self-supervision task f on $\mathbf{x}_i$ to estimate the inherent noise/variations in $\mathbf{x}_i$. We consider the vanilla reconstruction task as the self-supervision task. The estimated noise is then integrated into the learning process for task g to reduce the aleatoric uncertainty.

3.3.1 Modeling Uncertainty

The reconstruction based self-supervision task involves predicting $\mathbf{x}_i$ given the input $\mathbf{x}_i$ through an autoencoder (f, parameterized by ψ). Note that in the following discussion, due to the nature of the reconstruction task, $\mathbf{y}_i = \mathbf{x}_i$. We wish to capture both the aleatoric and epistemic uncertainty in the output of the reconstruction task. We employ the dropout variational inference [57] (also known as Monte Carlo dropout (MCD)) as an approximation to the posterior over the Bayesian Neural Network (BNN) to estimate these uncertainties. Drawing the model parameters $\hat{\psi} \sim q(\psi)$ from the approximate posterior we obtain the

output consisting of both $\hat{y}_{ij}$ and the aleatoric uncertainty $\hat{\sigma}(x_{ij})$. The illustration is shown in Figure 3.2.

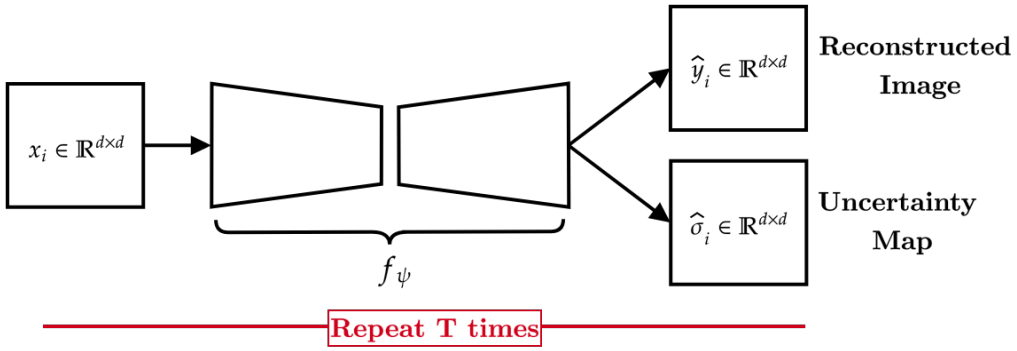


Figure 3.2: **Uncertainties in the reconstruction task.** The input is processed by a heteroscedastic NN with two output heads: **(a)***[top]* reconstruction, and **(b)***[bottom]* uncertainty head. The explicit uncertainty, denoted as $\hat{\sigma}_i$, represents the aleatoric uncertainty associated with the predictions, specifically every pixel of the reconstructed image. The NN is used with MC Dropout and the input image is passed through the model multiple times, resulting in multiple predictions and corresponding uncertainty maps. The variance across these predictions represents the epistemic uncertainty, while the average of the multiple uncertainty maps provides the mean aleatoric uncertainty (Eqn. 3.3).

Assuming a Gaussian likelihood to model the aleatoric uncertainty induces the following minimization objective:

$$\mathcal{L}_{\text{BNN}}(\psi) = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{2\hat{\sigma}(x_{ij})^2} ||y_{ij} - \hat{y}_{ij}||^2 + \frac{1}{2} \log \hat{\sigma}(x_{ij})^2 \tag{3.1}$$

where, $\hat{y}_{ij}$ is the regressed value of pixel j of image $\mathbf{x}_i$, $\hat{\sigma}$ is the noise observation parameter dependent on x_{ij} for $\hat{\psi}$ and N_i is the number of pixels in the image $\mathbf{x}_i$.

Intuitively — The heteroscedastic neural network is applied to predict the parameters of the underlying Gaussian distribution. These parameters were optimized using the negative log-likelihood (Eqn. 3.1), where the estimated mean and variance are substituted. A crucial aspect to note is that the observation noise (σ) varies with the input (x). In this approach, we perform Maximum A Posteriori (MAP) inference to estimate a single value for the model parameters (ψ). This method captures only aleatoric uncertainty, as epistemic uncertainty pertains to the model itself rather than the data.

The loss has two components - the residual error obtained through a stochastic sample, and an uncertainty regularization term. The aleatoric uncertainty is learned implicitly from the loss function. As suggested in [34] for achieving numerical stability, we train the network to predict the log variance resulting in the following minimization function, $s_i := \log \hat{\sigma}(x_{ij})^2$:

$$\mathcal{L}_{\text{BNN}}(\psi) = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{2} \exp(-s_i) ||y_{ij} - \hat{y}_{ij}||^2 + \frac{1}{2} s_i \tag{3.2}$$

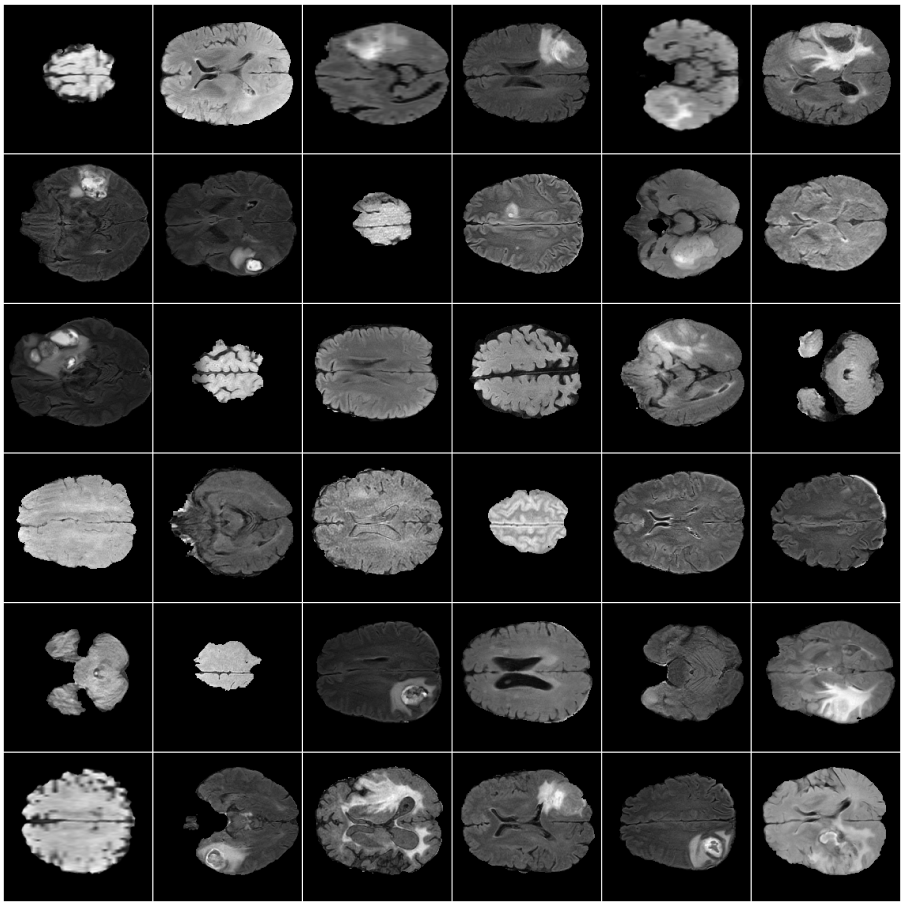
Thus, the predictive uncertainty for pixel x_{ij} can be approximated using:

$$\mathrm{Var}(y_{ij}) \approx \left(\frac{1}{T} \sum_{t=1}^{T} (\hat{y}_{ij})_t^2 - \left(\frac{1}{T} \sum_{t=1}^{T} (\hat{y}_{ij})_t \right)^2 \right) + \frac{1}{T} \sum_{t=1}^{T} (\hat{\sigma}_{ij})_t^2 \tag{3.3}$$

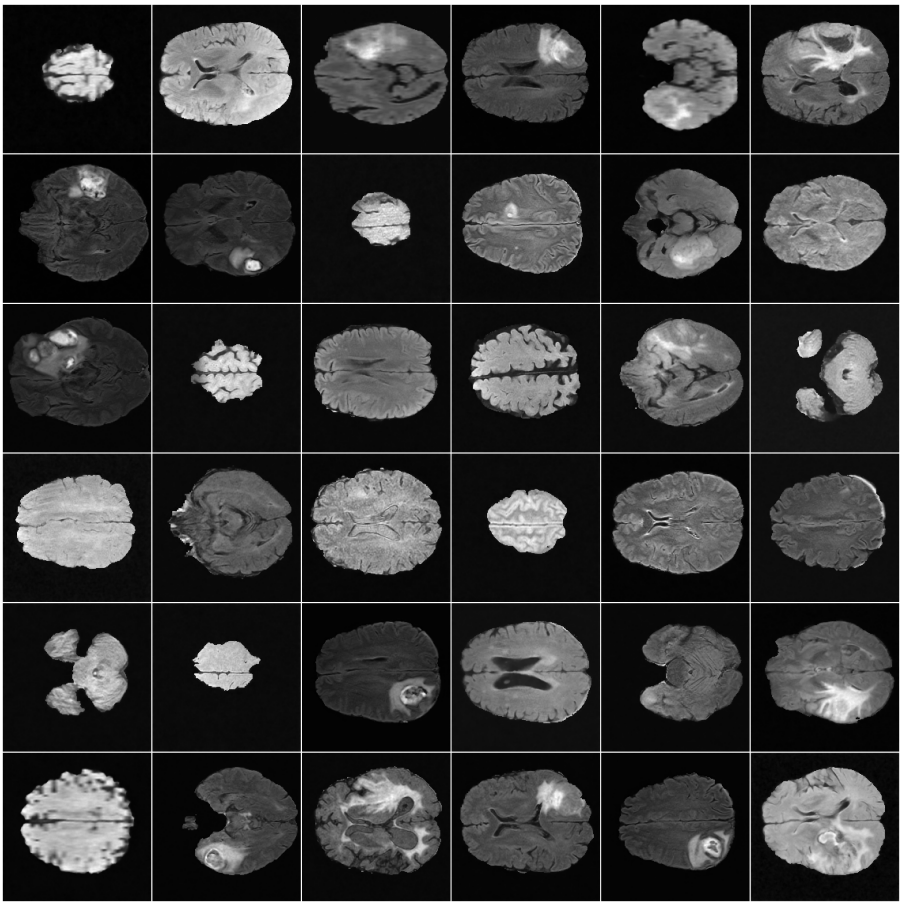
where, $(\hat{y}_{ij})_{t=1}^{T}$ and $(\hat{\sigma}_{ij})_{t=1}^{T}$ are the T sampled outputs for randomly masked weights $\hat{\psi} \sim q(\psi)$. The first and the second terms of the summation correspond to the epistemic and aleatoric uncertainties respectively.

Visualizing Reconstruction Task Output

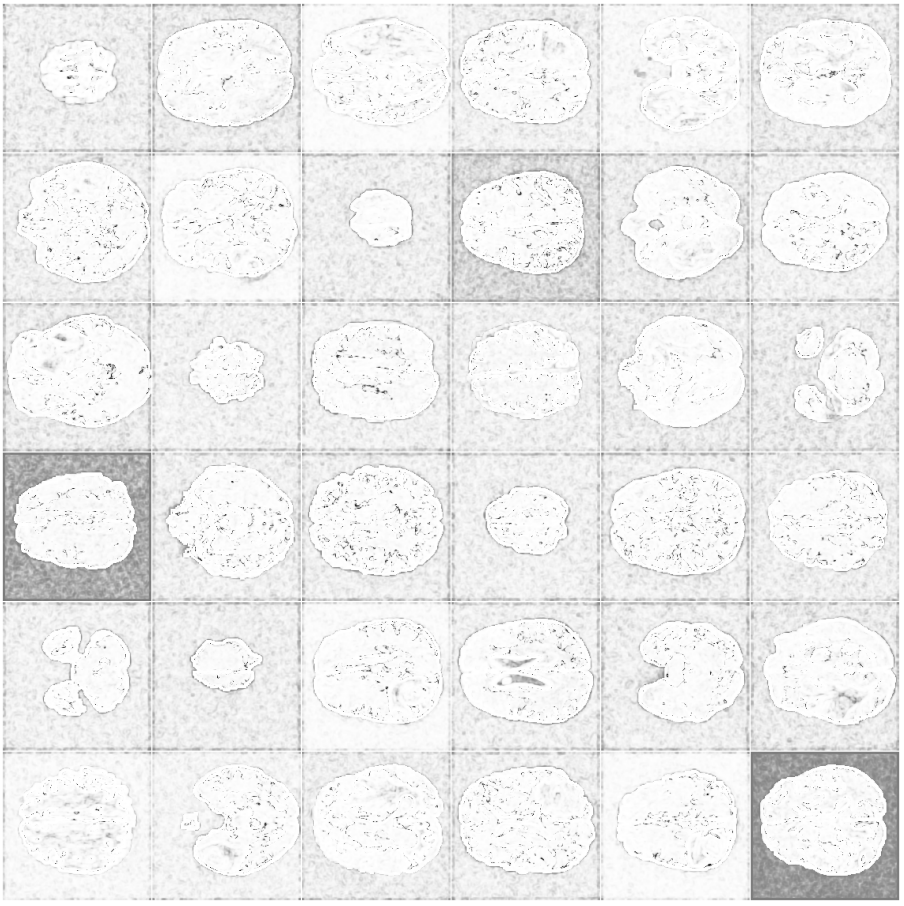
Figure 3.3 visualizes the reconstruction output of the model, which was trained to minimize the *L1* distance between the original and reconstructed images. The implementation details are provided in Section 3.4.1. As observed in Figure 3.3b, the model effectively reconstructs the original images with high fidelity. Furthermore, in Figure 3.3e, the aleatoric uncertainty is notably higher in regions where the tumor is located, indicating that the model identifies these areas as inherently noisy or ambiguous. This increased uncertainty aligns with the complexity and variability of tumor regions, which are often challenging to reconstruct accurately due to their heterogeneous nature and the inherent noise in the input data. Such insights emphasize the relevance of quantifying uncertainty in critical regions to enhance the reliability of downstream tasks, such as tumor segmentation.



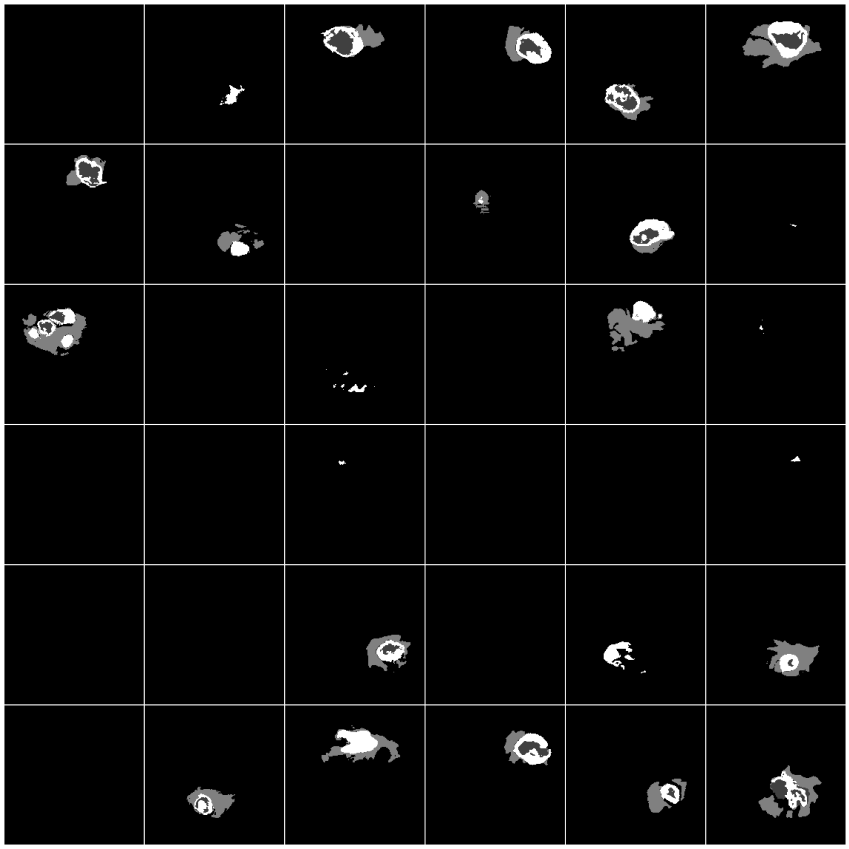
(a) Original images.



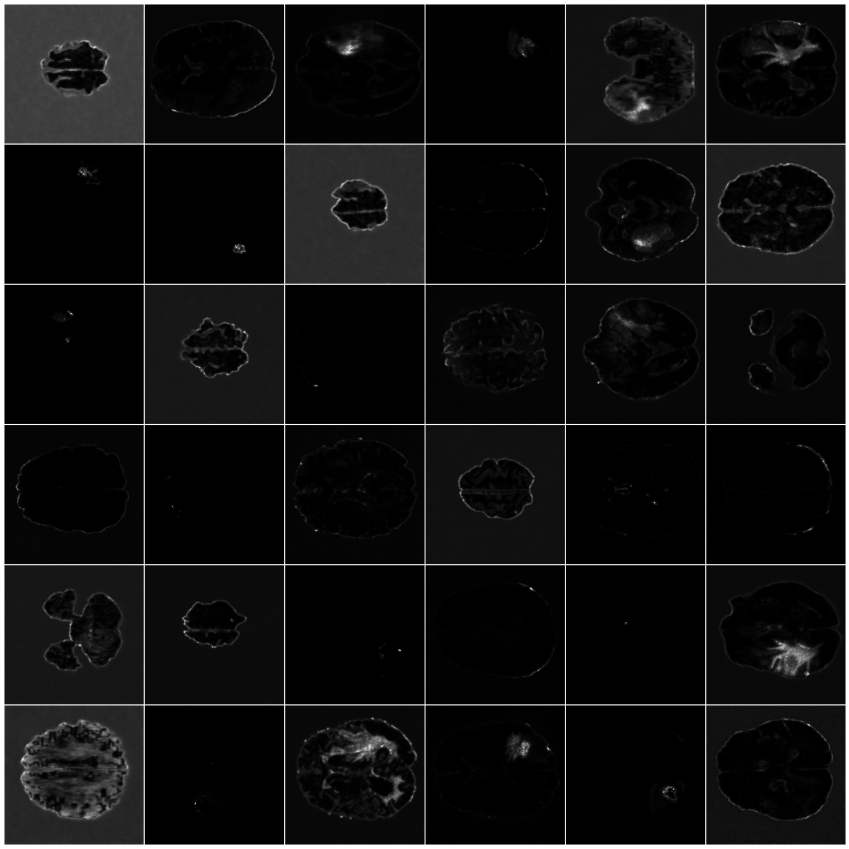
(b) Reconstructed images.



(c) SSIM loss between test and reconstructed images.



(d) Mask images



(e) Aleatoric uncertainty associated with reconstructed images.

Figure 3.3: **Visualization of the reconstruction output of the images (slices) from the test set.** Reconstructed images show high fidelity, corroborated by Structural Similarity Index Measure (SSIM) loss images. Aleatoric uncertainty is prominent where the reconstruction model finds ambiguity.

Modeling Segmentation Task Uncertainty

Segmentation is a pixel-level classification task. For each pixel j, we predict the class label and the uncertainty in the prediction. We estimate the *aleatoric uncertainty* using heteroscedastic classification neural network (NN) [34]. The heteroscedastic classification NN, g (parametrized by $\mathbf{W}$) predicts k-dimensional logit and uncertainty vectors for a k-class segmentation task. Assuming a Gaussian distribution over the logits, a sample logit vector can be obtained as shown in Eqn. 3.4. The sampled vector is squashed with the softmax function to obtain the classification probabilities.

$$\hat{\mathbf{y}}_{ij}|\mathbf{W} \sim \mathcal{N}(g_{ij}^{\mathbf{W}}, (\boldsymbol{\sigma}_{ij}^{\mathbf{W}})^2); \quad \mathbf{p}_{ij} = \text{Softmax}(\hat{\mathbf{y}}_{ij}) \tag{3.4}$$

where, $g_{ij}^{\mathbf{W}}$ and $(\boldsymbol{\sigma}_{ij}^{\mathbf{W}})^2$ are the model output and variance. *Epistemic uncertainty* of the probability vector $\mathbf{p}_{ij}$ is summarized by measuring the entropy. $\mathbf{p}_{ij}$ is approximated using Monte Carlo integration, which averages the softmax predictions for a given input over T sampled masked weights $\{\hat{\mathbf{W}} \sim q(\mathbf{W})\}_{t=1}^{T}$ where, at any given step $\mathbf{W} := \hat{\mathbf{W}}$. The illustration is shown in Figure 3.4.

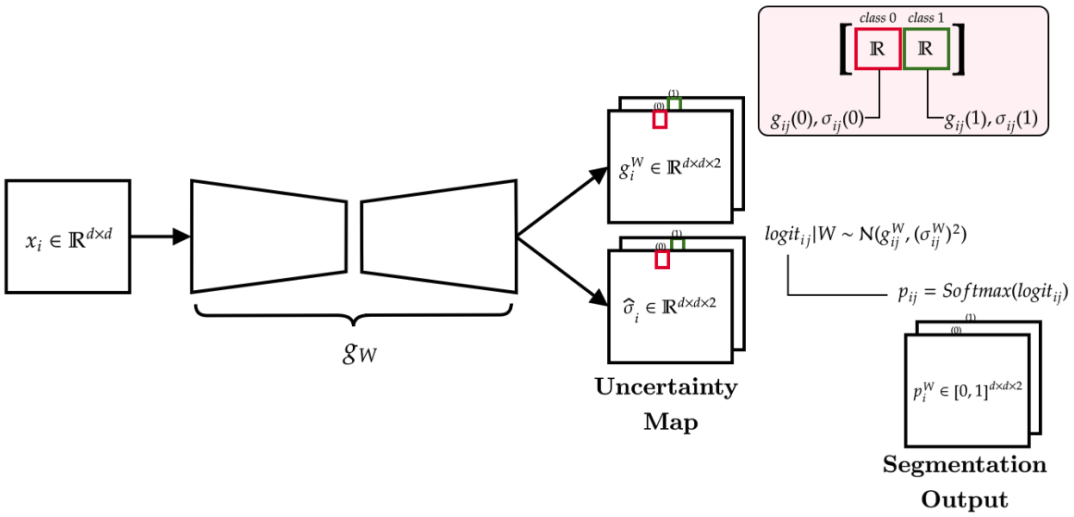


Figure 3.4: **Uncertainties in the segmentation task.** The heteroscedastic regression NN is used to estimate the distribution over the *logit space*. Here, $g_i^{\mathbf{W}}$ represents the raw prediction (real values), and $\hat{\sigma}_i$ denotes the aleatoric uncertainty associated with these predictions (logits). From the learned distribution, a new logit is sampled and subsequently passed through the softmax function to generate the segmentation mask. Here, both the segmentation output and uncertainty map are computed channel-wise.

3.3.2 Interpreting Aleatoric Uncertainty from Reconstruction Task as a Noise Model

By definition, aleatoric uncertainty captures variations in the output due to changes/noise in the input. The choice of the reconstruction as the self-supervision task gives a unique interpretation to the aleatoric uncertainty estimated at every pixel – models the noise at the pixel location. The statistics of the pixel intensities at each location can also be used

to model this noise. However, modeling dependencies between neighboring pixels becomes a challenge. On the other hand, the reconstruction task inherently models the interaction between adjacent locations and, therefore, can provide a richer model characterizing the uncertainty of the pixel intensities at each location. More concretely, the aleatoric uncertainty $\sigma(x_{ij})$ for the reconstruction task is interpreted as the variance in the pixel intensities at the j^{th} location for the image $\mathbf{x}_i$.

3.3.3 Using the Noise Model to Reduce Aleatoric Uncertainty in the Segmentation Task

Our previous interpretation of the aleatoric uncertainty allows us to explicitly account for the data noise when learning a different task using the same dataset. The heteroscedastic aleatoric uncertainty quantification from the reconstruction task provides variance in the pixel intensities at every location. We propose to augment the training data for the segmentation task by sampling pixel intensities from the noise model. Specifically, for every image, and at every pixel location, we sample the intensities from the Gaussian distribution $\mathcal{N}(\hat{x}_{ij}, \hat{\sigma}(x_{ij}))$, where, $\hat{x}_{ij}, \hat{\sigma}(x_{ij})$ are the outputs of reconstruction model. The image created through this sampling process is associated with the original image's ground truth. The augmented dataset is used to train the segmentation model. We hypothesize that this process reduces the aleatoric uncertainty in the segmentation task. The same is shown in Figure 3.5

3.4 Results and Discussion

3.4.1 Experimental Setting

The proposed aleatoric uncertainty reduction method is modeled as a data augmentation technique. Therefore, we compare our method with other standard data augmentation techniques, pixel-level augmentation (adding Gaussian noise) and structure-level augmentation (Full Augmentation) [50]. We use [34] as the framework for uncertainty estimation.

