DOCTOR OF PHILOSOPHY

Deep Learning-Based Medical Image Analysis Methods for Reducing Annotation Costs and Predictive Triage Misdiagnosis

Carse, Jacob

Award date:
2023

Licence:
CC BY-NC-ND

Deep Learning-Based Medical Image Analysis Methods for Reducing Annotation Costs and Predictive Triage Misdiagnosis

Jacob Carse

2023

A thesis submitted to The University of Dundee in accordance with the requirements for the award of

Doctorate of Philosophy

Based on research carried out under the supervision of

Professor Stephen McKenna

Professor Frank Carey

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AMDIM	Augmented Multiscale Deep InfoMax
BAD	British Association of Dermatologists
BALD	Bayesian Active Learning Disagreement
BFGS	Broyden–Fletcher–Goldfarb–Shanno Algorithm
CEAL	Cost-Effective Active Learning
CEREALS	Cost-Effective Region-based Active Learning for Semantic Segmentation
CNN	Convolutional Neural Network
CPC	Contrastive Predictive Coding
DBAL	Diverse Mini-Batch Active Learning
ECE	Expected Calibration Error
EC-SelectiveNet	Expected Cost SelectiveNet
ELBO	Evidence Lower Bound
GAN	Generative Adversarial Network
H&E	Hematoxylin and Eosin
ISIC	International Skin Imaging Collaboration
KDE	Kernel Density Estimator
MAP	Maximum A Posteriori
MCE	Maximum Calibration Error
mIoU	Mean Intersection Over Union
MRI	Magnetic Resonance Imaging
NCE	Noise-Contrastive Estimation
NHS	National Health Service
PCam	Patch Camelyon
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SWIN	Shifted Window Transformer
TCAL	Triple Criteria Active Learning

Acknowledgements

I dedicate this thesis to all the people who have helped me along my PhD journey and would like to express my sincere gratitude to them.

I want to thank my partner **Tasnim Hassan** first and foremost for her unwavering support, encouragement, and tolerance. Without her, I would not have been able to complete this. She has been my pillar and my motivation. She has always encouraged me to pursue my goals and has always had faith in me. She is an incredible blessing in my life.

I also want to express my gratitude to my supervisor, **Professor Stephen McKenna**, for his advice and assistance over the years. He has taught me how to conduct amazing research and communicate clearly. He has been a great mentor and role model for me. He has also given me numerous chances to develop both personally and as a researcher. I have learned so much from him and I am honoured to be his student.

Also, I want to express my gratitude to the entire **Computer Vision and Image Processing (CVIP)** group for their camaraderie and cooperation. They have developed a fun and engaging environment for research, and I have profited from their insightful comments and recommendations. It is a pleasure to work with such lovely people, so I'd also like to thank the **Dermatology research group** members for their kind remarks and ongoing collaboration.

Last but not least, I want to express my gratitude to my family and friends for their unfaltering support and inspiration. They have always supported me through the highs and lows of this trip and have experienced both my joys and my frustrations alongside me. They have also given me the courage and assurance I needed to face and overcome the obstacles I encountered along the way.

Declaration

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it contains no material previously published or produced by another party in fulfilment, partial or otherwise, of any other degree or diploma at another University or Institute of higher education, except where due acknowledgement is made in the text.

Jacob Carse 2023

Abstract

Medical image analysis is a critical and challenging field that can be significantly enhanced using deep learning techniques. However, these models require large amounts of annotated data, which can be costly and time-consuming to obtain. Additionally, deep learning models often suffer from overconfidence and poor generalisation, leading to incorrect diagnoses and negative clinical outcomes. The primary objective of this thesis is to address these challenges by proposing methods that reduce annotation costs and improve diagnostic accuracy using deep learning.

The first contribution of this thesis is an active learning framework designed to increase annotation throughput for histopathology patches. Histopathology is the gold standard for cancer diagnosis, but it requires manual examination by pathologists, which is labour-intensive and prone to errors. To address this issue, this thesis proposes an active learning framework that selects regions for annotation composed of multiple patches, which is expected to increase annotation throughput. This framework is evaluated with various query strategies for nuclei classification using convolutional neural networks (CNN) trained on small patches containing single nuclei.

This thesis proposes a multi-directional modification to the contrastive predictive coding (CPC) method for unsupervised representation learning for histopathology patches. Recent advancements in deep learning have had a significant impact on digital pathology, however a significant challenge is the large amounts of annotated data needed. Unsupervised representation learning aims to learn meaningful and transferable features from unannotated data, which can be useful for downstream tasks such as classification. The proposed method uses an alternative mask to construct a latent context and a multi-directional PixelCNN autoregressor, to learn effective deep feature representations for improved classification accuracy in digital pathology compared to the standard implementation of CPC.

The third contribution of this thesis is a study on calibration techniques evaluated on a multi-class dermatology dataset and a binary histopathology dataset. Calibration is critical for medical image analysis, where overconfident or underconfident predictions can have serious consequences for patient care. The study applied