We evaluate our method, quantitatively and qualitatively, on brain tumor segmentation (BraTS 2018) with $k = 2$ classes (background versus whole tumor). We partitioned the dataset into train (60%), validation (20%) and test set (20%). As a pre-processing step, we applied intensity normalization to each MRI slice from each patient independently by subtracting the mean and dividing by the standard deviation of the brain region computed at the patient level. We cropped the input image from 240×240 to 188×188, removing the background pixels as much as possible. We have used modified UNet architecture with a dropout probability of 0.5 applied throughout the network for all the experiments. The same architecture is used for both reconstruction and segmentation tasks. We used AdamW optimizer with a learning rate of 10^{-3} and weight decay of 10^{-2}. During training, the learning rate is reduced by the factor of 0.1 with a patience of 20. In practice, we used Laplacian prior, as opposed to the Gaussian prior. The resulting loss applies an $L1$ distance

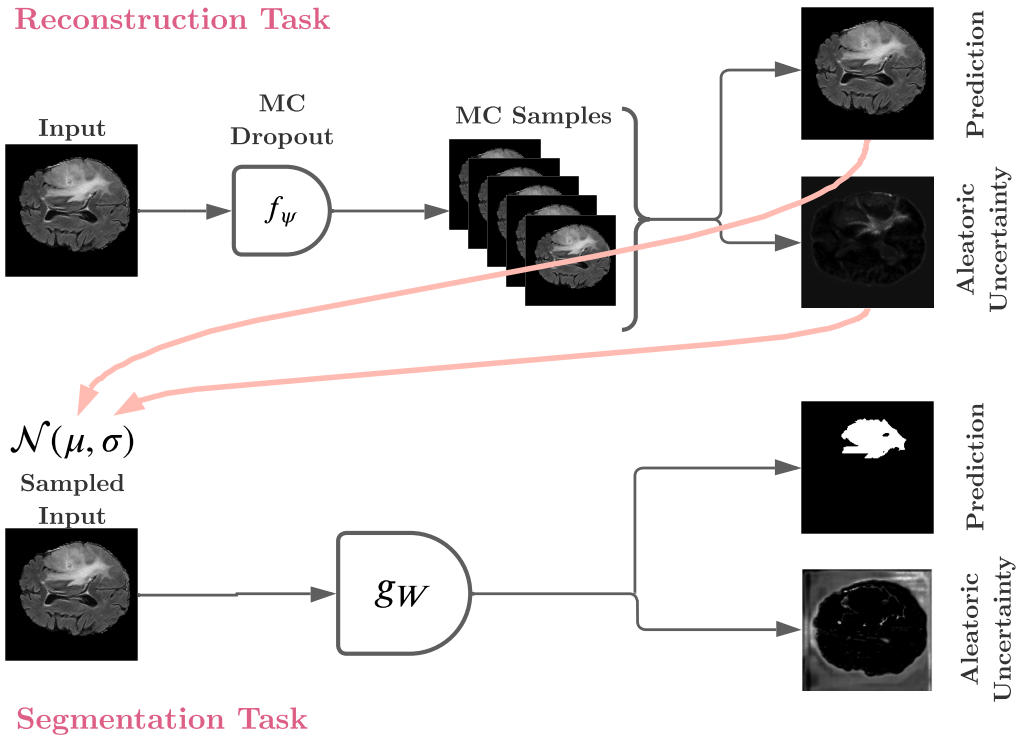


Figure 3.5: **Proposed Approach.** `Stage 1`: **(a)** Estimate aleatoric uncertainty in an auxiliary self-supervised task (image reconstruction). **(b)** Aleatoric uncertainty is associated with the prediction of the model. But in this task, prediction is the reconstructed input. So, we interpret the aleatoric uncertainty as noise present in the input data. `Stage 2`: (a) We leverage aleatoric uncertainty (interpreted as noise in the input) estimated from `Stage 1` to generate new images. It can be viewed as a data augmentation process. **(b)** The augmented dataset is used to train the segmentation model which reduces the aleatoric uncertainty in the segmentation task.

on the residuals. Based on our experimentation, we found this to perform better than $L2$ loss for the reconstruction task [34]. We used dice loss for the segmentation task. NOTE: Results of the reconstruction task are shown in Section 3.3.1.

3.4.2 Quantitative Results

By definition, the *uncertainties* are defined over the predictions, but instead of just evaluating them in the tumor region, we also compare how different augmentations perform in the non-tumor area of the brain. To evaluate the *quality of the segmentation*, we used six performance metrics: dice, precision, recall, F1, Jaccard index, and specificity, each applied to the brain region. To compare the *calibration* of the model, we used two calibration metrics: expected calibration error (ECE) and Brier score (loss). The results are presented in Table 3.3. Since the proposed approach generates new samples from the estimated Gaussian distribution (using reconstruction step), it can be regarded as a data augmentation technique. To ensure a fair comparison, we used data augmentation methods like injecting Gaussian noise into the input and full augmentation. We can see that our method shows significantly less aleatoric uncertainty than all other augmentations and the baseline, although the full

augmentation constitutes both pixel-level (elastic transform) and structure-level (vertical flip, horizontal flip, scale, shear, rotation) augmentations [50]. Our method has also outperformed the baseline and Gaussian noise augmentation on almost all the performance and calibration metrics.

Table 3.3: **Comparison between different implementations.** i) Baseline: Segmentation model with an uncertainty estimation framework. (ii) Gaussian: Gaussian noise is added to the baseline model. (iii) Full Augmentation: Baseline model with four augmentations - (a) affine image transformations - scale, shear, rotate, vertical, horizontal flipping (b) elastic transformations. ↑: Higher is better, ↓: Lower is better; Best results shown in **bold**. Statistical difference between ours and best/2nd best: ** $p < 0.001$ (highly statistically significant) & $p > 0.05$ (statistically non-significant) shown in *Italics*.

Metrics		Baseline	Gaussian	Ours	Full Augmentation
Uncertainty	Aleatoric (Tumor) ↓	$0.01760 \pm 5.4e^{-5}$	$0.00773 \pm 1.7e^{-5}$	$\mathbf{^{**}0.00154 \pm 1.5e^{-6}}$	$0.01361 \pm 3.6e^{-4}$
	Epistemic (Tumor) ↓	$0.08636 \pm 4.2e^{-3}$	$0.09399 \pm 5.9e^{-3}$	$0.08607 \pm 4.2e^{-3}$	$\mathbf{^{**}0.07601 \pm 4.6e^{-3}}$
	Predictive (Tumor) ↓	$0.10397 \pm 5.0e^{-3}$	$0.10173 \pm 6.3e^{-3}$	$\mathbf{^{**}0.08761 \pm 4.3e^{-3}}$	$0.08962 \pm 7.0e^{-3}$
	Aleatoric (Non-Tumor) ↓	$0.00369 \pm 1.4e^{-6}$	$0.00167 \pm 6.6e^{-7}$	$\mathbf{^{**}0.00050 \pm 1.6e^{-7}}$	$0.07141 \pm 7.3e^{-4}$
	Epistemic (Non-Tumor) ↓	$0.00051 \pm 1.4e^{-6}$	$0.00066 \pm 1.4e^{-6}$	$0.00043 \pm 7.3e^{-7}$	$\mathbf{^{**}0.00039 \pm 8.5e^{-7}}$
	Predictive (Non-Tumor) ↓	$0.00421 \pm 2.7e^{-6}$	$0.00233 \pm 2.1e^{-6}$	$\mathbf{^{**}0.00093 \pm 9.3e^{-7}}$	$0.07181 \pm 7.3e^{-4}$
Performance	Dice ↑	0.784 ± 0.083	0.784 ± 0.081	*0.790 ± 0.081*	**0.801 ± 0.079**
	Precision ↑	0.603 ± 0.188	0.573 ± 0.172	***0.618 ± 0.186***	0.606 ± 0.184
	Recall(Sensitivity/TPR) ↑	0.531 ± 0.168	**0.559 ± 0.175**	*0.535 ± 0.168*	0.552 ± 0.177
	F1 ↑	0.543 ± 0.165	0.544 ± 0.161	*0.548 ± 0.165*	**0.558 ± 0.169**
	Jaccard ↑	0.716 ± 0.091	0.714 ± 0.088	*0.723 ± 0.089*	**0.736 ± 0.086**
	Specificity(TNR) ↑	$0.9940 \pm 1.3e^{-4}$	$0.990 \pm 2.4e^{-4}$	$\mathbf{^{**}0.9943 \pm 1.1e^{-4}}$	$0.992 \pm 1.8e^{-4}$
Calibration	ECE ↓	0.0161	0.0168	0.0150	**0.0146**
	Brier Score ↓	0.0174	0.0183	0.0162	**0.0157**

3.4.3 Qualitative Results

In the aleatoric uncertainty maps, presented in Figure 3.6, we see higher uncertainty around the tumor boundary for the baseline and Gaussian noise-based augmentation. Our method shows minimum uncertainty compared to other augmentations (evident from the intensity scale). Although the full-augmentation uncertainty is also very less (vs. baseline/Gaussian), we see higher uncertainty in the region outside the tumor area.
We believe that the sampling process from the aleatoric uncertainty regions estimated by the reconstruction task helped to generate features, which assisted the segmentation model to reduce the aleatoric uncertainty significantly.

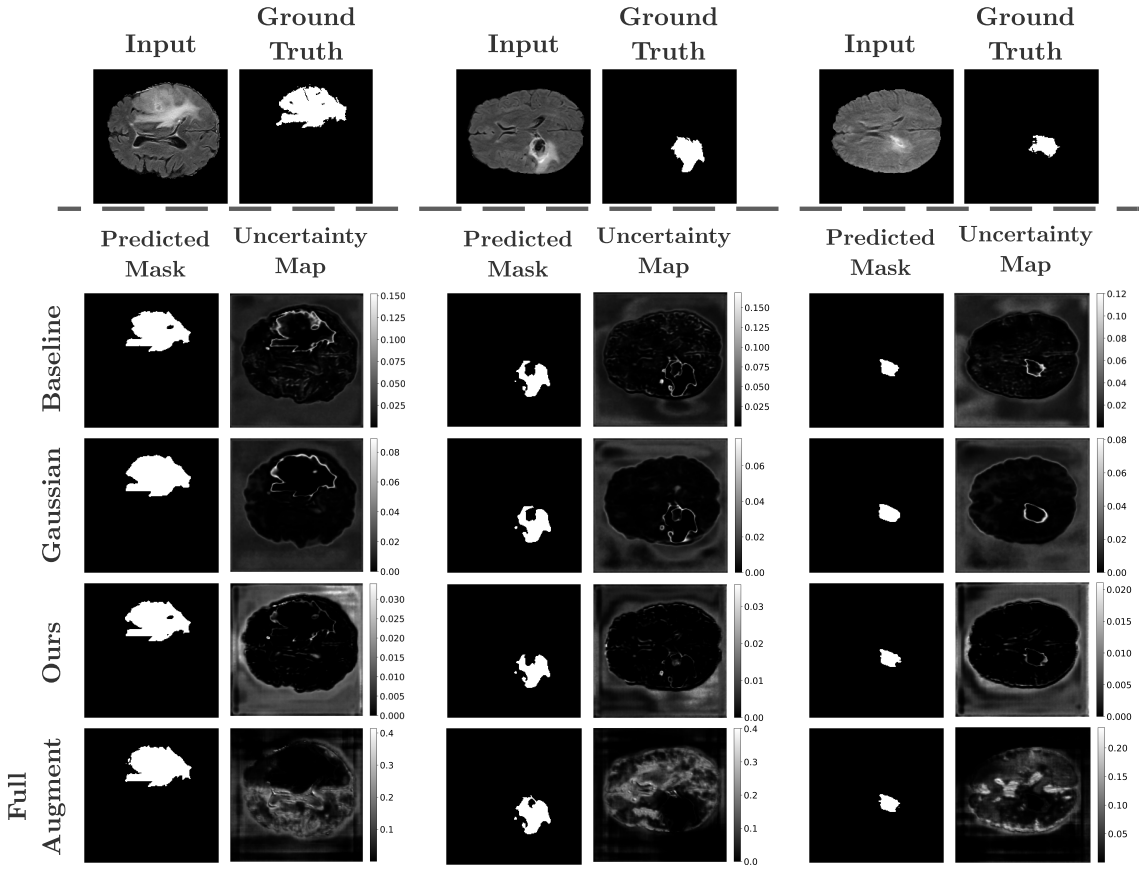


Figure 3.6: **Comparison of prediction and aleatoric uncertainty maps.** *Row 1*: Input and Ground Truth Images; *Row 2-5*: Predicted mask and Uncertainty Map of Baseline, Gaussian, Ours, Full Augmentation methods.

3.4.4 Discussion

The proposed approach effectively mitigates aleatoric uncertainty in segmentation tasks, thereby improving the reliability of model predictions. A notable advantage lies in its reconstruction step, which estimates the underlying distribution, offering a richer understanding of uncertainty by identifying the regions where the model is uncertain about the reconstruction. Leveraging this information in the downstream task (segmentation) helps generate new samples focusing on these specific areas, further reinforcing the robustness of predictions. However, these benefits come at the cost of increased computational overhead, as the additional reconstruction step adds to the framework's complexity. Despite this drawback, the method remains highly relevant for safety-critical applications, such as medical image analysis, where precise uncertainty quantification is crucial for ensuring reliable decision-making.

3.5 Conclusion

We propose a novel interpretation of aleatoric uncertainty estimated from an auxiliary self-supervised task as the noise or randomness inherent to the data and utilize it to reduce aleatoric uncertainty in other tasks performed on the same dataset. Experiments were performed on the benchmark BraTS dataset with image reconstruction as the self-supervised task and segmentation as the image analysis task. Data uncertainty estimated from reconstruction was used for data augmentation in segmentation by sampling images from the pixel predictive distribution. Our results show that the proposed approach significantly reduces aleatoric uncertainty in tumor segmentation compared to other standard augmentation methods. We further observe that the model's performance across many quantitative metrics is either better or on-par with other techniques, establishing it as a potentially reliable mechanism for addressing aleatoric uncertainty.

4 Understanding Calibration of Deep Neural Networks for Medical Image Classification

4.1 Introduction

Recent advances in deep neural networks have shown remarkable improvement in performance for many computer vision tasks like classification, segmentation, and object detection [3, 4]. However, it is essential that model predictions are not only accurate but also well calibrated [37]. Model calibration refers to the accurate estimation of the probability of correctness or uncertainty of its predictions. As calibration directly relates to the trustworthiness of a model's predictions, it is an essential factor for evaluating models in safety-critical applications like medical image analysis [153, 35, 67, 154].

Probabilities derived from deep learning models are often used as the basis for interpretation because they provide a measure of confidence or certainty associated with the predictions. When a deep learning model assigns a high probability to a particular class, it indicates a stronger belief in that prediction. For example, in medical diagnosis, a high probability assigned to a certain disease can indicate a higher likelihood of its presence based on the observed input data. However, it is important to note that the reliability of interpretation based on probabilities depends on the calibration of the model [58, 37, 59]. Calibration ensures that the assigned probabilities reflect the true likelihood of events, allowing for accurate interpretation. Without proper calibration, the interpretation based solely on probabilities may be misleading or unreliable.

Apart from directly interpreting the probabilities as confidence for decision process, several explainability methods [89] have been proposed that depend on the information extracted from the model predictions like weighting random masks [155], perturbation [96, 41], prediction difference analysis [156], contribution scores [157]. The contribution of calibration to the model's explainability lies in providing reliable probability estimates, which aid in understanding the model's decision-making process and associated uncertainties. It is observed that the improved calibration has a positive impact on the saliency maps obtained as interpretations, also improving their quality in terms of faithfulness and are more human-friendly [43]. This interplay between explainability and calibrated predictions emerges as a pivotal factor in establishing a trustworthy model for medical decision support

systems.

In healthcare, even minor errors in model prediction can carry life-threatening consequences. Therefore, incorporating uncertainty assessment into model predictions can lead to more principled decision-making that safeguards patient well-being. For example, human expertise can be sought in cases with high uncertainty. A model's predictive uncertainty is influenced by noise in data, incomplete coverage of the domain, and imperfect models. Effectively estimating or minimizing these uncertainties can markedly enhance the overall quality and reliability of the results [7, 8]. Considerable endeavors have been dedicated to mitigating both data and model uncertainty through strategies like data augmentation [158, 55], Bayesian inference [56, 57, 33], and ensembling [36, 25]

Modern neural networks are known to be miscalibrated [37] (overconfident, i.e., high confidence but low accuracy, or underconfident, i.e., low confidence but high accuracy). Hence, model calibration has drawn significant attention in recent years. Approaches to improve the calibration of deep neural networks include post-hoc strategies [60, 37], data augmentation [159, 145, 61] and ensembling [25]. Similar strategies have also been utilized in the domain of medical image analysis to explore calibration with the primary goal of alleviating miscalibration [73, 38, 80, 160]. Furthermore, recent research has also investigated the impact of different training approaches on the model's performance and calibration. These include the use of focal loss [79], self-supervised learning [69], and fully-supervised networks with pretraining [144]. However, the scope of these studies has been limited to exploring calibration in the context of generic computer vision datasets like CIFAR10, CIFAR100, and ImageNet [161, 162]. Moreover, the majority of these studies have only utilized Expected Calibration Error (ECE) as the calibration metric. Unfortunately, ECE has several drawbacks, rendering it unfit for tasks like multi-class classification and inefficient due to bias-variance trade-off [63]. Nevertheless, as reliable and accurate estimation of predictive uncertainty is important, measuring calibration is an ongoing active research area resulting in many new metrics [63, 64, 145, 37, 142].

Model calibration is tied to the training process that is inherently challenging for medical image analysis applications. The scarcity of labeled training datasets is a major cause for concern [65, 2]. Gathering labeled data for the medical domain is a daunting task due to the complex and intricate annotating process requiring domain expertise. *Transfer learning* is a popular learning paradigm to circumvent the labeled training data scarcity [66, 67]. Although transfer learning improves model accuracy, especially for smaller datasets, it also improves the quality of various complementary model components like adversarial robustness, and uncertainty [144]. Remarkably, the literature suggests that the advantages of popular methods such as transfer learning on classical computer vision datasets do not extend to medical imaging applications [68]. *Self-supervised learning (SSL)* is another promising training regime when learning from scarce labeled data in classical computer vision applications [71, 72]. Though fully-supervised (pretrained) and self-supervised approaches seem to improve various model performance measures like accuracy, robustness, and uncertainty [69, 70], the impact of the training regime(s) on model calibration is

under-explored.

Our current work addresses these crucial gaps in the literature – understanding the calibration of deep neural networks for medical image analysis in the context of different training regimes and several calibration metrics. Accordingly, our main contributions are:

1. We study the effect of different training regimes on the performance and calibration of models used for medical image analysis. Specifically, we compare three different training paradigms: Fully-Supervised with random initialization (FS_r), Fully-Supervised with pretraining (FS_p), and Rotation-based Self-Supervision with pretraining (SSL_p).

2. We leverage several complementary calibration metrics to provide an accurate, unbiased, and comprehensive evaluation of the predictive uncertainty of models.

3. We assess the influence of varying dataset sizes, architecture capacities, and task complexity on the performance and calibration of the models.

4. We identified some of the potential factors that are correlated with the observed changes in the calibration of models. These include layer-wise learned representations as well as the weight distribution of the model parameters.

In general, we observe that the rotation-based self-supervised pretrained training approach provides better calibration for medical image analysis tasks than its fully supervised counterpart, with on-par or better performance. Additionally, our findings contradict recent literature [68] that remarked *"transfer offers little benefit to performance"* for medical datasets. Furthermore, both the weight distribution and the learned representation analysis indicate that self-supervised training provides implicit regularization that in-turn achieves better calibration.

4.2 Methods

4.2.1 Training Regimes

Fully-Supervised and Transfer Learning

In a fully-supervised training regime, we use the given input data and the corresponding target value to learn the task. We can train models using two different ways, learning from scratch, i.e., initializing model weights randomly, or pretraining, i.e., transferring knowledge from one task to another by using the learned weights. In the transfer learning approach, a model is first pretrained using supervised learning on a large labeled dataset [3, 163]. Then the learned generic representations are fine-tuned on the in-domain medical data [68, 164]. Generally, fine-tuning a pretrained model achieves better generalized performance and faster convergence than training a fully-supervised network from scratch [165, 166].

We have considered FS_r as a baseline in our experiments where the model is trained from scratch. ImageNet pretraining is used as the default pretraining approach, which has shown remarkable performance on medical imaging datasets [164].

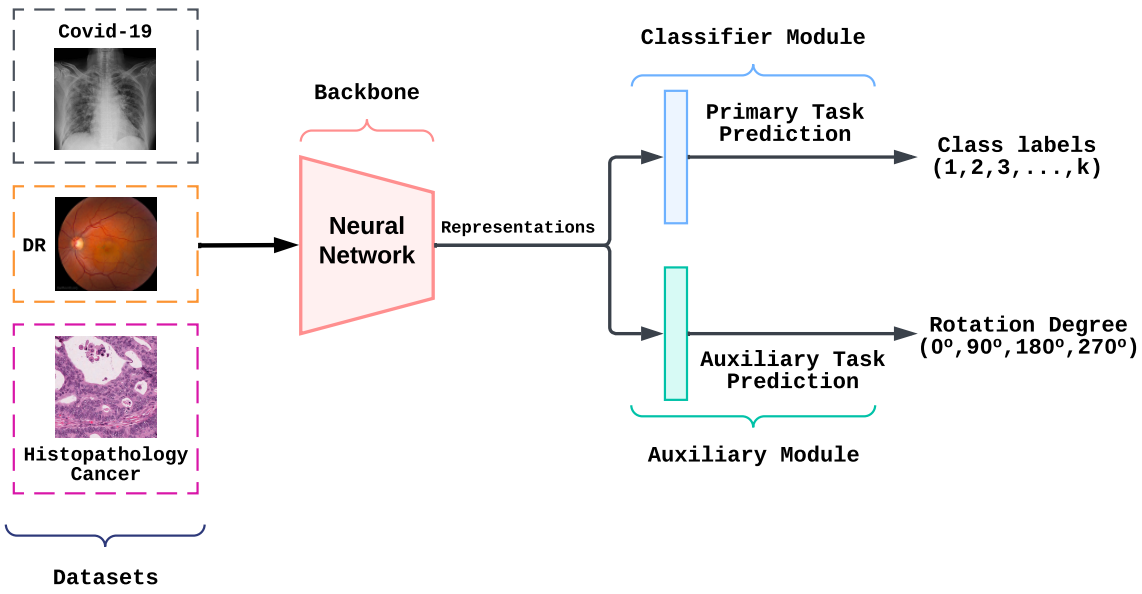


Figure 4.1: Self-Supervised Learning Framework

Self-Supervised Learning

In self-supervised training regime [69, 167], Figure 4.1, we train a classifier network with a separate auxiliary head to predict the induced rotation in the image. The output of the penultimate layer is given to both the classifier and the auxiliary module. The classifier predicts a k-way softmax output vector based on the chosen task/dataset, whereas the auxiliary module predicts a 4-way softmax output vector indicating the rotation degree (0°, 90°, 180°and 270°). Given a dataset $\mathcal{D}$, of N training examples, $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$, the goal is to learn representations using a self-supervised regime. The overall loss during training is the weighted sum of vanilla classification and the auxiliary task loss

$$\mathcal{L}(\theta) = \mathcal{L}(y, p(y|R_r(x)); \theta) + \lambda \mathcal{L}_{aux}(r, p(r|R_r(x)); \theta) \tag{4.1}$$

where, $R_r(x)$ is a rotation transformation on input image x and $r \in \{0°, 90°, 180°, 270°\}$ is the ground truth label for the auxiliary task. Note that the auxiliary component does not require ground truth training label y as input. $\mathcal{L}_{aux}$ is the cross-entropy between r and the predicted rotation.

4.2.2 Calibration Metrics

Perfect Calibration: In a multi-class classification problem, let the input be x and the label $y^* \in \{1, \cdots, K\}$ and f the learned model. The model's output is $f(x) = (\hat{y}, \hat{p})$ where $\hat{y}$ is a class prediction and $\hat{p}$ is its associated confidence. If $\hat{p}$ is always the true probability, then we call the model perfectly calibrated as defined in (4.2).

$$\mathbb{P}\left(\hat{y} = y^* \mid \hat{p} = p\right) = p, \quad \forall p \in [0, 1] \tag{4.2}$$

The difference between the true confidence (accuracy) and the predicted confidence (output probability), $|\mathbb{P}(\hat{y} = y^* \mid \hat{p} = p) - p|$ for a given p is known as calibration error or miscalibration. Note that $\hat{p}$ is a continuous random variable, the probability in (4.2) cannot be computed using finitely many samples resulting in different approximations for the calibration error as discussed below.

Expected Calibration Error (ECE)

The most common miscalibration measure is the ECE [138, 37], which computes the difference in the expectation between confidence and accuracy. It is a scalar summary statistic of calibration.

$$\mathbb{E}_{\hat{p}}\left[|\mathbb{P}(\hat{y} = y^* \mid \hat{p} = p) - p|\right] \tag{4.3}$$

In practice, we cannot estimate ECE without quantization; therefore, the confidence scores for the predicted class are divided into m equally spaced bins. For each bin, the average confidence (conf) and accuracy (acc) are computed. The difference between the average confidence and accuracy weighted by the number of samples summed over the bins gives us the ECE measure. Formally,

$$\text{ECE} = \sum_{m=1}^{M} \frac{n_m}{N} |\operatorname{acc}(m) - \operatorname{conf}(m)| \tag{4.4}$$

where n_m is the number of predictions in bin m. While ECE is used extensively to measure calibration, it has some major drawbacks [63]:

(i) Structured around binary classification, ECE only considers the class with maximum predicted probability. As a result, it discounts the accuracy with which the model predicts other class probabilities in a multi-class classification setting.

(ii) Deep neural network predictions are typically overconfident, causing skewness in the output probabilities. Consequently, equal-interval binning metrics like ECE is impacted by only a few bins.

(iii) The number of bins, as a hyperparameter, plays a crucial role in the quality of calibration estimation. However, determining the optimal number of bins is challenging due to the bias-variance tradeoff.

(iv) In a static binning scheme like ECE, overconfident and underconfident predictions occurring in the same bin result in a reduction of calibration error. In such cases, it is difficult to infer the true cause of improvement in model calibration.

These issues have resulted in the development of novel calibration metrics discussed in the following subsections.

Adaptive Calibration Error (ACE)

As *ECE* suffers from skewness in the output predictions, ACE mainly focuses on the regions where the predictions are made. It uses an adaptive binning scheme to ensure an equal number of predictions in each bin [142, 63]. Formally,

$$\text{ACE} = \frac{1}{KR}\sum_{k=1}^{K}\sum_{r=1}^{R}|\text{acc}(r,k) - \text{conf}(r,k)| \tag{4.5}$$

where, $\text{acc}(r,k)$ and $\text{conf}(r,k)$ represent the accuracy and confidence for the adaptive calibration range or bin r and class label k, respectively. Due to adaptive binning, the bin spacing can be unequal; wide in the areas where the number of data points is less, and narrow otherwise.

Maximum Calibration Error (MCE)

It refers to the upper-bound estimate of miscalibration useful in safety-critical applications. MCE [138, 37] captures the worst-case deviation between confidence and accuracy by measuring the maximum difference across all bins m, as shown below:

$$\text{MCE} = \max_{m\in\{1,\ldots,M\}}|\text{acc}(m) - \text{conf}(m)| \tag{4.6}$$

Overconfidence Error (OE)

Modern deep neural networks provide high confident outputs despite being inaccurate. Thus a metric that captures the model's overconfidence provides better model insights. OE [145] captures the overconfidence in the model prediction by penalizing the confidence score only when the model confidence is greater than the accuracy.

$$\text{OE} = \sum_{m=1}^{M}\frac{n_m}{N}\Big[\text{conf}(m)\times\max\Big(\text{conf}(m) - \text{acc}(m), 0\Big)\Big] \tag{4.7}$$

Brier or Quadratic Score

It is a strictly proper scoring rule that measures the accuracy of the probabilistic predictions [147, 148, 168]. It is the mean squared difference between one-hot encoded true label and predicted probability. Formally,

$$\text{Brier} = \sum_{k=1}^{K}(\mathbb{1}_{[y^*=k]} - \hat{p}(y^* = k \mid x))^2 \tag{4.8}$$

Negative Log Likelihood (NLL)

For safety-critical applications, using a probabilistic classifier that predicts the correct class and gives the probability distribution of the target classes is encouraged. Using NLL, we

can evaluate models with the best predictive uncertainty by measuring the quality of the probabilistic predictions [149, 169, 170]. Formally,

$$\text{NLL} = -\sum_{k=1}^{K} \mathbb{1}_{[y^*=k]} \log[\hat{p}(y^* = k \mid x)] \tag{4.9}$$

Additionally, *Root Mean Square Calibration Error (RMSCE)* [142, 144, 143] measures the square root of the expected squared difference between confidence and accuracy. As it defines the magnitude of miscalibration, it is highly correlated to ECE. Similar to ACE, *Static Calibration Error (SCE)* [63], extends ECE by measuring calibration over all classes in each bin for a multi-class setting but does not use an adaptive binning approach. As a result, we exclude these metrics from our experimental analysis.
It can be observed from the above definitions that none of the individual metrics takes a holistic approach. Hence, it is important to recognize that individual metrics are limited in their ability to provide accurate estimates of calibration. Consequently, a collective evaluation of these metrics is necessary for a better or unbiased understanding of calibration performance.

4.2.3 Experimental Setup

Datasets

We used three different datasets to investigate the classification performance and calibration of models trained under different regimes. The datasets have varying characteristics such as different imaging modalities, and sizes.

- The Diabetic Retinopathy (DR) dataset contains 35K high-resolution ($\sim 5000 \times 3000$) retinal fundus scans [171]. Each image is rated for the severity of diabetic retinopathy on a scale of 0-4, which makes it a five-class classification problem. The images are captured under varying imaging conditions, like different models and camera types.

- The Histopathologic Cancer dataset contains 220K images (patches of size 96×96) extracted from larger digital pathological scans [172, 173, 174]. Each image is annotated with a binary label indicating the presence of tumor tissue in the histopathologic scans of lymph node sections.

- The COVID-19 is a small dataset consisting of 317 high-resolution ($\sim 4000 \times 3000$) chest X-rays images [175, 176, 177]. This dataset corresponds to a three-class classification problem.

Both DR and Histopathology cancer datasets are segregated into four training datasets of sizes: 500, 1000, 5000, and 10000; and a common test dataset of 2000 images. The *Covid-19* dataset is partitioned into 60/20 train/validation split and a separate 20% test set for evaluation. The images in all the datasets are resized to 224×224, which is the standard input resolution for ResNet architectures.

Implementation Details

Architectures – Due to the popularity of ResNet architectures in medical imaging for classification tasks [165, 164, 66], we choose the standard ResNet18, ResNet50 [178], and WideResNet [179] architectures as the network backbone to simulate small, medium, and large architecture sizes, respectively. The details are mentioned in Table 4.1. For the training regimes relying on a pretrained model, we initialize the backbone architectures using ImageNet-pretrained weights, and the classifier and self-supervised modules using the Kaiming uniform initialization [180] variant. *WideResNet (*WRN-d-k*:)* It is a variant of residual networks to simulate large architecture size. The depth and width of WideResNet are regulated by a deepening factor d and a widening factor k. We used *WRN-50-2* for our experiments, i.e., WideResNet with 50 convolutional layers and a widening factor of 2.

Table 4.1: Overview of the models used in this study.

Model Name	Number of Layers	Parameters
ResNet18	18 layers	11M
ResNet50	50 layers	23M
WideResNet	ResNet50, 2×width	66M

Evaluation Metrics – We use two performance metrics - *Accuracy* and *Area under the Receiver Operating Characteristic curve (ROC AUC)*; and six calibration metrics - *ECE, MCE, ACE, OE, Brier* and *NLL*.

4.3 Results

4.3.1 Effect of Training Regimes on Calibration

In this study, we investigate the performance and calibration of three different architectures - *ResNet18, ResNet50 & WideResNet* using three different training regimes - *Fully-Supervised with random initialization* (FS_r), *Fully-Supervised with pretraining* (FS_p) and *Rotation-based Self-Supervision with pretraining* (SSL_p).
For medical image analysis, both the accuracy and reliability of the models are crucial. In this context, there are two key scenarios we need to consider:

1. *High accuracy and high calibration error* – When a model has high accuracy but is miscalibrated, the model's predictions may not be trustworthy. Both incorrect predictions with high confidence and correct predictions with low confidence are detrimental in healthcare applications. Reliance on accuracy alone is hazardous.

2. *High accuracy and low calibration error* – This is the ideal scenario, where a model has high accuracy and well-calibrated confidence scores. Predictions from such a model can be trusted in the decision-making process.

Hyperparameter Details

We used the batch size=16, epochs=300, optimizer=SGD, learning rate, lr=0.001, momentum=0.9, and weight decay=0.0005 parameter values for FS_r, FS_p, and SSL_p training regimes across all datasets and architecture sizes for our experiments. For pretrained setups, FS_p and SSL_p, we trained the classifier and auxiliary module for the first 30 epochs with a lr=0.001 and then fine-tuned the complete network with lr=$1e^{-5}$. In SSL_p training, $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ is empirically chosen based on the best validation accuracy.

4.3.2 Effect of Architecture and Dataset Size

In this section, we present the findings of our analysis of the DR dataset. The performance and calibration scores of various architectures, as well as the effects of increasing training dataset size, are depicted in Figure 4.2 for the three different training regimes. Similar results and analysis on the Histopathology dataset is presented in Figure 4.3.

Owing to the difficulty of the task, the performance of all training regimes across all the models is not very high ($\leq 75\%$). However, we do see a clear improvement in performance as the training dataset size increases across all architectures and regimes. Additionally, we observe that initializing models with pretrained weights (with SSL_p having an edge over FS_p) offer a significant advantage over random initialization, which contradicts existing assumptions that transfer learning from ImageNet models is not beneficial. Both FS_p and SSL_p result in similar performance when using larger models [68].

Comparing the effect of FS_p and SSL_p training regimes on calibration, we see that SSL_p significantly improves calibration across all metrics for all architectures and training dataset sizes as illustrated in Figures 4.2(c)-(h). The gap in the calibration metrics for SSL_p and FS_p is highest when using the largest architecture (WideResNet). While a randomly initialized model (FS_r) results in marginally better calibration (sometimes even better than SSL_p), the performance is significantly poor. Overall, we observe that models trained using self-supervision with pretrained weights show better or similar performance with a significant improvement in calibration error compared to fully-supervised pretraining. These results suggest that self-supervised training can help improve both performance and calibration, leading to more robust and reliable models for medical image analysis.

We discuss the results on the Covid-19 dataset separately owing to its small size. Figure 4.4 depicts that all the models result in high performance on this dataset indicating the ease of learning the task. The superior performance of FS_p and SSL_p indicate a definite advantage of transfer through pretrained over random init, contradicting the recent findings [68]. It is also evident that larger models result in better performance than shallow models. The negative impact of training from a random init (FS_r) for over-parameterized models is also evident from the drop in the performance and calibration with the increase in architecture size. While we observe a significant difference in the performance, there is only a marginal change in the calibration metrics. There is no definite trend in the calibration across the three training regimes. Thus, while transfer seems to have a positive impact on performance, calibration does not enjoy a commensurate impact.

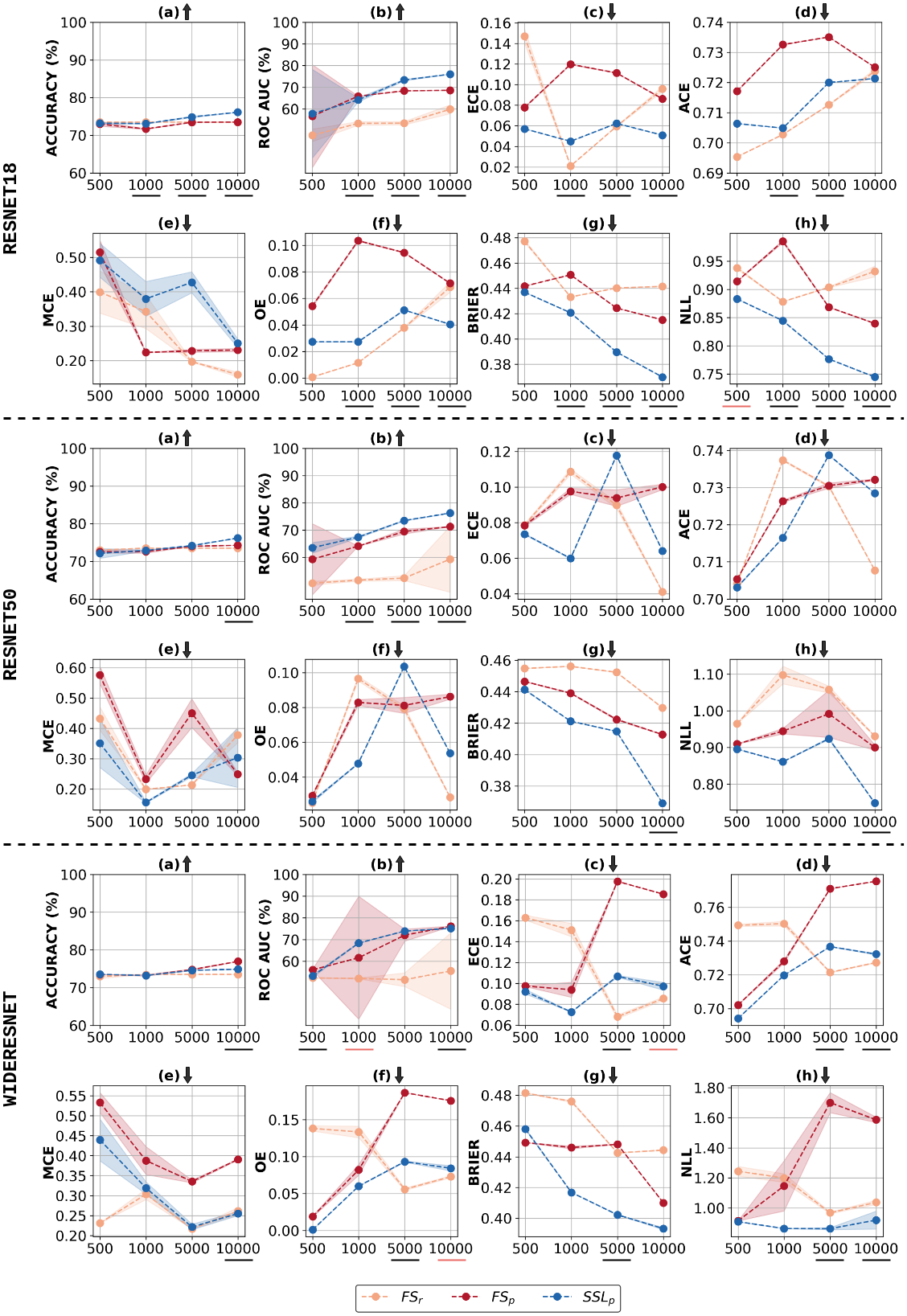


Figure 4.2: **Joint evaluation for performance and calibration across different dataset sizes (x-axis) and architectures for DR dataset.** The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials. The underline shows the statistical significance between FS_p and SSL_p. Black and Pink color signifies $p < 0.05$ and $0.05 < p < 0.1$ level of significance, respectively. ↑: higher is better, ↓: lower is better.

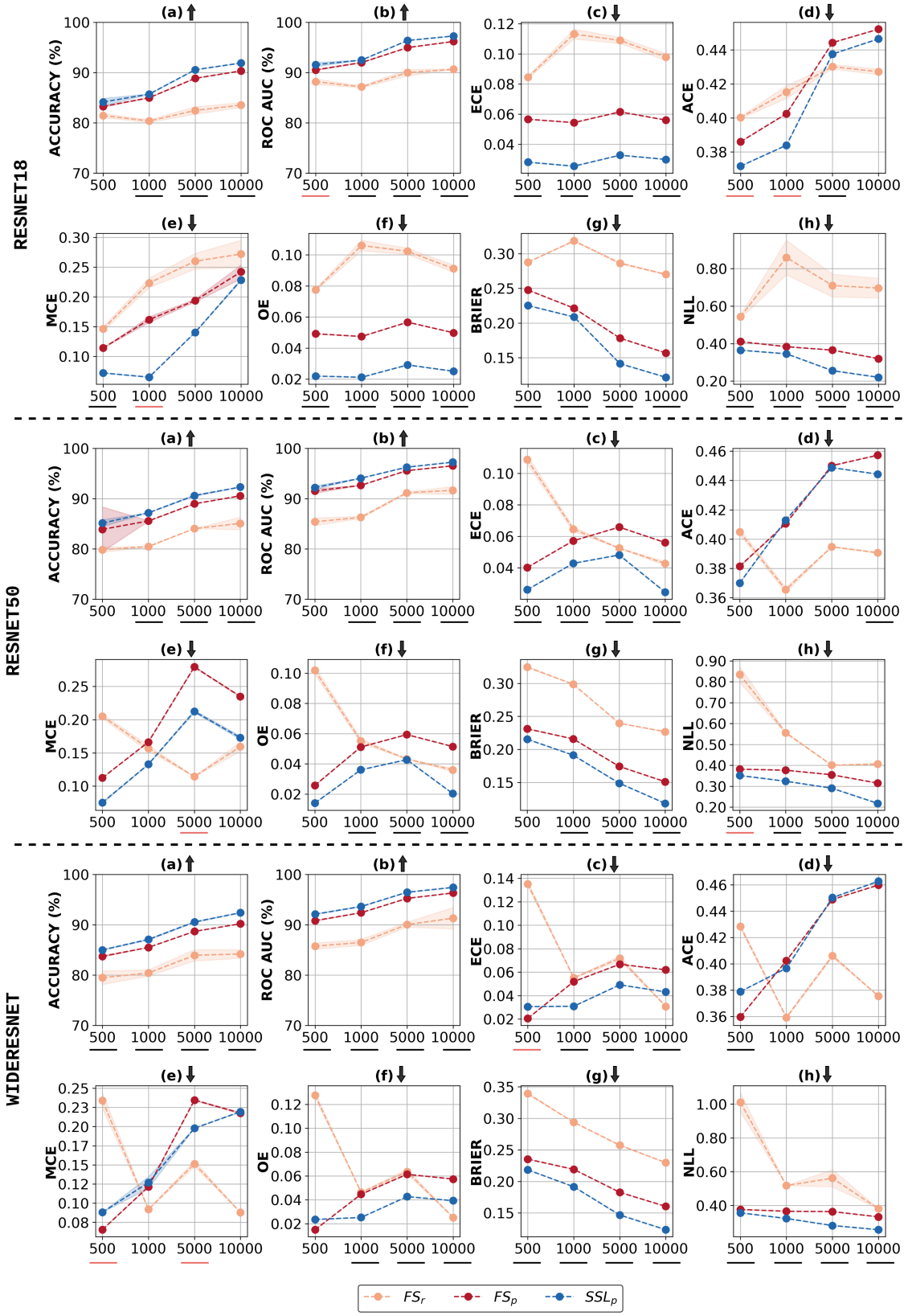


Figure 4.3: **Joint evaluation for performance and calibration across different dataset sizes (x-axis) and architectures for Histopathology Cancer dataset.** The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials. The underline shows the statistical significance between FS_p and SSL_p. Black and Pink color signifies $p < 0.05$ and $0.05 < p < 0.1$ level of significance, respectively.

4.3.3 Issues with using Single Calibration Metric

In this section, we discuss the importance of collective evaluation of calibration metrics. For this purpose, let's consider the question - *Does transfer learning improve calibration?* In the context of DR dataset, we analyze the results in Figure 4.2. Comparing FS_r and FS_p using only *Brier* for all architectures and dataset sizes, the general trend we observe is that transfer learning improves calibration. However, this observation fails when we chose *ECE* metric, which gives us mixed results. Similarly, incorrect conclusions could be drawn when using individual metrics like *NLL* and *ACE*. Likewise, we consider the

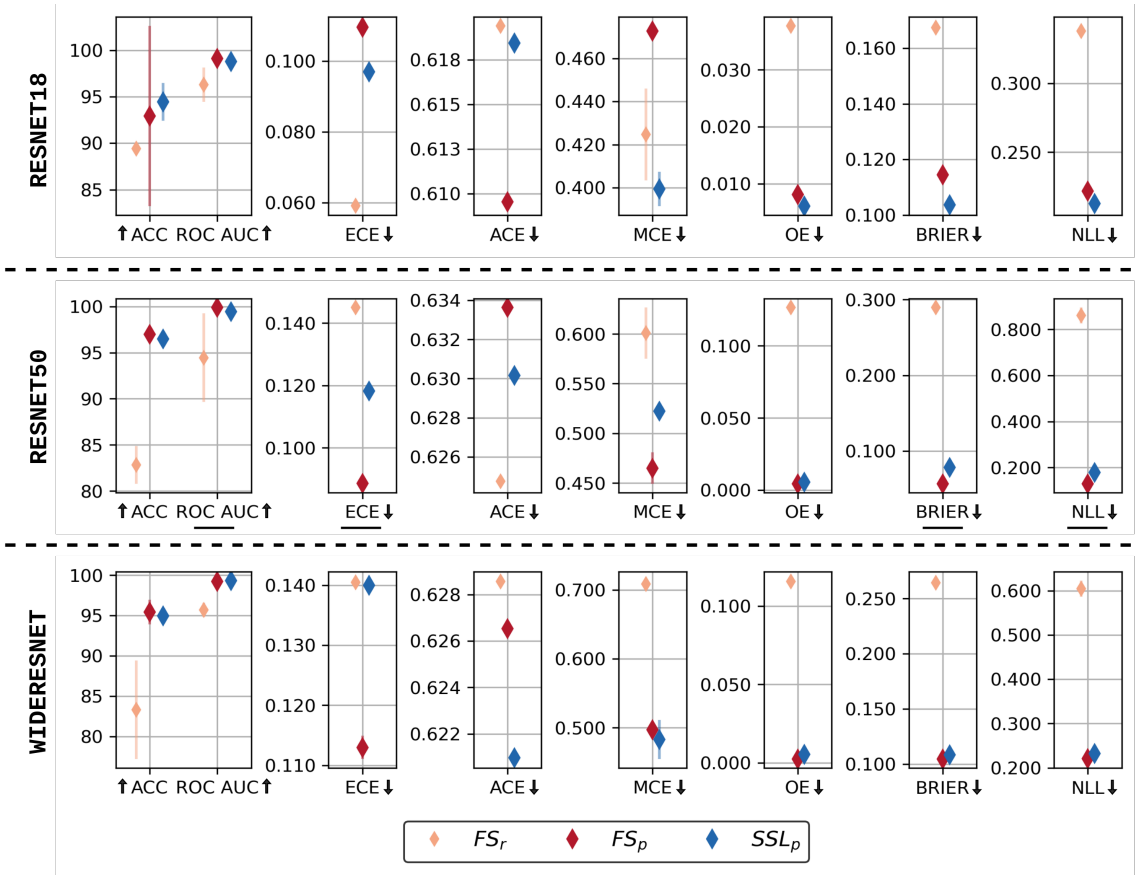


Figure 4.4: **Comparing performance and calibration across different architectures and training regimes for *Covid-19* dataset.** The error bars correspond to $\mu \pm \sigma$, estimated over 3 trials. Relying on a single calibration error metric, such as ECE or ACE, can lead to conflicting conclusions when it comes to model selection. By considering a combination of metrics, we gain a more comprehensive understanding of the model's calibration performance. ↑: higher is better, ↓: lower is better.

effect of architecture on performance and calibration in the context of the small Covid-19 dataset. From Figure 4.4, we observe that FS_p and SSL_p have comparable performances with nominal improvement with increasing architecture size. In this case, using only *ECE* as the calibration metric would lead us to infer that FS_p provides better calibration than SSL_p for large capacity models. In contrast, *ACE* suggests the opposite. However, these two training regimes are quite similar across most other metrics.

These examples further highlight that in scenarios where models provide mixed calibration

results, selecting the best model is non-trivial/subjective. In section 4.4, we discuss some potential model selection criteria to address this issue.

4.3.4 Factors affecting Performance and Calibration

In this section, we explore two potential factors linked to the enhanced calibration of the self-supervised training regime. Firstly, we examine the standard deviation of weight distributions and calibration metrics across different training regimes. Secondly, we investigate the similarity of learned representations in the activations.

Weight Distribution

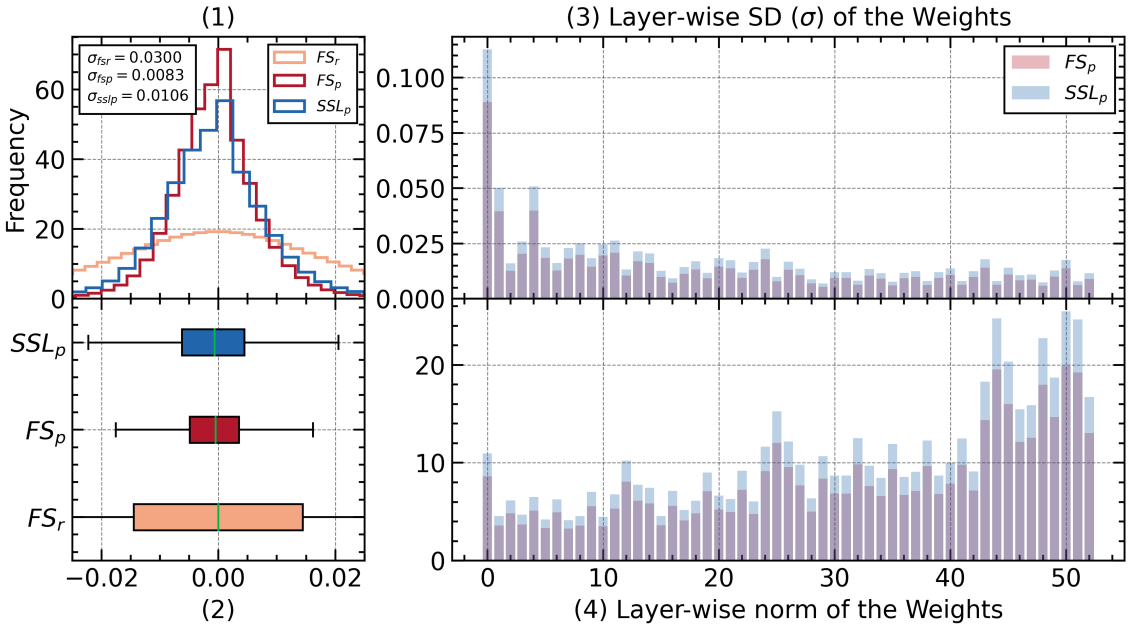
The weight distribution of a neural network can provide useful insights into the model's performance. Regularization schemes like $\mathcal{L}_1$, $\mathcal{L}_2$, dropout [181, 115] are often employed to find optimal parameters of a model with low generalization error. By adding a parameter norm penalty term to the objective function, the $\mathcal{L}_1$ and $\mathcal{L}_2$ norms encourage sparse weights with many zero values and small weight values respectively. Weighting the contribution of the penalty term controls the regularization effect. For instance, with $\mathcal{L}_2$ norm, the histogram of weights tends to a zero-mean normal distribution with a high penalty that causes the model to underestimate the weights and hence leads to underfitting. In contrast, a low penalty yields a flatter histogram that causes the model to overfit the training data. To strike the right balance, careful hyperparameter tuning is needed to determine the data-dependent optimal penalty term contribution for better generalization. Based on this intuition, we attempt to interpret the performance and calibration of networks trained using different regimes using weight distribution analysis. To the best of our knowledge, the calibration of a model has not been explained in the context of the weight distribution of a network, especially for medical image analysis.

The comparison of weight distributions between the models trained using FS_r, FS_p, and, SSL_p for the DR dataset in Figure 4.5a-(1),(2) reveals some interesting observations. The weight distribution of the model trained with FS_p exhibits a higher peak than SSL_p, indicating that most of the weights are small. Conversely, the FS_r model exhibits the highest standard deviation, resembling a uniform distribution. Now, the question arises: which distribution is preferable, and which scenario leads to better generalization with improved calibration? To address this, we analyze the impact of weight distribution on the performance and calibration of FS_p and SSL_p models using Figure 4.2 and Figure 4.5. We observe that both models show similar AUC performance, with SSL_p displaying a smaller peak in the weight distribution. This difference in weight distribution influences the calibration metrics, where SSL_p demonstrates significantly lower calibration error across most metrics. In other words, the predicted probabilities align more closely with the true probabilities using the SSL_p model.

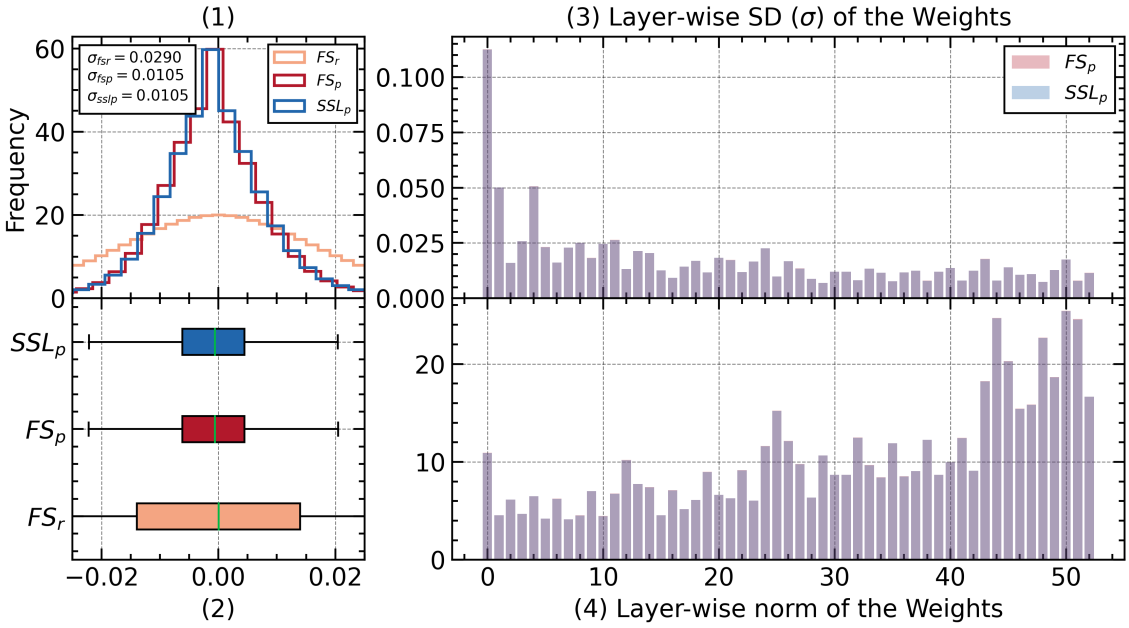
For Histopathology dataset, the weight distribution of the SSL_p model is similar to that of the FS_p, as seen in Figure 4.5b-(1),(2). This similarity in weight distribution could be attributed to an easier task, leading to higher test performance. However, despite the

similarity in weight distribution, the SSL_p model still provides better-calibrated outputs compared to the FS_p, but the difference in calibration error between these training regimes is now smaller. Considering the standard deviation of the weight distributions, it is suggested that a balance in the spread of weights is important for achieving good performance and calibration. It is important to note that the FS_r model has the highest standard deviation and comparable calibration error, it exhibits low AUC performance, making it inconsequential among other training regimes.

In Figure 4.5-(3),(4), we analyze the layer-wise standard deviation and Frobenius norm of the weights. In Figure 4.5a, we observe SSL_p influence on the standard deviation and weight magnitudes in every layer of the network. Additionally, we notice that the standard deviation tends to be higher in the initial layers and decreases as we move towards higher layers of the network. In Figure 4.5b, the standard deviation and magnitude of weights are similar for both SSL_p and FS_p training regimes. This suggests that the features extracted by each layer of the network are similar, which could be attributed to the high performance achieved by both training regimes. Despite the similarity, the SSL_p training regime still produces a better-calibrated model than the FS_p, indicating the additional benefits of self-supervised training.



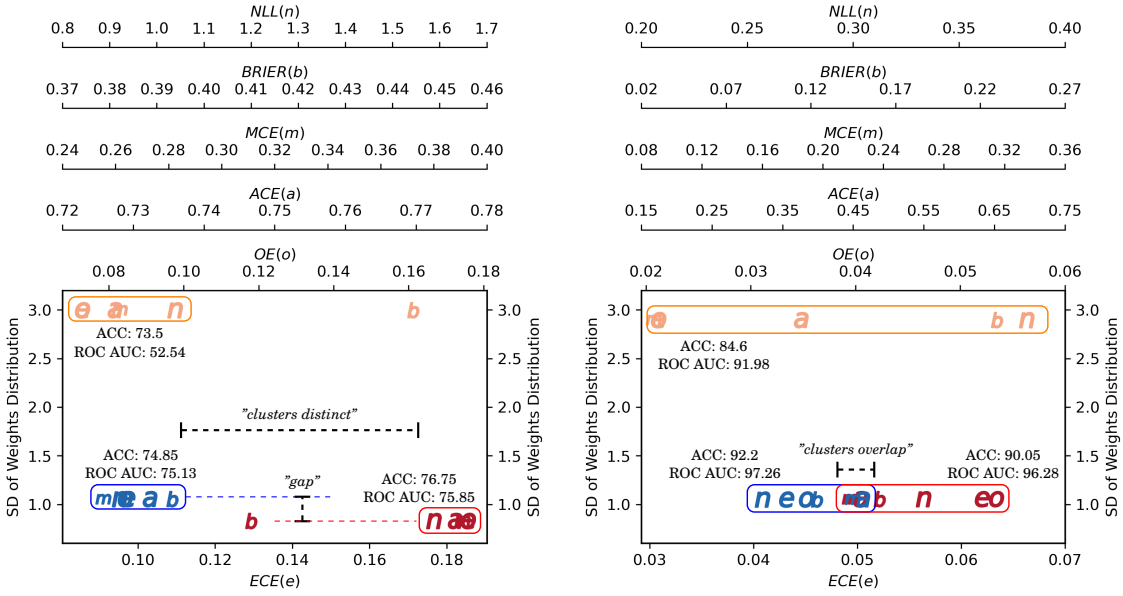
(a)



(b)

Figure 4.5: **Comparing different aspects of WideResNet learned weights for dataset size 10000 on DR-(a) and Histopathology Cancer-(b) datasets.** (1) and (2) the normalized histogram of weights of three training regimes. (3) Layer-wise comparison of standard deviation (SD) between FS_p and SSL_p. (4) Layer-wise comparison of Frobenius norm between FS_p and SSL_p.

For a more comprehensive analysis, Figure 4.6 further consolidates the trends between performance, the standard deviation of the weights, and model calibration. The figure highlights that achieving good performance and calibration in a model necessitates finding a balance in the spread of weights, a balance which the SSL_p training regime was able to achieve successfully. Due to the different scales of the calibration metrics, we plot them on multiple axes. The weight values and their standard deviation are very small; therefore, we scaled them by 10^2. In Figure 4.6a, FS_r (top left, orange) has the highest standard deviation (wide distribution) and gives us the best calibration error (x-axis) but the worst performance compared to other training regimes. The standard deviation for FS_p (bottom right, red) is the lowest, but the calibration error is still high, which is not ideal. On the other hand, SSL_p has a low standard deviation but yields the best performance and calibration. So, when we encounter the gap in the standard deviation of weights between different training regimes (SSL_p and FS_p), we observe the calibration error metrics are well separated (Figure 4.6a). Alternatively, when the gap is negligible, the calibration error metrics overlap (Figure 4.6b). In summary, we observe that the $SSLp$ training regime consistently provides better calibration than the FS_p regime for both datasets. The magnitude of improvement or change in calibration is directly related to the differences in weight distributions.



(a) (b)

Figure 4.6: **Comparing calibration metrics (x-axis) vs. standard deviation (SD, y-axis)** of WideResNet architecture for dataset size 10000 on DR and Histopathology cancer datasets. Colors represent training regimes (orange for FS_r, blue for SSL_p, and red for FS_p), and markers are the lowercase initials of each calibration metric; e – ECE, o – OE, a – ACE, m – MCE, b – Brier, n – NLL. Alongside each calibration error cluster, the performance is also reported. Ideally, the metrics should be at the bottom left with comparable performance. **(a)** SSL_p has less calibration error with on-par performance than FS_p training regime, indicating it to be a suitable choice. Calibration error metrics clusters of SSL_p and FS_p are noticeably well separated, correlating with the gap in their SD. **(b)** Here, SSL_p seems to be the best in calibration and performance compared to other training regimes. The noticeable difference we observed here is that the calibration error metrics clusters of SSL_p and FS_p are close (somewhat overlapping) when the SD of their weight distributions are similar.

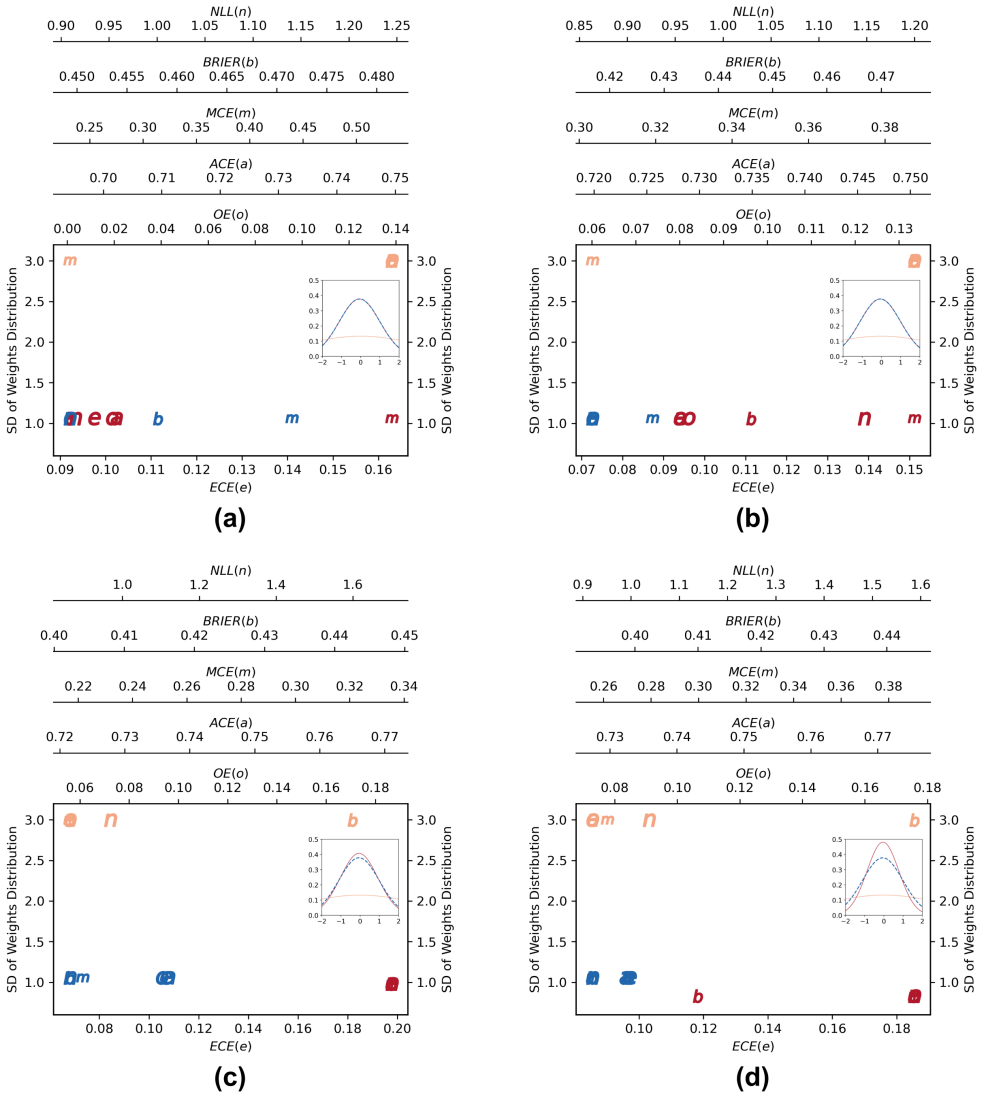


Figure 4.7: **Standard Deviation of Weights distribution vs. Calibration scores analysis.** (a), (b), (c), and (d) depict the relationship between the SD of weights distribution and calibration metrics from the smallest dataset size to the largest one (500, 1000, 1000, 10000), respectively of the DR dataset. Additionally, the corresponding weight distribution plots have been overlaid for convenience of reference. Considering the four plots, we can observe the trend that the calibration metrics of different regimes are segregated when there is a difference in the spread of their distributions (as shown in plots c & d) and overlapping when there is no difference in the SD of weights distribution (as shown in plots a & b). Based on the characteristics of SSL_p (shown in blue), it can be remarked that a balance in the spread of weights is necessary to achieve both good performance and calibration.

Learned Representation

In addition to the diversity of the whole weight space, we explore the impact of layer-wise, learned neural representations on performance and calibration. Towards this end, we use the widely popular Centered Kernel Alignment (CKA) [182] metric that measures the similarity between the activations of hidden layers in a neural network. Literature suggests that high representational similarity across layers indicates redundancy in learned representations of a network. Furthermore, redundant representations impact the generalizability due to the influence of regularized training [183], which in turn improves the model calibration [37]. CKA analysis of WideResNet's layer representations for different training regimes on the DR dataset is shown in Figure 4.8. The CKA plots for FS_p and SSL_p depict comparatively similar patterns. However, the higher layers of FS_p show a significant decrease in representational similarity (darker region shown in blue box) with increasing dataset size. The relatively high CKA values of the deeper layers of SSL_p depict redundancy of learned representations lighter regions) that provides implicit regularization. This in turn explains the reduced calibration error of SSL_p compared to FS_p as seen in Figure 4.2. A similar pattern is observed for ResNet18 and ResNet50 architectures as depicted in Figure 4.9. For the Histopathology dataset, the CKA plots in Figure 4.10 for FS_p and SSL_p show very similar patterns that explain comparable performance and calibration afforded by these training regimes.

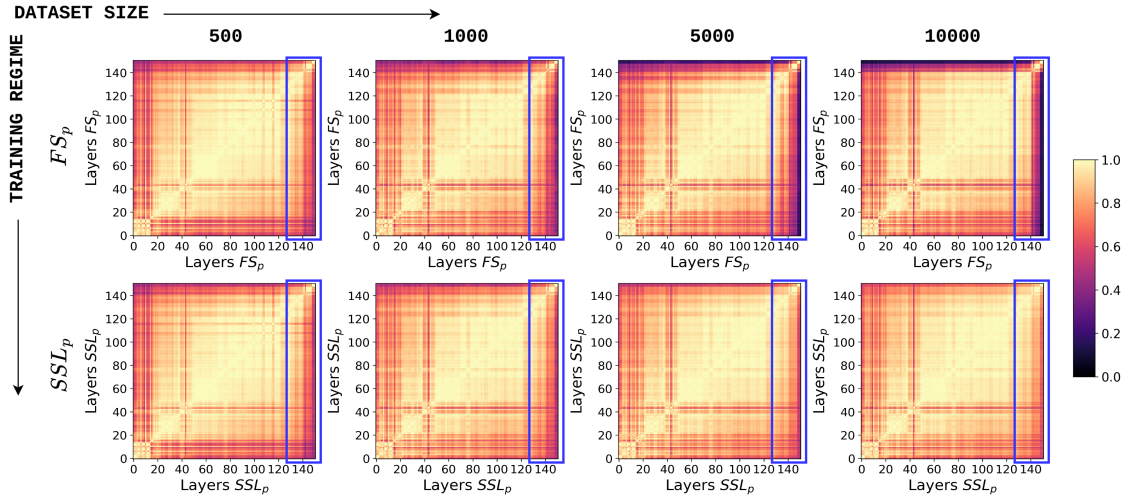


Figure 4.8: **CKA plots of trained WideResNet architecture using fully-supervised (pretrained, FS_p) and self-supervised (pretrained, SSL_p) regime for DR dataset.** The plots represents similarity between representations of features. The range of the CKA metric is between 0 and 1, with 0 indicating two completely distinct activations (not similar) and 1 indicating two identical activations (similar).

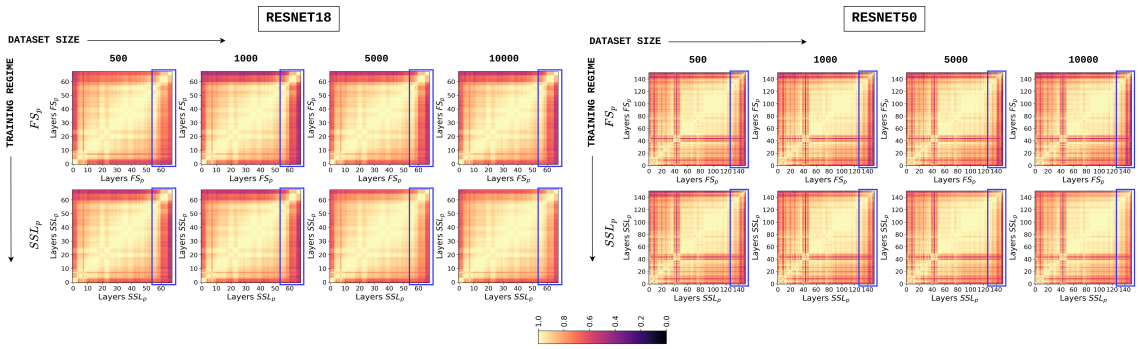


Figure 4.9: CKA plots of trained ResNet18 and ResNet50 architectures using FS_r, FS_p, and SSL_p regimes for DR dataset.

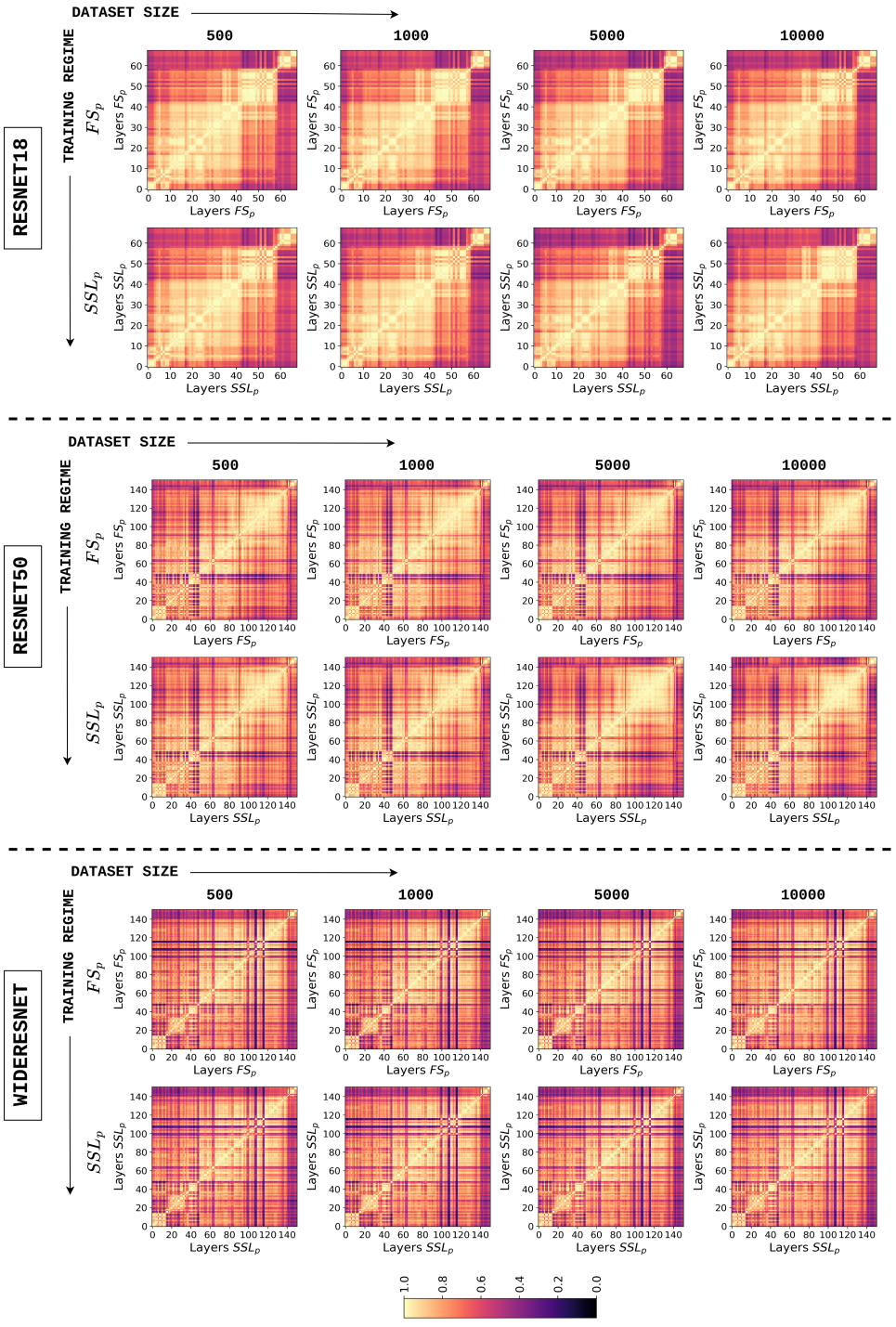


Figure 4.10: CKA plots of trained architectures using different regimes for Histopathology Cancer dataset.

Quantitative Comparison CKA

Table 4.2 presents the quantitative results of the CKA analysis, using mean CKA values. These findings align with the trends observed in Figure 4.8. In the case of the DR dataset, the mean CKA values of SSL_p rise as the dataset size increases. This supports our previous findings, where the calibration of SSL_p is superior to that of FS_p, and this distinction grows more pronounced as the dataset size becomes larger (Figure 4.6a). In the context of the Histopathology dataset, previous observations also indicated that SSL_p outperforms FS_p in terms of calibration, although the difference in calibration metrics values' magnitude is less (4.6b). Consequently, we notice that there is no significant difference in the mean CKA values between the two training approaches indicating the representations learned are quite similar.

Table 4.2: Mean CKA values of different training regimes across varying architectures, datasets and their sizes.

Architecture	Training Regime	Diabetic Retinopathy				Histopathology Cancer			
		500	1000	5000	10000	500	1000	5000	10000
ResNet18	FS_p	0.86	0.84	0.85	0.85	0.76	0.76	0.75	0.75
	SSL_p	0.85	0.85	0.88	0.88	0.76	0.75	0.74	0.75
ResNet50	FS_p	0.84	0.84	0.84	0.84	0.74	0.75	0.74	0.73
	SSL_p	0.85	0.85	0.86	0.87	0.75	0.74	0.74	0.73
WideResNet	FS_p	0.84	0.83	0.81	0.81	0.70	0.69	0.69	0.71
	SSL_p	0.84	0.85	0.86	0.87	0.69	0.70	0.69	0.71

4.3.5 RadImageNet Pretraining

To investigate the effect of domain-specific transfer learning, we conducted experiments using RadImageNet [66] a pretrained neural network (ResNet50) trained only on medical imaging datasets shown in Figure 4.11. Overall, we notice consistent patterns in calibration, where SSL_p either outperforms or matches FS_p, in line with our observations from other experiments. In this context, we observe that FS_p and SSL_p exhibit comparable performance in (a) and (b). However, in the MCE plot (e), SSL_p demonstrates superior calibration compared to FS_p. For the remaining metrics, SSL_p tends to show marginal improvement or comparable calibration. Taken together, these findings provide additional evidence that SSL_p consistently delivers calibration models on par with, or sometimes even superior to, those produced by FS_p.

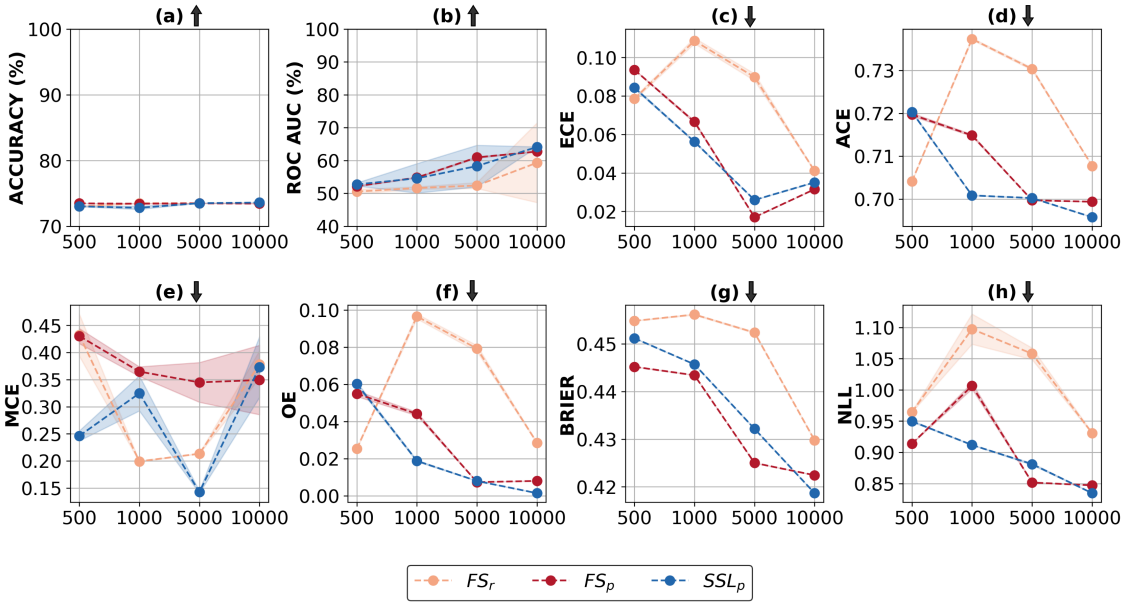


Figure 4.11: **Joint evaluation for performance and calibration across different dataset sizes (x-axis) of DR dataset using ResNet50 architecture with RadImageNet pretraining.** The shaded region corresponds to $\mu \pm \sigma$, estimated over 3 trials. ↑: higher is better, ↓: lower is better.

4.3.6 Comparison of Fully-Supervised and Reconstruction-Based Self-Supervised Task

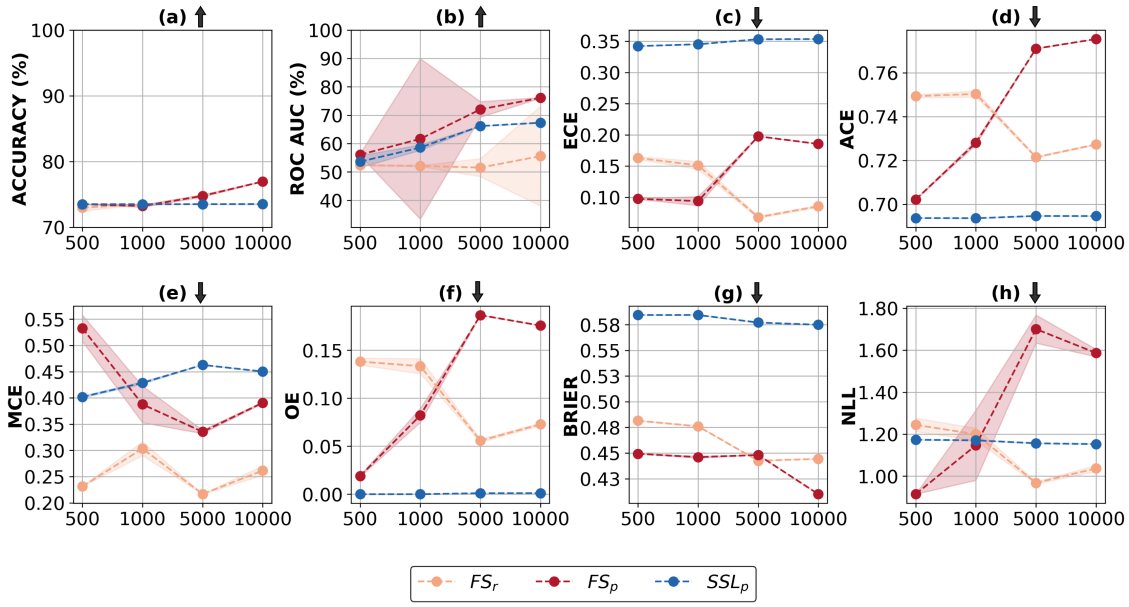


Figure 4.12: **Comparison of fully supervised (FS_r, random initialization), fully supervised (FS_p, pretraining), and reconstruction-based auxiliary SSL task (SSL_p, pretraining) on DR dataset.** Notably, the calibration of models achieved through the auxiliary task does not precisely align with that of the rotation task. Remarkably, the plots reveal a notable contrast: very low OE (f) but high ECE (c). This discrepancy could hint at potential underconfidence, stemming from substantial regularization induced by the reconstruction-based auxiliary SSL task. However, drawing definitive conclusions is premature, as further experiments, encompassing various architectures and hyperparameter tuning, are necessary. Relying solely on the plots, we abstain from making a judgment regarding the superiority of either FS_p or reconstruction-based SSL_p.

4.4 Discussion and Conclusion

For safety-critical applications like medical image analysis, it is imperative to choose models with high accuracy and low calibration errors. In this study, we investigate the performance and calibration of three different architectures using three different training regimes on medical imaging datasets of varying sizes and task complexities. Furthermore, we use six complementary calibration metrics that collectively provide a comprehensive evaluation of the predictive uncertainty of the models.

Model selection with mixed calibration results – While using multiple calibration metrics provides a more comprehensive evaluation, deciding on the best model can still be challenging as observed in Section 4.3.3. There are a few strategies that can be employed to aid in the decision-making process. One approach is to use a voting-based scheme, where each model is assigned a vote based on its performance across the calibration metrics. The model with the maximum number of votes is then selected as the best choice. This approach

treats all metrics equally and can be useful when there is no significant variation in the importance of different metrics.

Domain specific metric relevance – However, it is important to consider that different calibration metrics may have different objectives and importance in specific domains. For example, metrics like OE (Overconfidence Error) explicitly measure the overconfidence of the model predictions, while MCE (Maximum Calibration Error) provides an upper bound on the mistakes made by the model. In such cases, it might be necessary to assign more weightage to these important metrics during the voting process. The determination of metric importance is subjective and can vary depending on the application. Expert knowledge and domain expertise play a crucial role in assigning relative importance to different metrics. By incorporating the opinions of experts, the voting process can be tailored to reflect the specific requirements of the application.

Margin for model selection – In addition to assigning weights to metrics, introducing a margin or threshold in the voting scheme can help refine the model selection process. This threshold represents the minimum difference in calibration error between two training regimes that must be surpassed for a metric to be considered in the model selection. By setting a threshold, the metrics can be filtered out that do not exhibit significant differences and focus on those that have a substantial impact on model calibration.

It is worth noting that the difficulty of choosing a model also arises when one model has higher accuracy but poorer calibration while another model has lower accuracy but better calibration. This dilemma has been discussed in the literature [184], highlighting the need for careful consideration of calibration metrics during model selection. *Selective prediction* is one scenario where we abstain the classifier that gives us low-confident predictions based on some threshold or cost structure of the specific application [185]. In such cases, low-confidence predictions are referred to an expert for further analysis or diagnosis. This approach allows for cautious decision-making when the model's confidence is not sufficient for reliable predictions. Overall, the selection of the best model with mixed calibration results requires a combination of objective evaluation, subjective judgment of metric importance, and consideration of domain-specific requirements.

Calibration Metrics – While we have elaborated on the drawbacks of ECE, it provides an intuitive and straightforward interpretation, is simple to implement, and captures pure calibration. Additionally, ECE is associated with the reliability diagram - a powerful tool to visualize model calibration. It's also worth noting that alternative calibration metrics have their own shortcomings. The majority of the existing metrics suffer from challenges like scale-dependent interpretation, lack of normalized range, arbitrary choice of number of bins, etc. [62]. Moreover, composite measures like NLL and Brier blend calibration and refinement, making it challenging to isolate calibration effects. Multiclass settings introduce additional complexity due to the multitude of classes, their diverse interrelations, and the absence of a universally accepted metric for gauging refinement. Moreover, the choice of calibration metric can also be domain or application-dependent. As

there is no universally applicable or acceptable calibration metric, we proposed collective evaluation of these metrics for a better or unbiased understanding of calibration performance.

Limitations – Our current study focused on medical image classification tasks across three different benchmark datasets. However, due to limited computational resources, we selected datasets with 2D images. Extending this work to 3D datasets as well as other tasks like medical image segmentation and registration, can help broaden our understanding of calibration in the general context of medical image analysis. Additionally, our study highlights that using the rotation-based self-supervised learning (SSL) approach gives better-calibrated results compared to the usual fully-supervised learning. A comparison of other SSL techniques, such as contrastive SSL or generative SSL, would be interesting.

Conclusion – In general, for medical image classification tasks, we observe that training regimes have a varying impact on model calibration. Overall, we observe that across different architectures, training regimes, datasets, and sample sizes, (a) transfer learning through pretraining helps improve performance over random-initialized models and (b) pretrained self-supervised approach provides better calibration than its fully supervised counterpart, with on-par or better performance. While we notice a sizeable increase in performance with dataset sizes, only nominal improvement is realized with increasing model capacity.
Furthermore, we identified weight distribution and learned representations of a neural network as potential confounding factors that provide useful insights into model calibration, in particular, to explain the superiority of a rotation-based self-supervised training regime over fully supervised training.

Broader Impact – We anticipate that this analysis will offer significant insights into calibration across datasets of varying sizes and models of different complexities. This work raises a broader question regarding the search for a unified metric that can provide a comprehensive understanding of model calibration, thereby reducing the need to evaluate models based on multiple criteria. Ensuring accurate and reliable probabilistic predictions is vital for effective risk management and decision-making. It is particularly important when relying on the outputs of probabilistic models that require trust. Additionally, developing well-calibrated models is essential for promoting the widespread acceptance of machine learning methods, especially in fields like AI-driven medical diagnosis, as it directly influences the level of trust in new technologies and improves their explainability.

5

LS+: Informed Label Smoothing for Improving Calibration in Medical Image Classification

5.1 Introduction

Deep neural networks (DNNs) have demonstrated outstanding performance across various medical image tasks, including classification, segmentation, and detection [3]. However, modern DNNs are prone to miscalibration, compromising the reliability and trustworthiness of their predictions – critical factors in healthcare applications [35]. Therefore, addressing the issue of miscalibration and enhancing model calibration is of utmost importance.
Various approaches including data augmentation [145], ensemble [25], label smoothing [74, 75], focal loss [79], entropy-based regularization and feedback calibration during training [186, 38], have been proposed to mitigate DNN miscalibration. While some of these approaches involve varying the inputs to the DNN [145], others focus on changing the true label distribution [134, 80]. Studies have demonstrated the effectiveness of smoothing true labels during training for improving calibration [74]. Probabilities from DNNs serve as confidence indicators for predictions; High probabilities signify stronger belief in a predicted class, crucial in fields like medical diagnosis. However, interpreting the DNN results is incomplete without taking into account the model calibration [8, 187]. Calibration ensures that assigned probabilities accurately reflect the true likelihood of events. Without proper calibration, interpretations based solely on probabilities may be misleading or unreliable.
Contribution. Miscalibration [37] is defined as the disparity between the true confidence (accuracy) and the predicted confidence (output probability). Achieving perfect calibration entails bringing the predicted confidence score close to accuracy. To address this, we propose Label Smoothing plus (LS+) a novel and simple extension to label smoothing that substitutes the hard labels with informed smoothened versions computed from the validation set. The contributions of the paper are outlined as follows:

1. We introduce a simple yet effective approach to enhance model calibration by altering the true label distribution with a surrogate distribution computed from the class-wise accuracy on the validation set.

2. Our proposed method improves calibration with better or on par-performance when

compared to other popular approaches on three medical imaging datasets.

3. Using retention curves and density plots of correct and incorrect predictions, we observed that our method provides reliable and interpretable scores for model reject/second opinion, which is essential for safety-critical applications.

5.2 Related Work

Post-hoc calibration — Use a hold-out data set (calibration/validation set) to calibrate the confidence scores of a neural network. Several well-studied calibration methods include Platt scaling [60], isotonic regression [130], and temperature scaling (TS) [37]. Weight scaling [73] is an alternative version of TS for medical imaging tasks that explicitly optimizes the ECE measure to improve calibration. Additionally, class-distribution-aware vectors [133] for TS and label smoothing are used to address class-wise overconfidence. Meta-calibration [188] proposes differentiable ECE-driven calibration to obtain well-calibrated and highly accurate models.

Train-time calibration — An alternative approach that directly generates calibrated DNN models. Explicit confidence penalty (ECP) [76] leverages the entropy of the predicted distribution to regularize the loss function. Both Label smoothing (LS) [75, 74] and Focal loss (FL) [79] implicitly regulate the network output probabilities, encouraging their distribution to closely resemble the uniform distribution. Furthermore, auxiliary loss functions in conjunction with negative log-likelihood (NLL) are used to improve calibration. The difference between Confidence and Accuracy (DCA) [77] serves as an auxiliary loss, penalizing the model when the cross-entropy loss is reduced but the accuracy remains unchanged. Multi-class Difference in Confidence and Accuracy [78] broadens the scope of DCA by considering the calibration of every class, not solely the top-predicted class.

Our current work proposes a more informed strategy to enhance model calibration; the alignment of predicted probabilities (confidence) with accuracy is achieved by incorporating class specific priors derived from a separate validation set to account for current calibration level of the model.

5.3 Methodology

5.3.1 Preliminaries

Consider a multi-class classification problem comprising of K classes. Let $\hat{\mathbf{p}} = [\hat{p}_1, \ldots, \hat{p}_K]$ and $\mathbf{y} = [y_1, \ldots, y_K]$ be the predicted class distribution (confidence scores) of a deep neural network (DNN) and the ground truth one hot label encoding for an instance x respectively.

Calibration — A well-calibrated classifier generates confidence scores that align with the actual frequency of correct predictions. Formally, we can define calibration for a perfectly calibrated model for all classes as, $\mathbb{P}(y = y^*|\hat{\mathbf{p}}[y] = \hat{p}) = \hat{p}$, where, $y \in \text{argmax}_k\ y_k$, $y^* \in \{1, \cdots, K\}$, $\hat{\mathbf{p}}[y]$ is the confidence that sample x belongs to class y.[78]

Hard Labelling (HL) — DNN is conventionally trained using only the cross entropy (CE) loss defined as $CE(\mathbf{y}, \hat{\mathbf{p}}) = -\sum_k y_k \log \hat{p}_k$, which reduces to $\log \hat{p}_k$ if x is labeled k. Minimizing CE loss is equivalent to maximizing the log-likelihood of the correct label. Often, the optimization is continued until $\hat{p}_k$ is very close to y_k. As a result the DNN may suffer from over-fitting causing over confident predictions, leading to poor generalization and miscalibration.

Label Smoothing (LS) — An approach to mitigate miscalibration is to replace the one-hot encoded (*hard*) label vector with a smoothened (*soft*) label vector $\mathbf{y}' = (1-\alpha)\mathbf{y} + \alpha\mathbf{u}$, where $\mathbf{u}$ is a fixed distribution (typically uniform). Thus, label smoothing strategy involves minimizing $\mathcal{L}_{LS}$ defined as

$$\mathcal{L}_{LS} = H(\mathbf{y}', \hat{\mathbf{p}}) = -\sum_{k=1}^{K} y'_k \log \hat{p}_k = (1-\alpha)\ CE(\mathbf{y}, \hat{\mathbf{p}}) + \alpha\ CE(\mathbf{u}, \hat{\mathbf{p}}) \tag{5.1}$$

As the $CE(\mathbf{u}, \hat{\mathbf{p}})$ term penalizes the deviation between prediction ($\hat{\mathbf{p}}$) and prior ($\mathbf{u}$) distributions, it can be expressed using Kullback-Leibler (KL) divergence: $CE(\mathbf{u}, \hat{\mathbf{p}}) = D_{KL}(\mathbf{u}, \hat{\mathbf{p}}) + H(\mathbf{u})$. As $H(\mathbf{u})$, the entropy of $\mathbf{u}$, is a constant, the label smoothing cost function simplifies to [75]:

$$\mathcal{L}_{LS} = (1-\alpha)\ CE(\mathbf{y}, \hat{\mathbf{p}}) + \alpha\ D_{KL}(\mathbf{u}, \hat{\mathbf{p}}) \tag{5.2}$$

5.3.2 Label Smoothing Plus (LS+)

There are two drawbacks with vanilla label smoothing. Firstly, the approach does not take into account the DNN's current calibration level. As a result, forcible application of label smoothing to an already well-calibrated DNN may worsen its calibration. Secondly, the uniform prior does not take into account class-wise calibration levels (poorly and well-calibrated classes are treated alike). We propose Label Smoothing Plus (LS+) that addresses these two drawbacks in one go. LS+ replaces the uniform prior $\mathbf{u}$, with an informed class specific prior $\mathbf{v}^k = [v_1^k, \ldots, v_K^k]$ for $k = \{1, \ldots, K\}$, that is estimated on a separate validation set. In particular, the element v_j^k in the informed prior $\mathbf{v}^k$ for class k is defined as

$$v_j^k = \begin{cases} \mathcal{V}_k^{acc} & \text{if } j == k \\ (1 - \mathcal{V}_k^{acc}) \cdot \frac{1}{K-1} & \text{otherwise} \end{cases} \tag{5.3}$$

where, $\mathcal{V}_k^{acc}$ is validation set accuracy for class k using the pretrained (without label smoothing) model $\mathcal{M}$. *Example (Illustrated in Figure 5.1): For a pre-trained, four-class classification model with 76% validation accuracy for a specific class creates a label vector* $[0.76, 0.08, 0.08, 0.08]$*, which coerces the model to generate class prediction probabilities to match the validation accuracy.*

Learning the priors on the validation set ensures unbiased estimates and takes into account the current model calibration status. Furthermore, the smoothening of the prior is also dependent on the class accuracy. Priors of classes that are already accurately predicted by

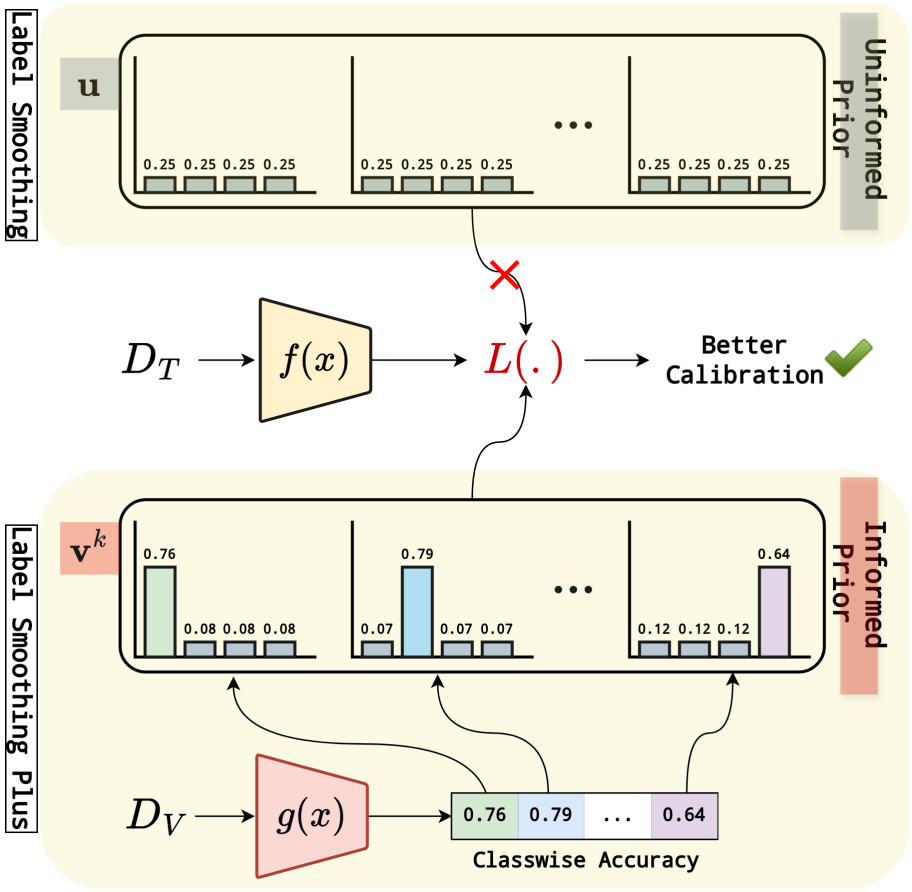


Figure 5.1: $\mathbf{D_T}$: Training Data, $\mathbf{D_V}$: Validation Data, $\mathbf{g(x)}$: Pre-trained model using Hard Labels, $\mathbf{f(x)}$: Model to be calibrated, $\mathbf{v^k}$: class-specific prior, $\mathbf{u}$: uniform prior. *(Top)* LS uses the same prior for all classes. *(Bottom)* LS+ uses class-specific priors computed from the validation set's class-wise accuracy based on the pre-trained model.

the model are smoothened to lesser extent than those of classes that are not accurately predicted. During training, the informed prior $\mathbf{v}^k$ corresponding to the ground truth class label for the instance x is used in place for a fixed uniform prior $\mathbf{u}$. In theory, $\mathbf{v}^k$ may be computed periodically after every few training iterations. However, we compute it only once before LS+ is applied. The complete pseudo-code for LS+ is presented in Algorithm 1.

5.4 Experiments and Results

5.4.1 Experimental Setting

Datasets — We evaluate LS+ using three benchmark datasets curated for medical image classification: (i) *Chaoyang - Histopathological dataset* [189] consists of colon slides with a patch size of 512×512. It is a multiclass ($K = 4$) dataset that is divided into training and testing sets consisting of 4021 and 2139 images respectively. Furthermore, we partitioned the training set into train (90%) and validation (10%). (ii) A *Minimalist Histopathology Image Analysis (MHIST) dataset* [190] comprises of 3,152 histopathological images of colorectal

Algorithm 1 Pseudocode of LS+

Input: A training dataset $\mathcal{D}_T = \{(x_i, y_i)\}_{i=1,\cdots,N}$, a validation dataset $\mathcal{D}_\mathcal{V}$, number of classes K, number of training epochs T, pre-trained model $\mathcal{M}$
Class-wise accuracy vector: $\mathcal{V}^{acc} = \mathcal{M}(\mathcal{D}_\mathcal{V})$, where $\mathcal{V}^{acc} \in \mathbb{R}^K$ and each component of the vector corresponds to the accuracy associated with the class $k \in K$
Compute new, class-specific label distribution set $\{\mathbf{v}^1, \ldots, \mathbf{v}^K\}$ *using Eqn (5.3)*
Minimize $\mathcal{L}_{LS+}$ over training data using the new distribution computed from $\mathcal{D}_\mathcal{V}$
for $t = 0$ **to** $T - 1$ **do**
    For each training instance i that belongs to class k, choose the corresponding informed prior $\mathbf{v}^k$
    $\mathcal{L}_{LS+} = (1 - \alpha) \cdot CE(\mathbf{y}, \hat{\mathbf{p}}) + \alpha \cdot D_{KL}(\mathbf{v}^k, \hat{\mathbf{p}})$
end for

polyps. It is a binary class ($K = 2$) dataset with images of size 224×224. The training and test sets consist of 2175 and 977 samples, respectively. Here, we partitioned the training dataset into train (80%) and validation (20%). (iii) *International Skin Imaging Collaboration (ISIC - 2018)* [191, 192] is a multi-class dataset ($K = 7$; highly imbalanced) of dermoscopic images of skin with a size of 600×450. It consists of separate train/validation/test sets with 10015/193/1512 samples, respectively. *The performance on the separate test set in all the three datasets facilitates an unbiased evaluation of LS+ and other approaches.*

Network Architectures and Implementation Details — We used two popular image classification architectures: ResNet-34 and ResNet-50, implemented using Tensorflow 2.4. These models are ImageNet pretrained and were specifically chosen for their effectiveness on small biomedical datasets [165, 164]. During training, all images are resized to 224×224 dimension. We used Adam optimizer with a learning rate set to $1e-3$, batch size of 8 and standard data augmentation techniques [193].

Baseline Methods — We compare LS+ with the following models: (a) Conventional classification using cross-entropy loss with one-hot encoded labels (Hard Labels), (b) cross-entropy loss with label smoothing (LS) [75], (c) focal loss ($\gamma = 3$) (FL) [79] that provides implicit regularization and two auxiliary loss methods - (d) difference between confidence and accuracy (DCA) [77], and (e) multi-class difference in confidence and accuracy (MDCA) [78].

Evaluation Metrics — We use several metrics to evaluate the models. Performance of the models is measured using accuracy (ACC), area under receiver operating characteristic (AUROC), precision, recall, F1-score. Similarly, a comprehensive comparison of calibration is achieved using expectation calibration error (ECE), adaptive calibration error (ACE), static calibration error (SCE), cross-entropy error (CE) and brier loss (Brier) [187].

5.4.2 Calibration performance comparison with SOTA

Table 5.1 provides a quantitative comparison of our method with SOTA approaches on Chaoyang, MHIST and ISIC-2018 datasets, respectively. These results demonstrate that validation accuracy-based label smoothing provides significant and consistent reduction across all calibration error metrics. Remarkably, this improvement in calibration was

achieved without compromising performance. In fact, marginal improvement can be observed in majority of the performance metrics across different architectures and datasets. Even for the highly imbalanced ISIC dataset, our model provides notable enhancement across all calibration metrics with minimal effect on performance, further solidifying the effectiveness of our approach.

Table 5.1: **Quantitative Results.** Performance and Calibration results on the test set of three benchmark datasets. The reported values are the average of 3 runs and given as percentages (%) with SD (σ) as subscript. ↑: Higher is better, ↓: Lower is better. Architectures: R34 *(ResNet-34)*, R50 *(ResNet-50)*; Datasets: D1 *(Chaoyang)*, D2 *(MHIST)* and D3 *(ISIC)*.

D1	Method	ACC ↑	AUROC ↑	Precision ↑	Recall ↑	F1 ↑	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓
R34	HL	$81.50_{1.2}$	$\mathbf{94.08}_{0.5}$	$75.91_{1.1}$	$74.58_{1.3}$	$75.10_{1.3}$	$11.33_{2.5}$	$11.21_{2.7}$	$06.26_{1.2}$	$74.21_{15.2}$	$29.74_{2.5}$
	LS [75]	$81.91_{0.4}$	$93.84_{0.6}$	$76.94_{0.9}$	$74.59_{0.8}$	$75.51_{0.7}$	$03.67_{0.9}$	$03.80_{0.8}$	$03.70_{0.2}$	$50.65_{2.8}$	$26.43_{1.1}$
	FL [79]	$81.89_{1.1}$	$94.07_{0.5}$	$76.58_{2.2}$	$\mathbf{75.68}_{1.1}$	$75.68_{1.6}$	$08.34_{6.2}$	$08.34_{6.1}$	$05.78_{3.3}$	$52.57_{4.2}$	$28.16_{2.3}$
	DCA [77]	$81.91_{0.7}$	$93.86_{0.3}$	$76.63_{0.6}$	$73.41_{1.3}$	$74.57_{1.1}$	$09.27_{1.4}$	$09.06_{1.6}$	$04.94_{0.8}$	$60.34_{3.8}$	$27.75_{1.3}$
	MDCA [78]	$81.52_{1.5}$	$93.13_{1.2}$	$76.55_{1.5}$	$74.91_{1.6}$	$75.45_{1.4}$	$10.72_{2.7}$	$10.58_{2.9}$	$05.99_{1.2}$	$81.92_{22.9}$	$29.42_{2.4}$
	Ours	$\mathbf{82.28}_{0.7}$	$94.02_{0.2}$	$\mathbf{77.30}_{1.3}$	$75.36_{1.4}$	$\mathbf{75.99}_{0.7}$	$\mathbf{02.81}_{0.7}$	$\mathbf{03.13}_{1.1}$	$\mathbf{03.49}_{0.4}$	$\mathbf{49.76}_{2.2}$	$\mathbf{25.66}_{0.9}$
R50	HL	$80.79_{0.5}$	$93.20_{0.5}$	$75.51_{0.4}$	$73.98_{0.3}$	$74.56_{0.2}$	$09.26_{4.5}$	$09.16_{4.6}$	$05.71_{1.6}$	$72.81_{22.7}$	$29.93_{2.3}$
	LS [75]	$80.62_{1.5}$	$93.04_{0.6}$	$75.41_{1.4}$	$73.92_{1.4}$	$74.48_{1.4}$	$03.51_{0.4}$	$04.27_{0.6}$	$\mathbf{03.66}_{0.3}$	$53.77_{2.2}$	$27.91_{1.4}$
	FL [79]	$80.52_{3.0}$	$93.47_{1.0}$	$76.30_{2.0}$	$72.60_{2.6}$	$73.73_{2.6}$	$04.16_{0.6}$	$04.26_{0.9}$	$04.06_{1.6}$	$53.98_{8.1}$	$27.70_{3.5}$
	DCA [77]	$79.82_{0.4}$	$92.75_{0.3}$	$74.82_{1.0}$	$72.32_{1.7}$	$73.18_{1.1}$	$13.89_{0.4}$	$13.88_{0.4}$	$07.43_{0.2}$	$92.42_{3.1}$	$33.30_{0.4}$
	MDCA [78]	$79.88_{2.0}$	$92.44_{0.6}$	$75.22_{2.5}$	$71.03_{3.2}$	$72.26_{3.0}$	$11.85_{4.3}$	$11.70_{4.5}$	$06.76_{2.0}$	$86.82_{27.9}$	$32.88_{4.2}$
	Ours	$\mathbf{81.44}_{1.7}$	$\mathbf{93.56}_{0.6}$	$\mathbf{76.39}_{2.0}$	$\mathbf{74.76}_{1.2}$	$\mathbf{75.20}_{1.9}$	$\mathbf{03.33}_{0.5}$	$\mathbf{03.45}_{0.7}$	$04.26_{0.8}$	$\mathbf{53.06}_{5.2}$	$\mathbf{27.17}_{2.7}$

D2	Method	ACC ↑	AUROC ↑	Precision ↑	Recall ↑	F1 ↑	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓
R34	HL	$77.14_{3.6}$	$84.32_{2.7}$	$78.27_{5.8}$	$73.63_{1.8}$	$74.14_{2.4}$	$17.40_{3.6}$	$17.24_{3.7}$	$17.98_{3.7}$	$101.40_{24.6}$	$38.97_{6.8}$
	LS [75]	$78.68_{1.5}$	$87.16_{1.0}$	$78.45_{3.4}$	$76.31_{2.2}$	$76.51_{1.4}$	$06.50_{1.6}$	$06.78_{1.4}$	$08.13_{2.2}$	$45.92_{2.4}$	$29.71_{1.2}$
	FL [79]	$80.32_{1.0}$	$87.10_{1.4}$	$80.13_{1.1}$	$76.63_{1.2}$	$77.70_{1.2}$	$12.01_{2.0}$	$12.11_{2.3}$	$11.76_{2.5}$	$48.25_{0.6}$	$31.55_{1.2}$
	DCA [77]	$77.83_{1.1}$	$85.83_{1.0}$	$77.10_{0.8}$	$73.94_{1.7}$	$74.86_{1.6}$	$08.51_{1.5}$	$08.46_{1.9}$	$09.01_{2.1}$	$51.02_{5.2}$	$31.45_{1.8}$
	MDCA [78]	$80.25_{1.7}$	$87.45_{1.4}$	$79.38_{1.9}$	$77.44_{2.2}$	$78.12_{2.0}$	$12.10_{2.4}$	$11.91_{2.2}$	$12.28_{2.5}$	$63.28_{11.6}$	$31.36_{0.8}$
	Ours	$\mathbf{81.48}_{0.9}$	$\mathbf{87.69}_{0.7}$	$\mathbf{80.77}_{1.3}$	$\mathbf{78.70}_{0.6}$	$\mathbf{79.47}_{0.8}$	$\mathbf{05.68}_{0.8}$	$\mathbf{06.42}_{0.8}$	$\mathbf{06.75}_{1.0}$	$\mathbf{44.78}_{0.8}$	$\mathbf{28.43}_{0.7}$
R50	HL	$77.21_{4.5}$	$83.63_{4.5}$	$75.71_{4.7}$	$74.22_{6.1}$	$74.65_{5.8}$	$12.17_{3.0}$	$11.89_{3.2}$	$12.33_{2.6}$	$61.24_{11.1}$	$34.38_{5.8}$
	LS [75]	$80.73_{1.4}$	$86.84_{1.5}$	$80.43_{2.6}$	$77.95_{0.5}$	$78.62_{0.7}$	$04.57_{1.9}$	$05.33_{1.6}$	$06.17_{1.4}$	$45.38_{3.5}$	$28.60_{2.2}$
	FL [79]	$77.25_{1.2}$	$84.01_{1.0}$	$77.60_{2.3}$	$72.81_{2.8}$	$73.65_{2.7}$	$10.61_{3.6}$	$10.74_{3.5}$	$11.93_{2.5}$	$51.62_{1.5}$	$34.07_{1.0}$
	DCA [77]	$79.40_{3.0}$	$85.50_{3.1}$	$78.97_{3.2}$	$75.61_{3.7}$	$76.61_{3.7}$	$07.99_{1.1}$	$08.08_{1.1}$	$09.14_{1.0}$	$60.43_{11.5}$	$30.94_{4.1}$
	MDCA [78]	$77.62_{2.4}$	$84.22_{2.8}$	$76.22_{2.7}$	$74.97_{2.3}$	$75.44_{2.4}$	$09.91_{3.0}$	$09.82_{3.3}$	$10.05_{3.0}$	$60.54_{10.5}$	$33.08_{4.0}$
	Ours	$\mathbf{81.45}_{0.9}$	$\mathbf{88.29}_{0.4}$	$\mathbf{80.57}_{1.2}$	$\mathbf{79.04}_{1.2}$	$\mathbf{79.60}_{1.0}$	$\mathbf{04.03}_{0.6}$	$\mathbf{04.27}_{0.7}$	$\mathbf{05.80}_{1.2}$	$\mathbf{42.40}_{0.9}$	$\mathbf{26.87}_{0.5}$

D3	Method	ACC ↑	AUROC ↑	Precision ↑	Recall ↑	F1 ↑	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓
R34	HL	$74.25_{0.3}$	$\mathbf{92.37}_{0.8}$	$63.76_{2.9}$	$52.91_{0.9}$	$55.85_{1.1}$	$15.38_{4.7}$	$15.30_{4.7}$	$04.78_{1.2}$	$112.51_{30.2}$	$40.77_{3.1}$
	LS [75]	$73.19_{1.1}$	$88.61_{3.6}$	$63.69_{3.6}$	$49.94_{2.9}$	$52.69_{4.3}$	$08.84_{3.0}$	$09.52_{2.9}$	$03.37_{0.5}$	$85.98_{8.4}$	$39.45_{1.8}$
	FL [79]	$74.01_{1.8}$	$90.69_{1.4}$	$64.95_{4.4}$	$52.02_{3.9}$	$55.90_{3.9}$	$04.19_{2.1}$	$04.44_{2.4}$	$03.01_{0.7}$	$77.03_{5.4}$	$37.06_{2.3}$
	DCA [77]	$\mathbf{74.37}_{1.5}$	$91.13_{1.3}$	$\mathbf{65.64}_{3.8}$	$53.27_{5.0}$	$\mathbf{57.38}_{3.8}$	$12.42_{3.3}$	$12.25_{3.4}$	$04.08_{0.8}$	$90.67_{13.8}$	$38.70_{4.0}$
	MDCA [78]	$72.62_{1.5}$	$90.15_{2.9}$	$61.78_{4.4}$	$\mathbf{53.41}_{5.2}$	$55.74_{3.2}$	$13.44_{5.5}$	$13.45_{5.4}$	$04.56_{1.2}$	$100.40_{20.4}$	$41.70_{4.7}$
	Ours	$74.03_{0.8}$	$90.02_{0.4}$	$61.28_{5.9}$	$48.36_{2.4}$	$50.23_{1.6}$	$\mathbf{03.72}_{0.5}$	$\mathbf{03.36}_{0.6}$	$\mathbf{02.04}_{0.2}$	$\mathbf{76.85}_{0.7}$	$\mathbf{36.66}_{0.2}$
R50	HL	$72.84_{0.3}$	$\mathbf{89.14}_{0.8}$	$61.75_{1.6}$	$49.38_{1.7}$	$52.92_{2.1}$	$15.41_{3.1}$	$15.36_{3.0}$	$04.91_{0.9}$	$137.89_{18.8}$	$43.10_{2.2}$
	LS [75]	$73.28_{1.8}$	$88.24_{2.4}$	$60.08_{2.5}$	$49.56_{6.0}$	$51.41_{5.3}$	$\mathbf{05.31}_{0.2}$	$\mathbf{06.45}_{0.9}$	$02.80_{0.5}$	$85.35_{1.1}$	$38.44_{1.1}$
	FL [79]	$71.96_{1.8}$	$88.73_{2.2}$	$59.11_{1.5}$	$\mathbf{51.07}_{3.6}$	$53.48_{2.9}$	$06.54_{4.1}$	$06.74_{3.8}$	$03.28_{0.7}$	$105.18_{19.7}$	$41.11_{3.5}$
	DCA [77]	$72.67_{0.8}$	$88.77_{3.0}$	$61.35_{3.7}$	$48.97_{2.8}$	$52.12_{2.1}$	$14.02_{6.7}$	$13.81_{6.5}$	$04.38_{1.8}$	$112.82_{36.0}$	$42.49_{5.1}$
	MDCA [78]	$73.68_{1.0}$	$89.13_{2.1}$	$61.09_{3.6}$	$50.56_{2.8}$	$53.95_{2.9}$	$17.65_{6.2}$	$17.62_{6.2}$	$05.44_{1.5}$	$143.59_{43.3}$	$43.32_{4.8}$
	Ours	$\mathbf{73.77}_{0.9}$	$89.01_{1.5}$	$\mathbf{63.12}_{1.4}$	$51.04_{1.4}$	$\mathbf{55.03}_{1.5}$	$06.96_{0.6}$	$06.95_{1.0}$	$\mathbf{02.63}_{0.2}$	$\mathbf{83.61}_{3.6}$	$\mathbf{37.38}_{1.1}$

5.4.3 Uncertainty-based Retention Curves

To assess the reliability of the models, we plot the accuracy of a model as a function of its retention rate. As the fraction of predictions retained is increased, ground truth labels are replaced with predicted labels in decreasing order of prediction scores, providing a comprehensive view of error distribution across the dataset. *For a zero retention fraction, we opt for the predicted label vector ($\boldsymbol{\Omega}$) to be the same as the ground truth ($\boldsymbol{G}$), resulting in 100% accuracy. As we increase the retention fraction, we replace the label vector $\boldsymbol{\Omega}$ with the fraction of the original predicted labels from samples having the highest predicted probability. We continue the substitution process until the entire label vector is replaced with the predicted labels.* The area under this accuracy-retention curve (R-AUC) serves as a metric for evaluating the quality of uncertainty estimates (predicted confidence scores) [194], with a higher value indicating models with better predictions. Figure 5.2 exhibits superior reliability of our proposed validation accuracy based label smoothing model, making it more suitable for medical image analysis.

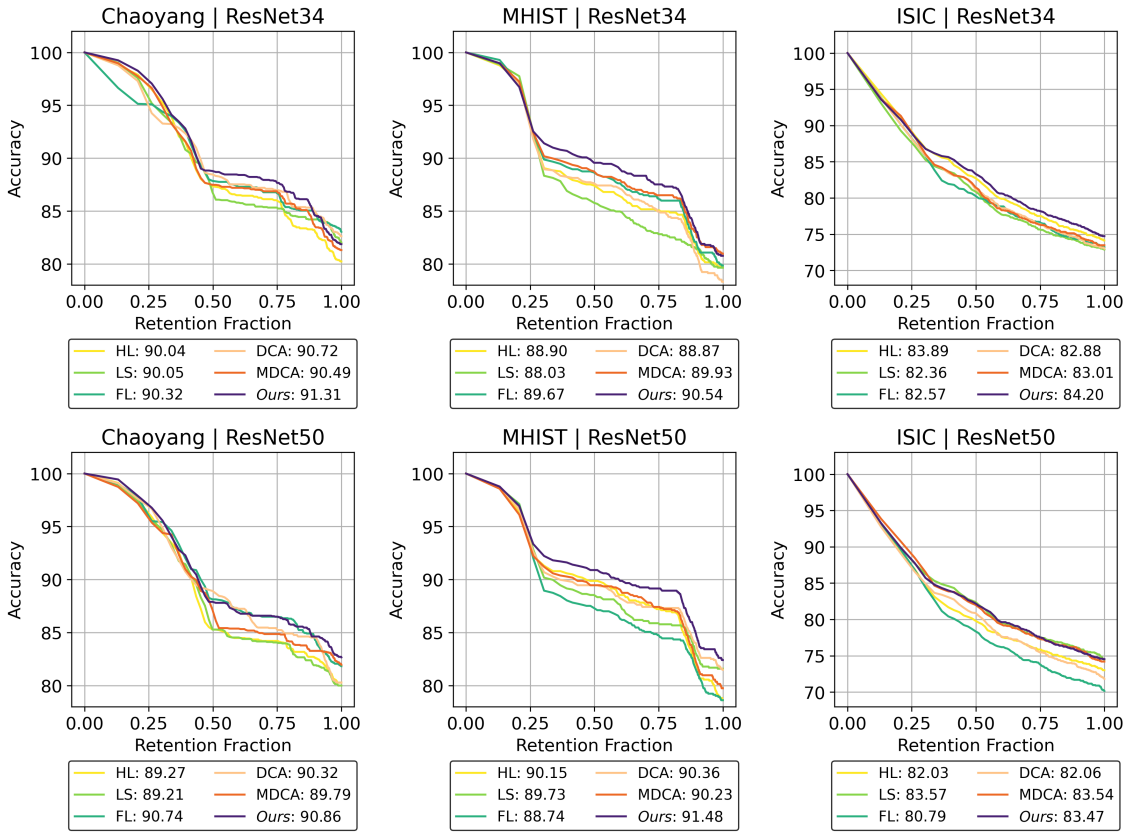


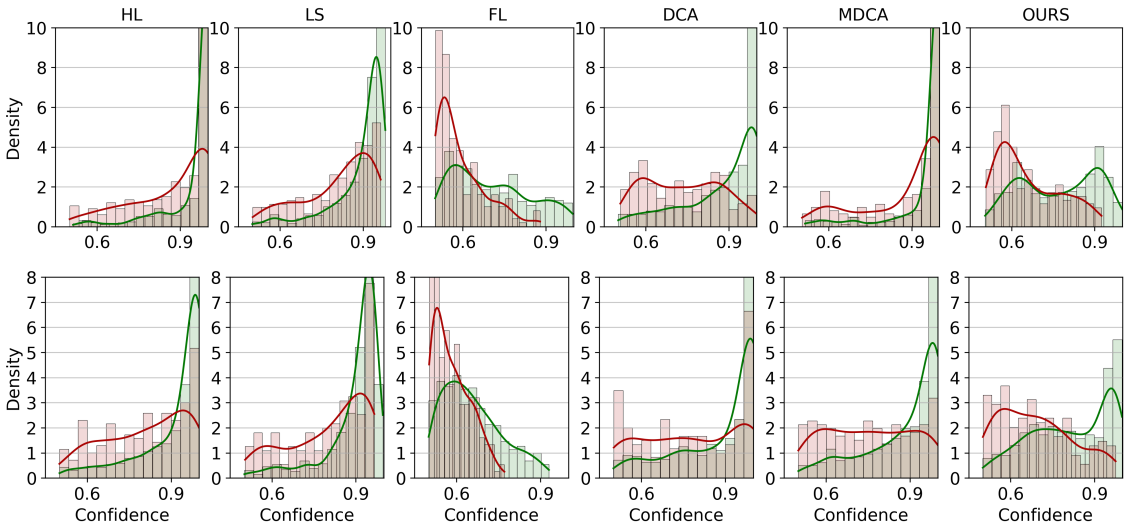
Figure 5.2: **Retention Curves.** Accuracy as a function of retention fraction along with the Area Under the Retention Curve (R-AUC) values using ResNet-34/50 for all three datasets. HL - *Hard Labels*, LS - *Label Smoothing*, FL - *Focal Loss*, DCA - *Difference between Confidence and Accuracy* and MDCA - *Multi-class Difference in Confidence and Accuracy*.

5.4.4 Clinical Significance of Predicted Confidence Scores

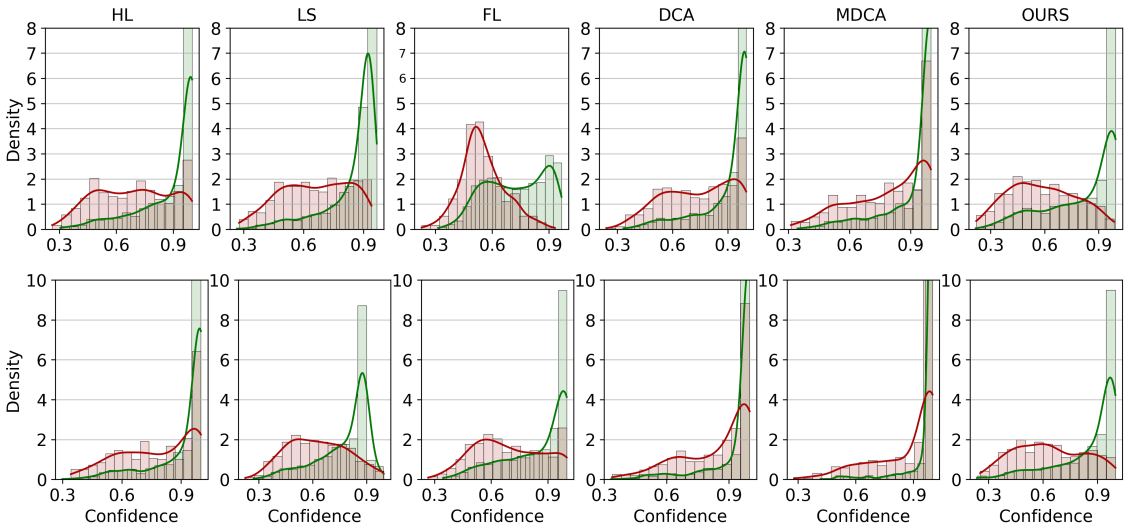
To gain deeper insights into model calibration, we distinguish between the confidence scores assigned to correct and incorrect classified samples in Figure 5.3. Ideally, the confidence scores of the correctly predicted samples should be close to 1 (indicating high certainty), while incorrect classified samples should move away (reflecting uncertainty). The density plots associated with the majority of SOTA approaches (except *FL*) exhibit left-skewed distributions for correct predictions (green), signifying high confidence levels of these models. Undesirably, these models also express high confidence in their incorrect predictions (red). *FL* exhibits contrasting behavior with right-skewed distributions for both correct and incorrect predictions indicating overall low confidence levels. Our proposed approach strikes the right balance by assigning relatively high scores for correctly classified samples while adeptly conveying uncertainty associated with incorrectly classified samples with low scores. This nuanced approach positions our model as a reliable and trustworthy solution that increases the likelihood of expert medical intervention when the model lacks confidence.

5.4.5 Discussion

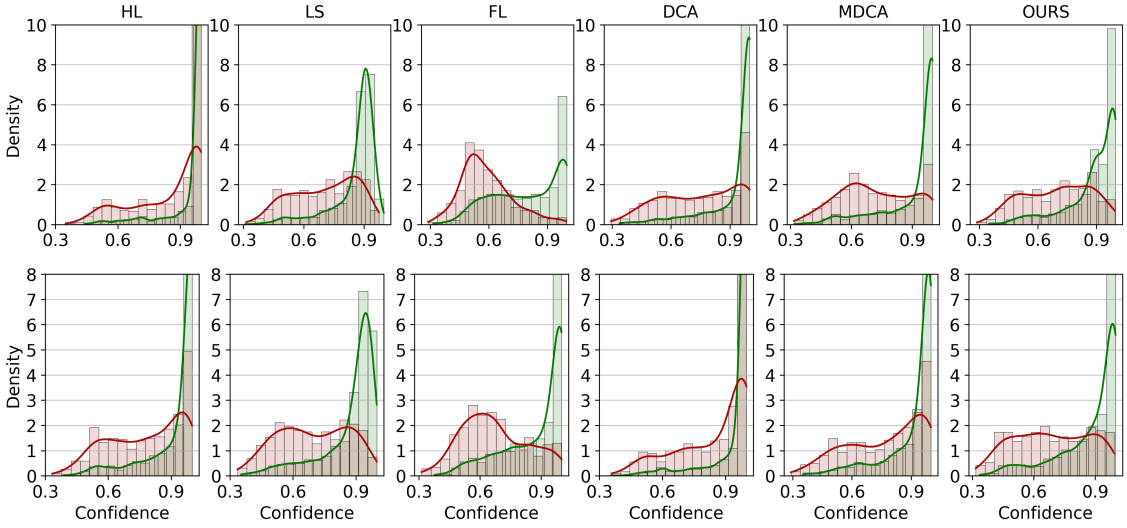
The proposed approach offers a simple and intuitive method for improving model calibration, making it accessible and easy to integrate into existing deep learning pipelines. By refining the confidence estimates of the model, it improves the alignment between predicted probabilities and actual outcomes, leading to more reliable decisions. However, while the method effectively improves both performance and calibration, it does not explicitly evaluate model robustness, particularly in detecting out-of-distribution (OOD) samples. This limitation is critical, as models deployed in real-world settings often encounter unseen data distributions, where a lack of robustness could lead to erroneous predictions with high confidence. Despite its simplicity, the framework remains effective for applications where interpretability and calibration are prioritized over complex architectural modifications. Its relevance is particularly pronounced in safety-critical domain such as medical image analysis, where accurate confidence estimation is crucial for ensuring that decisions are made with a well-calibrated level of certainty, reducing the risk of overconfident mispredictions.



(a) **MHIST Dataset.**



(b) **ISIC Dataset.**



(c) **Chaoyang Dataset.**

Figure 5.3: **Comparison of density plots for correct (green) and incorrect (red) classification confidences for ResNet-34 (top) and ResNet-50 (bottom) on MHIST, Chaoyang and ISIC datasets.** For incorrect predictions, LS, FL and Ours provide low confidence which is desirable. However, methods like HL, DCA and MDCA exhibits higher confidence even when they are wrong making them unreliable. The area under the histogram integrates to 1. We have clipped the y-axis in all the plots to better visualize the trends.

5.5 Ablation Studies

5.5.1 Comparison with Temperature Scaling

Table 5.2 shows the comparison of calibration metrics after applying temperature scaling, evaluated on the Chaoyang and MHIST datasets using ResNet-34 and ResNet-50 architectures. Notably, LS+ consistently achieves superior calibration performance. *NOTE: Temperature scaling is a post-hoc method; therefore, the performance metrics for each baseline remain unchanged, as reported in Table 5.1.*

Table 5.2: **Calibration comparison** ***(+Temperature Scaling)*** results on the test set of Chaoyang & MHIST datasets. The reported values are the average of 3 runs and given as percentages (%) with SD (σ) as subscript. ↑: Higher is better, ↓: Lower is better.

Chaoyang Dataset										
Method	ResNet-34					ResNet-50				
(+TS)	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓
HL	$09.47_{2.90}$	$09.47_{2.80}$	$05.53_{1.40}$	$63.51_{11.20}$	$28.67_{2.30}$	$07.20_{4.80}$	$07.09_{4.90}$	$05.14_{1.40}$	$64.36_{16.30}$	$29.12_{2.00}$
LS [75]	$04.05_{0.20}$	$04.53_{0.60}$	$\mathbf{03.30}_{0.10}$	$50.58_{2.70}$	$26.69_{1.00}$	$03.55_{0.50}$	$04.20_{0.60}$	$\mathbf{03.65}_{0.30}$	$53.72_{2.20}$	$27.87_{1.50}$
FL [79]	$04.09_{3.40}$	$04.23_{3.50}$	$04.52_{2.00}$	$50.26_{1.70}$	$26.95_{1.30}$	$03.01_{0.80}$	$03.19_{0.90}$	$03.95_{1.80}$	$53.30_{7.30}$	$27.52_{3.50}$
DCA [77]	$06.90_{1.60}$	$06.76_{1.80}$	$04.10_{0.70}$	$53.71_{2.20}$	$26.79_{1.10}$	$12.09_{0.30}$	$12.01_{0.40}$	$06.62_{0.20}$	$77.96_{3.00}$	$31.99_{0.30}$
MDCA [78]	$08.64_{3.40}$	$08.47_{3.30}$	$05.37_{1.30}$	$71.92_{17.80}$	$28.47_{2.10}$	$09.63_{4.60}$	$09.50_{4.70}$	$06.15_{1.80}$	$75.39_{21.40}$	$31.83_{3.80}$
Ours	$\mathbf{02.89}_{0.20}$	$\mathbf{02.97}_{0.50}$	$03.44_{0.40}$	$\mathbf{49.71}_{2.10}$	$\mathbf{25.65}_{0.80}$	$\mathbf{02.87}_{0.70}$	$\mathbf{03.00}_{0.50}$	$04.02_{0.90}$	$\mathbf{52.78}_{5.20}$	$\mathbf{27.06}_{2.70}$
MHIST Dataset										
Method	ResNet-34					ResNet-50				
(+TS)	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓
HL	$15.78_{3.60}$	$15.72_{3.70}$	$16.73_{3.90}$	$83.69_{19.70}$	$37.57_{6.60}$	$09.43_{3.1}$	$09.41_{3.1}$	$09.62_{2.8}$	$54.59_{9.1}$	$33.11_{5.5}$
LS [75]	$06.58_{0.80}$	$06.73_{1.00}$	$07.98_{1.80}$	$45.78_{2.10}$	$29.73_{1.30}$	$05.37_{1.4}$	$05.83_{1.4}$	$06.48_{1.0}$	$45.75_{4.4}$	$28.70_{2.5}$
FL [79]	$08.08_{0.70}$	$08.45_{0.30}$	$08.15_{0.60}$	$45.59_{2.20}$	$29.76_{2.20}$	$05.90_{3.3}$	$06.10_{3.2}$	$07.96_{3.0}$	$49.13_{0.6}$	$32.32_{0.3}$
DCA [77]	$06.94_{1.10}$	$06.56_{1.20}$	$\mathbf{07.59}_{1.60}$	$47.92_{3.20}$	$30.69_{1.50}$	$06.59_{1.6}$	$06.43_{1.4}$	$08.20_{1.6}$	$54.85_{9.5}$	$30.42_{4.0}$
MDCA [78]	$10.31_{3.10}$	$10.00_{2.80}$	$10.34_{3.20}$	$54.57_{7.40}$	$30.22_{0.80}$	$07.44_{3.4}$	$07.41_{3.2}$	$08.14_{2.4}$	$54.60_{7.7}$	$32.20_{3.7}$
Ours	$\mathbf{06.22}_{0.60}$	$\mathbf{05.96}_{1.30}$	$07.67_{0.70}$	$\mathbf{45.02}_{1.40}$	$\mathbf{28.34}_{0.60}$	$\mathbf{03.19}_{0.6}$	$\mathbf{03.69}_{0.8}$	$\mathbf{05.44}_{1.6}$	$\mathbf{42.28}_{0.7}$	$\mathbf{26.73}_{0.6}$

5.5.2 Standard training (HL) followed by fine-tuning with LS approach

In Table 5.3, we showed the results of the experiments with the HL-trained model and then fine-tuned using LS. We observed significantly degraded performance and calibration results compared to our approach on both the Chaoyang & MHIST datasets using ResNet-34 and ResNet-50 architectures.

5.5.3 Replacing validation set class-wise accuracy with constant value

In this study, we replaced V^{acc} with a constant value (0.4/0.6) for the correct class, distributing the remaining mass equally among the other classes to create the soft ground truth. The results, presented in Table 5.4, show a significant degradation in the model's calibration. Manually determining the optimal value is challenging and requires extensive trial and error, whereas LS+ automatically identifies the appropriate value using the validation set. Additionally, applying the same constant value to both poorly and well-calibrated classes can result in suboptimal calibration.

Table 5.3: **HL Trained Model $\xrightarrow{Fine-tune}$ LS.** Performance and Calibration results on the test set of two benchmark datasets. The reported values are the average of 3 runs and given as percentages (%) with SD (σ) as subscript. ↑: Higher is better, ↓: Lower is better. Architectures: R34 *(ResNet-34)*, R50 *(ResNet-50)*; Datasets: D1 *(Chaoyang)* and D2 *(MHIST)*.

Setup	ACC ↑	AUROC ↑	Precision ↑	Recall ↑	F1 ↑	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓
D1, LS, R34	$80.67_{1.6}$	$91.62_{0.7}$	$74.72_{2.2}$	$73.22_{2.4}$	$73.78_{2.3}$	$06.49_{1.2}$	$07.39_{0.9}$	$04.85_{0.2}$	$59.47_{3.3}$	$29.98_{2.2}$
D1, LS, R50	$79.60_{2.4}$	$91.76_{1.9}$	$76.08_{2.0}$	$73.67_{1.8}$	$74.23_{2.3}$	$06.23_{1.6}$	$06.07_{1.5}$	$05.06_{1.5}$	$60.56_{6.6}$	$30.56_{3.0}$
D2, LS, R34	$78.61_{1.1}$	$84.58_{1.1}$	$78.48_{2.8}$	$75.10_{1.1}$	$75.88_{0.4}$	$11.54_{0.5}$	$11.45_{0.3}$	$11.80_{0.6}$	$54.05_{1.6}$	$33.43_{1.2}$
D2, LS, R50	$81.34_{2.8}$	$87.01_{3.6}$	$81.82_{2.5}$	$77.34_{3.7}$	$78.58_{3.6}$	$06.06_{3.0}$	$06.23_{3.0}$	$07.47_{2.5}$	$45.19_{7.2}$	$28.37_{4.6}$

Table 5.4: **LS+ with constant values (0.4/0.6).** Performance and Calibration results on the test set of two benchmark datasets. The reported values are the average of 3 runs and given as percentages (%) with SD (σ) as subscript. ↑: Higher is better, ↓: Lower is better. Architectures: R34 *(ResNet-34)*, R50 *(ResNet-50)*; Datasets: D1 *(Chaoyang)* and D2 *(MHIST)*.

Method	ACC ↑	AUROC ↑	Precision ↑	Recall ↑	F1 ↑	ECE ↓	ACE ↓	SCE ↓	CE ↓	Brier ↓
D1, R34, 0.6	$80.42_{1.50}$	$92.28_{2.60}$	$74.32_{2.00}$	$73.80_{1.10}$	$73.94_{1.40}$	$10.56_{2.10}$	$10.84_{1.80}$	$07.81_{0.30}$	$62.02_{5.00}$	$30.63_{2.50}$
D1, R50, 0.6	$80.85_{1.90}$	$92.26_{2.50}$	$75.89_{3.10}$	$73.37_{1.20}$	$74.14_{1.70}$	$11.27_{4.70}$	$11.01_{4.90}$	$07.59_{1.20}$	$62.11_{5.10}$	$30.74_{2.70}$
D1, R34, 0.4	$80.54_{0.80}$	$93.52_{0.20}$	$75.40_{1.40}$	$74.12_{0.90}$	$74.52_{1.00}$	$19.97_{0.70}$	$19.87_{0.90}$	$12.28_{0.20}$	$69.04_{1.10}$	$33.96_{0.80}$
D1, R50, 0.4	$81.78_{0.50}$	$92.43_{1.60}$	$76.16_{0.50}$	$74.45_{1.00}$	$75.10_{0.80}$	$17.65_{0.90}$	$17.53_{0.90}$	$10.34_{0.30}$	$65.85_{2.30}$	$31.72_{1.00}$
D2, R34, 0.6	$77.35_{1.60}$	$84.43_{3.10}$	$76.26_{1.20}$	$74.12_{3.30}$	$74.68_{2.70}$	$70.28_{2.40}$	$07.86_{1.80}$	$08.71_{1.60}$	$49.62_{2.60}$	$32.12_{2.30}$
D2, R50, 0.6	$77.58_{3.40}$	$84.91_{4.00}$	$76.45_{4.20}$	$75.10_{4.50}$	$75.37_{4.00}$	$10.28_{1.80}$	$09.87_{2.20}$	$10.86_{2.20}$	$50.41_{3.00}$	$32.62_{2.70}$
D2, R34, 0.4	$80.93_{0.70}$	$86.75_{1.80}$	$80.14_{0.80}$	$78.08_{0.80}$	$78.84_{0.80}$	$15.20_{1.10}$	$15.63_{0.60}$	$15.48_{1.00}$	$51.29_{0.90}$	$32.82_{0.70}$
D2, R50, 0.4	$79.12_{3.30}$	$86.53_{2.40}$	$77.63_{3.40}$	$78.26_{3.00}$	$77.88_{3.30}$	$16.30_{1.30}$	$16.37_{1.20}$	$16.92_{1.10}$	$53.94_{2.50}$	$35.22_{2.40}$

5.6 Conclusion

We propose an informed label smoothing strategy (LS+) that addresses the shortcomings of the traditional version by taking into consideration the model's current calibration status as well as class-wise calibration levels. This is achieved by replacing the uniform prior with an informed class-specific prior estimated from the class accuracy on a separate validation set. Experimental results from three benchmark medical image classification tasks show that LS+ provides significant improvement in calibration. Consistent improvement across multiple performance and calibration metrics using two different architectures as well as higher R-AUC values along with density plots exhibit reliability and clinical readiness of LS+. Our present study assumes that both the validation and test sets stem from the same distribution. In medical imaging, heterogeneity of population, scanners and acquisition protocols presents a shift in distribution. Hence, our future efforts will be directed towards adapting LS+ to excel in out-of-distribution (OOD) scenarios.

6

Summary and Future Work

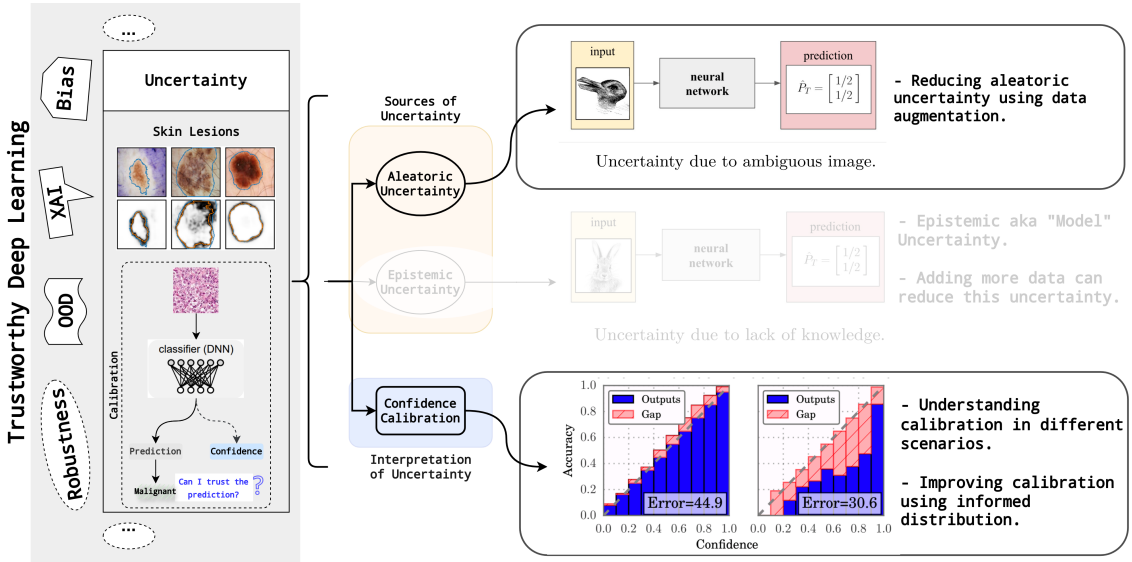


Figure 6.1: Illustration of Thesis Summary

6.1 Thesis Summary

Trustworthy Deep Learning involves several subdomains aimed at improving the reliability of DNNs in medical imaging. This thesis emphasizes trustworthiness by addressing two key aspects: uncertainty quantification and confidence calibration. Uncertainty quantification fosters trust by explicitly indicating the level of certainty in predictions. Confidence calibration, viewed as a form of uncertainty interpretation, further strengthens trust by ensuring that the model's predicted probabilities accurately reflect the likelihood of actual outcomes. The thesis summary is illustrated in Figure 6.1.

Uncertainty in predictions can be attributed to noise or randomness in data (aleatoric) and incorrect model inferences (epistemic). While model uncertainty can be reduced with more data or bigger models, aleatoric uncertainty is more intricate. In this work, we proposed a novel approach that interprets data uncertainty estimated from a self-supervised learning

(SSL) task as noise inherent to the data. Since we used a reconstruction task as the SSL task, where the input is reconstructed at the output, the nature of the task allows the uncertainty in the predictions to be interpreted as the noise present in the input. A heteroscedastic neural network (NN) was employed to model aleatoric uncertainty alongside the predictions, effectively learning a distribution over the output. From this learned data distribution, new data points were sampled, which can be considered a data augmentation process. The augmented data (additional features), was then used for training to reduce aleatoric uncertainty in the segmentation task. Evaluation on a Brain Tumor Segmentation (BraTS) dataset, using image reconstruction as the self-supervised task and segmentation as the image analysis task, demonstrated the effectiveness of the proposed approach. The results showed a significant reduction in aleatoric uncertainty for the image segmentation task while achieving performance that was either superior to or comparable with standard augmentation techniques.

Another key aspect of trustworthiness is the confidence scores of a DNN, which play a pivotal role in explainability by providing insights into the model's certainty, identifying cases that require attention, and establishing trust in its predictions. While there has been a significant effort towards training modern DNN to achieve high accuracy on medical imaging tasks, model calibration and factors that affect it remain under-explored. To address this, a comprehensive empirical study was conducted that explores model performance and calibration under different scenarios. Fully supervised training was considered, which is the prevailing approach in the community, as well as rotation-based self-supervised method with and without transfer learning, across various datasets and architecture sizes. Multiple calibration metrics were employed to gain a holistic understanding of model calibration. Our study revealed that:

a. Transfer learning helps improve performance and calibration over random-init models and smaller datasets.
b. Comparing self-supervised pretrained model is better calibrated than fully-supervised pretrained model, with on-par or better performance.
c. Dataset size has significant effect on calibration and performance. (Larger the better).
d. Increasing model capacity shows minimal improvement.
e. Factors such as weight distributions and the similarity of learned representations correlate with the calibration trends observed in the models.

This work shed light on the importance of model calibration in medical image analysis and highlights the advantages of incorporating self-supervised learning to improve both performance and calibration.

Understanding calibration in different scenarios is crucial, but improving it is just as important. While several train-time methods have been introduced to tackle this challenge, they often come at the cost of performance. Label smoothing, a widely used technique that uses soft targets during training, remains a popular strategy for improving calibration.

However, it fails to consider the existing calibration of the DNN and applies a uniform prior across all classes, which might not be ideal. To address this, Label Smoothing Plus (LS+) strategy was proposed which uses a class-specific prior estimated from the validation set to account for the model's calibration level. The effectiveness of our approach was evaluated by comparing it to state-of-the-art methods on various benchmark medical imaging datasets, different architectures, and several performance and calibration metrics for the classification task. Experimental results showed a notable reduction in calibration error metrics with a nominal improvement in performance compared to other approaches. Uncertainty-based retention curves and the analysis of confidence levels for correctly and incorrectly classified examples revealed that the predicted probabilities from LS+ are better suited for clinical decision-making, suggesting better reliability of the proposed method.

6.2 Impact and Applications

Uncertainty quantification and model calibration play a crucial role in ensuring the reliability of AI-driven medical applications. Below are some real-life scenarios illustrating their impact on medical interpretation and treatment refinement.

`Aleatoric Uncertainty Estimation`

- **Tumor Segmentation for Radiation Therapy Planning**
  When a radiologist relies on an AI model to delineate tumor boundaries on an MRI scan for radiation therapy planning, the model may identify regions of high aleatoric uncertainty. High aleatoric uncertainty indicates that the image data itself is ambiguous - perhaps due to motion artifacts, low contrast, or overlapping tissue structures. The radiologist can then cross-reference these areas with additional imaging modalities, such as PET scans or contrast-enhanced MRI, to improve the delineation accuracy. This reduces the risk of either under-treating the tumor or irradiating healthy tissues unnecessarily.

- **AI-Assisted Pathology in Cancer Diagnosis**
  When a deep learning model is used to assist pathologists in identifying cancerous cells in histopathology slides, it may highlight areas of high uncertainty in distinguishing between low-grade and high-grade cancerous cells. It suggests that the sample has ambiguous or borderline features. In this case, the pathologist might order additional staining techniques (e.g., immunohistochemistry) or seek a second expert opinion before finalizing the diagnosis. This helps reduce diagnostic errors and improves patient outcomes.

`Model Calibration`

- **AI-Guided Anesthesia Dosing**
  Suppose a deep learning model helps anesthesiologists determine the correct dosage

of anesthesia based on a patient's vitals, metabolism, and medical history. The following scenarios may arise concerning model calibration - An over-confident model that underestimates risk could lead to underdosing, potentially causing the patient to regain consciousness during surgery; Conversely, an under-confident model may overestimate risk, resulting in excessive anesthesia and increasing the likelihood of complications such as respiratory depression or prolonged recovery; A well-calibrated model, however, ensures precise, individualized dosage recommendations, improving surgical safety.

- **AI-Driven Decision Support for Medical Device Failures**
  In a scenario where AI monitors pacemakers or insulin pumps to predict failures before they occur, having a well-calibrated model is crucial to ensure timely and accurate alerts. Overconfident predictions that indicate no issues could cause potential device failures to go undetected, posing serious risks to patient safety. On the other hand, underconfident predictions may trigger frequent, unnecessary device replacements, leading to increased costs and patient inconvenience. A properly calibrated model strikes a balance by providing accurate failure risk estimates, enabling effective preventive maintenance while avoiding unnecessary interventions.

6.3 Future work

We addressed key challenges to improve the trustworthiness with a focus on uncertainty quantification and confidence calibration. However, several unresolved questions and existing gaps remain that warrant further investigation to drive meaningful progress in this domain.

6.3.1 Unexplored Questions/Existing Gaps

- *Limited Research on Calibration of Large Pretrained Models, such as Foundation Models* – Vision foundation models have gained significant traction in the computer vision community for their ability to improve performance across a range of downstream tasks. Despite their widespread adoption, the calibration of these models for medical image analysis tasks remains an underexplored area. Investigating how pretraining on foundation models impacts calibration is essential to ensuring their reliability in the healthcare domain.

- *Limited Understanding of Calibration Differences Between Pretraining on Natural and Medical Images* – As observed in Chapter 4 and supported by previous literature [68], dataset size plays a significant role in calibration. Spurious relationships learned during training can lead models to become overconfident in their predictions [195]. In Section 4.3.5, we observed peculiar behavior when comparing medical pretraining with ImageNet pretraining. The study revealed that the self-supervised model pretrained on RadImageNet exhibited very low overconfidence error but high ECE. This discrepancy could hint at potential underconfidence, stemming from

substantial regularization induced by the type of SSL task combined with medical image-based pretraining. However, drawing definitive conclusions is premature, as further experiments, encompassing various architectures and datasets, are necessary. Hence, investigating pretraining based on various datasets with respect to calibration is crucial for building well-calibrated models.

- *Overconfidence-to-Underconfidence*: Several methods have been developed to reduce overconfidence in modern DNNs; however, combining multiple techniques can sometimes result in an underconfident model [2, 196]. Therefore, it is essential to examine this shift from overconfidence-to-underconfidence [197] in DNN predictions and devise strategies that adjust the confidence dynamically – lowering it when predictions are overly confident and boosting it when they are underconfident.

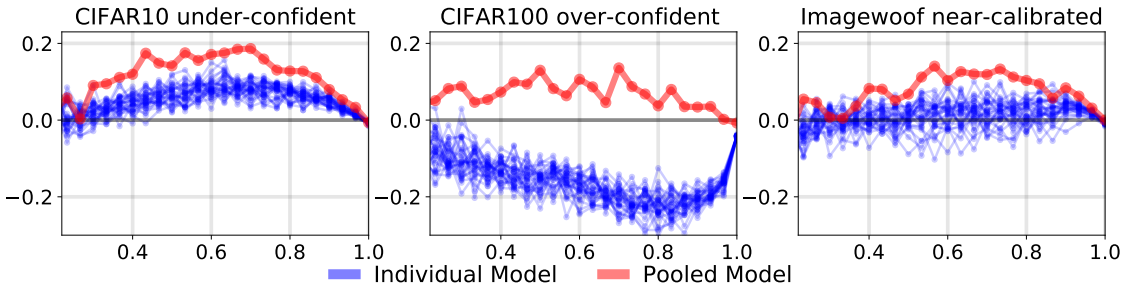


Figure 6.2: Representing three scenarios of miscalibration (a) under-confidence (b) over-confidence, and (c) near-calibrated. *Image Credits* [2]

As illustrated [2] in Figure 6.2, confidence of predictions can be categorized as under-confident, over-confident, or near-calibrated. Existing techniques as discussed in Chapter 5, when applied indiscriminately, may effectively address overconfidence but often fail in underconfident or near-calibrated scenarios. Future work should focus on developing a dynamic approach capable of adapting to the model's current calibration state. Building on the insights from Chapter 5, future methods are encouraged to improve calibration while also maintaining or boosting model performance relative to existing approaches. Additionally, they should be robust enough to handle out-of-distribution (OOD) scenarios effectively, ensuring reliability across both in-distribution and out-of-distribution data.

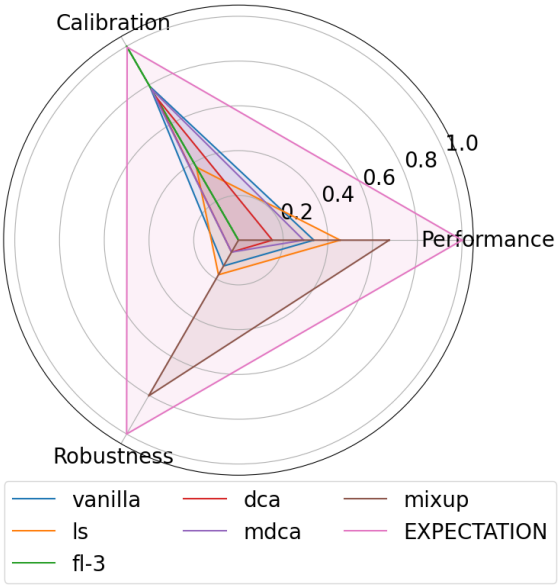


Figure 6.3: Future methods should perform well in all three aspects: Performance, Calibration and OOD Robustness.

Figure 6.3 highlights the anticipated advancements, demonstrating improvements across three critical aspects: *(1) performance, (2) calibration, and (3) OOD robustness.*

References

[1] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 2023. xv, 28

[2] Rahul Rahaman and alexandre thiery. Uncertainty quantification and deep ensembles. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, pages 20063–20075, 2021. xviii, 7, 44, 87

[3] A. Krizhevsky, I. Sutskever, and G. E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 1, 43, 45, 69

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 43

[5] Moloud Abdar, Abbas Khosravi, Sheikh Mohammed Shariful Islam, U. Rajendra Acharya, and Athanasios V. Vasilakos. The need for quantification of uncertainty in artificial intelligence for clinical data analysis: increasing the level of trust in the decision-making process. *IEEE Systems, Man, and Cybernetics Magazine*, 8(3):28–40, 2022. 1

[6] Moloud Abdar, Maryam Samami, Sajjad Dehghani Mahmoodabad, Thang Doan, Bogdan Mazoure, Reza Hashemifesharaki, Li Liu, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. Uncertainty quantification in skin cancer classification using three-way decision-based bayesian deep learning. *Computers in Biology and Medicine*, 135:104418, 2021. ISSN 0010-4825. 1

[7] Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and

challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in Neuroscience*, 14, 2020. 1, 6, 44

[8] A. Jungo and M. Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2019. 1, 6, 44, 69

[9] Yifei Zhang, Dun Zeng, Jinglong Luo, Zenglin Xu, and Irwin King. A survey of trustworthy federated learning with perspectives on security, robustness, and privacy, 2023. 1

[10] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9), jan 2023. ISSN 0360-0300. doi: 10.1145/3555803. 1

[11] Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene, and Michel Dojat. Trustworthy clinical ai solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*, 150:102830, 2024. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2024.102830. 1

[12] Min-Jen Tsai, Ping-Yi Lin, and Ming-En Lee. Adversarial attacks on medical image classification. *Cancers (Basel)*, 15(17), August 2023. 2

[13] Vera Sorin, Shelly Soffer, Benjamin S. Glicksberg, Yiftach Barash, Eli Konen, and Eyal Klang. Adversarial attacks in radiology - a systematic review. *European Journal of Radiology*, 167, Oct 2023. ISSN 0720-048X. 2

[14] Xiaoshuang Shi, Yifan Peng, Qingyu Chen, Tiarnan Keenan, Alisa T. Thavikulwat, Sungwon Lee, Yuxing Tang, Emily Y. Chew, Ronald M. Summers, and Zhiyong Lu. Robust convolutional neural networks against adversarial attacks on medical images. *Pattern Recognition*, 132:108923, 2022. ISSN 0031-3203. 2, 11

[15] Rahul Paul, Matthew Schabath, Robert Gillies, Lawrence Hall, and Dmitry Goldgof. Mitigating adversarial attacks on medical image understanding systems. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020. 2, 11

[16] Gerda Bortsova, Cristina González-Gonzalo, Suzanne C. Wetstein, Florian Dubost, Ioannis Katramados, Laurens Hogeweg, Bart Liefers, Bram van Ginneken, Josien P.W. Pluim, Mitko Veta, Clara I. Sánchez, and Marleen de Bruijne. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*, 73:102141, 2021. ISSN 1361-8415. 2, 12

[17] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23 (5):828–841, 2019. 2, 11

[18] Danilo Vasconcellos Vargas and Jiawei Su. Understanding the one pixel attack: Propagation maps and locality analysis. In *Proceedings of the Workshop on Artificial Intelligence Safety 2020 co-located with the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI 2020), Yokohama, Japan, January, 2021*, volume 2640, 2020. 2, 11

[19] Utku Ozbulak, Arnout Van Messem, and Wesley De Neve. Impact of adversarial examples on deep learning models for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 300–308, 2019. 2

[20] Antoine Sanner, Camila González, and Anirban Mukhopadhyay. How reliable are out-of-distribution generalization methods for medical image segmentation? In *Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021*, page 604–617, 2021. ISBN 978-3-030-92658-8. 3

[21] Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A. Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A survey, 2024. URL https://arxiv.org/abs/2310.08598. 3

[22] Tianshi Cao, Chin-Wei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection, 2020. URL https://arxiv.org/abs/2007.04250. 3

[23] Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, Gavin Siew Wei Tan, Leopold Schmetterer, Pearse A Keane, and Tien Yin Wong. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175, 2019. ISSN 0007-1161. doi: 10.1136/bjophthalmol-2018-313173. 3, 12

[24] Shujun Wang, Lequan Yu, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Patch-based output space adversarial learning for joint optic disc and cup segmentation. *IEEE Transactions on Medical Imaging (TMI)*, 38(11):2485–2495, 2019. doi: 10.1109/TMI.2019.2899910. 3

[25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 3, 6, 16, 27, 44, 69

[26] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR*, 2017. 3

[27] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework

for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. 3

[28] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 3

[29] Davood Karimi and Ali Gholipour. Improving calibration and out-of-distribution detection in deep models for medical image segmentation. *IEEE Transactions on Artificial Intelligence*, 4(2):383–397, 2023. 3

[30] Steffen Czolbe, Kasra Arnavaz, Oswin Krause, and Aasa Feragen. Is segmentation uncertainty useful? In *Information Processing in Medical Imaging*, pages 715–726, 2021. 3

[31] Ling Huang, Su Ruan, Yucheng Xing, and Mengling Feng. A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. *Medical Image Analysis*, 97:103223, 2024. ISSN 1361-8415. 3, 6

[32] Mehedi Hasan, Moloud Abdar, Abbas Khosravi, Uwe Aickelin, Pietro Lio', Ibrahim Hossain, Ashikur Rahman, and Saeid Nahavandi. Survey on leveraging uncertainty estimation towards trustworthy deep neural networks: The case of reject option and post-training processing, 2023. 3, 5

[33] Rohit Jena and Suyash P. Awate. A bayesian neural net to segment images with uncertainty estimates and good calibration. In *Information Processing in Medical Imaging (IPMI)*, 2019. 3, 6, 17, 27, 44

[34] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017. 3, 5, 6, 16, 27, 30, 35, 36, 37

[35] B. Kompa, J. Snoek, and A. L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 2021. 3, 8, 43, 69

[36] Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 2020. 3, 6, 44

[37] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017. 3, 7, 8, 20, 24, 43, 44, 47, 48, 61, 69, 70

[38] Agostina J. Larrazabal, César Martínez, Jose Dolz, and Enzo Ferrante. Maximum entropy on erroneous predictions: Improving model calibration for medical image

segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2023. 3, 8, 23, 44, 69

[39] Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *European Commission*, 2021. URL https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021PC0206. 4

[40] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 4

[41] Hristina Uzunova, Jan Ehrhardt, Timo Kepp, and Heinz Handels. Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. In *Medical Imaging 2019: Image Processing*, 2019. 4, 14, 43

[42] Matan Atad, David Schinz, Hendrik Moeller, Robert Graf, Benedikt Wiestler, Daniel Rueckert, Nassir Navab, Jan S Kirschke, and Matthias Keicher. Counterfactual explanations for medical image classification and regression using diffusion autoencoder. *arXiv preprint arXiv:2408.01571*, 2024. 4

[43] G. Scafarto, N. Posocco, and A. Bonnefoy. Calibrate to interpret. In Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas, editors, *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2023. 4, 7, 43

[44] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research (PMLR)*, pages 77–91, 23–24 Feb 2018. 4

[45] Esther Puyol-Antón, Bram Ruijsink, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Reza Razavi, and Andrew P. King. Fairness in cardiac mr image analysis: An investigation of bias due to data imbalance in deep learning based segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 413–423, 2021. ISBN 978-3-030-87199-4. 4, 14

[46] Tiarna Lee, Esther Puyol-Antón, Bram Ruijsink, Miaojing Shi, and Andrew P. King. A systematic study of race and sex bias in cnn-based cardiac mr segmentation. In *Statistical Atlases and Computational Models of the Heart. Regular and CMRxMotion Challenge Papers: 13th International Workshop, STACOM 2022, Held in Conjunction with MICCAI 2022*, page 233–244, 2023. ISBN 978-3-031-23442-2. 4, 14

[47] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2019. 4

[48] Melanie Ganz, Sune Hannibal Holm, and Aasa Feragen. Assessing bias in medical ai. In *International Conference on Machine Learning (ICML), Interpretable ML in Healthcare Workshop*, 2021. 4

[49] Jiyeong Kim, Zhuo Ran Cai, Michael L. Chen, Julia F. Simard, and Eleni Linos. Assessing Biases in Medical Decisions via Clinician and AI Chatbot Responses to Patient Vignettes. *JAMA Network Open*, 6(10):e2338050–e2338050, 10 2023. ISSN 2574-3805. 4

[50] Jakub Nalepa, Michal Marcinkiewicz, and Michal Kawulok. Data augmentation for brain-tumor segmentation: A review. *Frontiers in Computational Neuroscience*, 13: 83, 2019. ISSN 1662-5188. 5, 27, 36, 38

[51] Ryutaro Tanno, Daniel E. Worrall, Enrico Kaden, Aurobrata Ghosh, Francesco Grussu, Alberto Bizzi, Stamatios N. Sotiropoulos, Antonio Criminisi, and Daniel C. Alexander. Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion mri. *NeuroImage*, 225:117366, 2021. ISSN 1053-8119. 5, 27

[52] Stephen C. Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering System Safety*, 54(2):217–223, 1996. ISSN 0951-8320. Treatment of Aleatory and Epistemic Uncertainty. 5

[53] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009. ISSN 0167-4730. Risk Acceptance and Risk Communication. 5

[54] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110 (3):457–506, Mar 2021. ISSN 1573-0565. 5

[55] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 2019. 6, 18, 44

[56] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *International Conference on Machine Learning (ICML)*, 2015. 6, 15, 44

[57] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, page 1050–1059, 2016. 6, 15, 16, 18, 29, 44

[58] Allan H. Murphy and Robert L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1977. 7, 43

[59] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015. 7, 43

[60] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 1999. 7, 19, 44, 70

[61] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *International Conference on Learning Representations (ICLR)*, 2020. 7, 44

[62] Takuo Matsubara, Niek Tax, Richard Mudd, and Ido Guy. TCE: A test-based approach to measuring calibration error. In *Uncertainty in Artificial Intelligence*, 2023. 7, 67

[63] Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 7, 22, 24, 44, 47, 48, 49

[64] Aditya Singh, Alessandro Bay, Biswa Sengupta, and Andrea Mirabile. On the dark side of calibration for modern neural networks. In *Workshop on Uncertainty and Robustness in Deep Learning (UDL)*, 2021. 7, 44

[65] Curtis P. Langlotz et al. A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 nih/rsna/acr/the academy workshop. *Radiology*, 2019. 7, 44

[66] Xueyan Mei et al. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 2022. 7, 44, 50, 65

[67] Kai Ma, Siyuan He, Pengcheng Xi, Ashkan Ebadi, Stéphane Tremblay, and Alexander Wong. A trustworthy framework for medical image analysis with deep learning. *arXiv preprint arXiv:2212.02764*, 2022. 7, 43, 44

[68] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Neural Information Processing Systems (NeurIPS)*, 2019. 7, 44, 45, 51, 86

[69] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Neural Information Processing Systems (NeurIPS)*, 2019. 7, 44, 46

[70] Fernando Navarro, Christopher Watanabe, et al. Evaluating the robustness of self-supervised learning in medical imaging. *ArXiv*, 2021. 7, 44

[71] Atharva Tendle and Mohammad Rashedul Hasan. A study of the generalizability of self-supervised representations. *Machine Learning with Applications*, 2021. 7, 44

[72] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 7, 44

[73] L. Frenkel and J. Goldberger. Calibration of medical imaging classification systems with weight scaling. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2022. 8, 20, 44, 70

[74] R. Müller, S. Kornblith, and G. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 8, 21, 69, 70

[75] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 8, 21, 69, 70, 71, 73, 74, 79

[76] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. In *International Conference on Learning Representations (ICLR), Workshop Track Proceedings*, 2017. 8, 21, 70

[77] Gongbo Liang, Yu Zhang, Xiaoqin Wang, and Nathan Jacobs. Improved trainable calibration method for neural networks. In *British Machine Vision Conference (BMVC)*, 2020. 8, 22, 70, 73, 74, 79

[78] Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8, 22, 70, 73, 74, 79

[79] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, and P. Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 8, 21, 44, 69, 70, 73, 74, 79

[80] Balamurali Murugesan, Bingyuan Liu, Adrian Galdran, Ismail Ben Ayed, and Jose Dolz. Calibrating segmentation networks with margin-based label smoothing. *Medical Image Analysis*, 2023. 8, 22, 44, 69

[81] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572. 11

[82] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 11

[83] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, 2018. URL https://arxiv.org/abs/1712.04248. 11

[84] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *36th International Conference on Machine Learning (ICML)*, volume 97, pages 2484–2493. PMLR, 2019. 11

[85] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey, 2023. URL https://arxiv.org/abs/2108.13624. 12

[86] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, page 691–697, 2018. ISBN 9780999241127. 12

[87] Yue Zhang, Shun Miao, Tommaso Mansi, and Rui Liao. Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 599–607, 2018. ISBN 978-3-030-00934-2. 12

[88] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Transactions on Medical Imaging (TMI)*, 39(12):4237–4248, 2020. 13

[89] Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 2022. 13, 43

[90] Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable AI techniques in healthcare. *Sensors (Basel)*, 23(2):634, January 2023. 13

[91] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR), Workshop Track Proceedings*, 2013. 13

[92] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision (ECCV)*, pages 818–833, 2014. ISBN 978-3-319-10590-1. 13, 14

[93] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 13

[94] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 13

[95] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 13

[96] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *IEEE International Conference on Computer Vision (ICCV)*, 2017. 13, 43

[97] Avleen Malhi, Timotheus Kampik, Husanbir Pannu, Manik Madhikermi, and Kary Främling. Explaining machine learning-based classifications of in-vivo gastral images. In *Digital Image Computing: Techniques and Applications (DICTA)*, 2019. 13

[98] Dimitrios Lenis, David Major, Maria Wimmer, Astrid Berg, Gert Sluiter, and Katja Bühler. Domain aware medical image classifier interpretation by counterfactual impact analysis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020. 14

[99] Sam Maksoud, Arnold Wiliem, Kun Zhao, Teng Zhang, Lin Wu, and Brian Lovell. Coral8: Concurrent object regression for area localization in medical image panels. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2019. 14

[100] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 14

[101] Sven Koitka, Moon S. Kim, Ming Qu, Asja Fischer, Christoph M. Friedrich, and Felix Nensa. Mimicking the radiologists' workflow: Estimating pediatric hand bone age with stacked deep neural networks. *Medical Image Analysis*, 64:101743, 2020. ISSN 1361-8415. 14

[102] Satyananda Kashyap, Alexandros Karargyris, Joy Wu, Yaniv Gur, Arjun Sharma, Ken C. L. Wong, Mehdi Moradi, and Tanveer Syeda-Mahmood. Looking in the right

place for anomalies: Explainable ai through automatic location learning. In *IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020. 14

[103] Shiwen Shen, Simon X Han, Denise R Aberle, Alex A Bui, and William Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Systems with Applications*, 128:84–95, 2019. ISSN 0957-4174. 14

[104] Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt E. J. Michels, Gerard Schouten, and Veronika Cheplygina. Risk of training diagnostic algorithms on data with demographic bias. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 183–192. Springer International Publishing, 2020. 14

[105] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, Dec 2021. ISSN 1546-170X. 14

[106] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences of the United States of America*, 117(23):12592–12594, June 2020. 14

[107] Eike Petersen, Aasa Feragen, Maria Luise da Costa Zemsch, Anders Henriksen, Oskar Eiler Wiese Christensen, and Melanie Ganz. Feature robustness and sex differences in medical imaging: A case study in mri-based alzheimer's disease detection. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 88–98, 2022. 14

[108] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 2021. ISSN 0360-0300. 14

[109] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, page 429–440, 2021. ISBN 9781450385626. 14

[110] Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 12091–12103, 2021. 14

[111] J. M. Bernardo and A. F. Smith. Bayesian Theory. *John Wiley & Sons*, 405, 2009. 15

[112] D. J. MacKay. Bayesian methods for adaptive models. *PhD thesis, California Institute of Technology*, 1992. 15

[113] R. M. Neal. Bayesian Learning for Neural Networks. *Springer-Verlag New York, Inc.*, 1996. 15

[114] Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, 2011. 15

[115] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014. 15, 55

[116] Yarin Gal. Uncertainty in deep learning. In *PhD thesis, University of Cambridge*, 2016. 15

[117] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60, 1994. 16

[118] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective, 2020. URL https://arxiv.org/abs/1912.02757. 16

[119] Yifan Mao, Fei-Fei Xue, Ruixuan Wang, Jianguo Zhang, Wei-Shi Zheng, and Hongmei Liu. Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 529–538, 2020. ISBN 978-3-030-59725-2. 17

[120] Teresa Araújo, Guilherme Aresta, Luís Mendonça, Susana Penas, Carolina Maia, Ângela Carneiro, Ana Maria Mendonça, and Aurélio Campilho. Dr|graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, 63:101715, 2020. ISSN 1361-8415. 17

[121] Ishaan Bhat, Hugo J. Kuijf, Veronika Cheplygina, and Josien P.W. Pluim. Using uncertainty estimation to reduce false positives in liver lesion detection. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 663–667, 2021. 17

[122] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 17

[123] Vijay Badrinarayanan Alex Kendall and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 57.1–57.12, 2017. 17, 27

[124] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, 10 2017. ISSN 0008-5472. 18

[125] Haleh Akrami, Anand Joshi, Sergul Aydore, and Richard Leahy. Quantile regression for uncertainty estimation in vaes with applications to brain lesion detection. In *Information Processing in Medical Imaging (IPMI)*, pages 689–700, 2021. 18

[126] Jiawei Yang, Yuan Liang, Yao Zhang, Weinan Song, Kun Wang, and Lei He. Exploring instance-level uncertainty for medical detection. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 448–452, 2021. 18

[127] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018. 18

[128] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. 19

[129] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems*, volume 33, pages 14927–14937, 2020. 19

[130] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2002. 20, 70

[131] B. Ji, H. Jung, J. Yoon, K. Kim, and y. Shin. Bin-wise temperature scaling (bts): Improvement in confidence calibration performance through simple scaling techniques. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4190–4196, 2019. 20

[132] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 80–88, June 2022. 21, 22

[133] Mobarakol Islam, Lalithkumar Seenivasana, Hongliang Ren, and Ben Glocker. Class-distribution-aware calibration for long-tailed visual recognition. In *International Conference on Machine Learning (ICML), Uncertainty and Robustness in Deep Learning Workshop*, 2021. 21, 23, 70

[134] C. Zhang, P. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, and M. Cheng. Delving deep into label smoothing. *IEEE Transactions on Image Processing (TIP)*, 2021. 21, 69

[135] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning (ICML)*, 2018. 21

[136] Arthur Gretton. Introduction to rkhs, and some simple kernel algorithms. *Adv. Top. Mach. Learn. Lecture Conducted from University College London*, 16(5-3):2, 2013. 21

[137] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 21

[138] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015. 22, 24, 47, 48

[139] Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Meta-calibration: Learning of model calibration using differentiable expected calibration error. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. 23

[140] Sivaramakrishnan Rajaraman, Prasanth Ganesan, and Sameer Antani. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. *PLOS ONE*, 17(1):1–23, 01 2022. doi: 10.1371/journal.pone.0262838. 23

[141] Adrian Galdran, Johan W. Verjans, Gustavo Carneiro, and Miguel A. González Ballester. Multi-head multi-loss model calibration. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 108–117, 2023. ISBN 978-3-031-43898-1. 24

[142] Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2015. 24, 44, 48, 49

[143] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2019. 24, 49

[144] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning (ICML)*, 2019. 24, 44, 49

[145] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural

networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 24, 44, 48, 69

[146] Sebastian Gregor Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 25

[147] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1950. 25, 48

[148] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association (JASA)*, 2007. 25, 48

[149] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019. 25, 49

[150] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew B. Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, Feb 2024. ISSN 1548-7105. 25

[151] Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable lp canonical calibration error estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 7933–7946, 2022. 25

[152] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1613–1622, 2015. 27

[153] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association : JAMIA*, 2012. 43

[154] Christian Tomani and Florian Buettner. Towards trustworthy predictions from deep neural networks with fast adversarial calibration. In *AAAI Conference on Artificial Intelligence*, 2019. 43

[155] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018. 43

[156] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *International Conference on Learning Representations (ICLR)*, 2017. 43

[157] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning (ICML)*, 2017. 43

[158] Abhishek Singh Sambyal, Narayanan C Krishnan, and Deepti R Bathula. Towards reducing aleatoric uncertainty for medical imaging tasks. In *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022. 44

[159] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 44

[160] Skylar E. Stolte, Kyle Volle, Aprinda Indahlastari, Alejandro Albizu, Adam J. Woods, Kevin Brink, Matthew Hale, and Ruogu Fang. Domino: Domain-aware model calibration in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2022. 44

[161] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 44

[162] Dongdong Wang, Boqing Gong, and Liqiang Wang. On calibrating semantic segmentation models: Analyses and an algorithm. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 44

[163] Jeff Donahue, Yangqing Jia, et al. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning (ICML)*, 2014. 45

[164] Yang Wen, Leiting Chen, Yu Deng, and Chuan Zhou. Rethinking pre-training on medical imaging. *Journal of Visual Communication and Image Representation*, 2021. 45, 50, 73

[165] Shekoofeh Azizi et al. Big self-supervised models advance medical image classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 45, 50, 73

[166] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 45

[167] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 46

[168] Jochen Kruppa, Yufeng Liu, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal*, 2014. 48

[169] Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2015. 49

[170] Joaquin Quiñonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. Evaluating predictive uncertainty challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, 2006. 49

[171] EyePACS, Diabetic Retinopathy Detection. https://www.kaggle.com/competitions/diabetic-retinopathy-detection/. 49

[172] Babak Ehteshami Bejnordi et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 2017. 49

[173] Histopathologic Cancer Detection: Modified version of the PatchCamelyon (PCam) Benchmark Dataset. https://www.kaggle.com/competitions/histopathologic-cancer-detection/. 49

[174] Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018. 49

[175] Covid-19 Image Dataset. https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset. 49

[176] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv 2003.11597*, 2020. URL https://github.com/ieee8023/covid-chestxray-dataset. 49

[177] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv 2006.11988*, 2020. URL https://github.com/ieee8023/covid-chestxray-dataset. 49

[178] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 50

[179] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016. 50

[180] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *IEEE International Conference on Computer Vision (ICCV)*, 2015. 50

[181] Andrew Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *International Conference on Machine Learning (ICML)*, 2004. 55

[182] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019. 61

[183] Diego Doimo, Aldo Glielmo, Sebastian Goldt, and Alessandro Laio. Redundant representations help generalization in wide neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2022. 61

[184] Matthias Minderer et al. Revisiting the calibration of modern neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2021. 67

[185] José Hernández-Orallo, Peter Flach, and Cèsar Ferri. A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 2012. 67

[186] Y. Qin, X. Wang, B. Lakshminarayanan, Ed H. Chi, and A. Beutel. What are effective labels for augmented data? improving calibration and robustness with autolabel. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023. 69

[187] Abhishek Singh Sambyal, Usma Niyaz, Narayanan C. Krishnan, and Deepti R. Bathula. Understanding calibration of deep neural networks for medical image classification. *Computer Methods and Programs in Biomedicine*, 2023. 69, 73

[188] Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Meta-calibration: Learning of model calibration using differentiable expected calibration error. *Transactions on Machine Learning Research (TMLR)*, 2023. 70

[189] C. Zhu, W. Chen, T. Peng, Y. Wang, and M. Jin. Hard sample aware noise robust learning for histopathology image classification. *IEEE Transactions on Medical Imaging (TMI)*, 2022. 72

[190] J. Wei, A. Suriawinata, B. Ren, X. Liu, M. Lisovsky, L. Vaickus, et al. A petri dish for histopathology image analysis. In *International Conference on Artificial Intelligence in Medicine (AIME)*, 2021. 72

[191] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019. 73

[192] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161, Aug 2018. ISSN 2052-4463. 73

[193] Usma Niyaz, Abhishek Singh Sambyal, and Deepti R. Bathula. Leveraging different learning styles for improved knowledge distillation in biomedical imaging. *Computers in Biology and Medicine*, 2024. 73

[194] M. Andrey, B. Neil, G. Yarin, G. Mark, G. Alexander, C. German, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. 75

[195] Ujwal Krothapalli and Lynn Abbott. One size doesn't fit all: Adaptive label smoothing, 2021. URL `https://openreview.net/forum?id=wqRvVvMbJAT`. 86

[196] Yeming Wen, Ghassen Jerfel, Rafael Muller, Michael W Dusenberry, Jasper Snoek, Balaji Lakshminarayanan, and Dustin Tran. Combining ensembles and data augmentation can harm your calibration. In *International Conference on Learning Representations (ICLR)*, 2021. 87

[197] Shuang Ao, Stefan Rueger, and Advaith Siddharthan. Two sides of mis-calibration: Identifying over and under-confidence prediction for network calibration. In *The 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023. 87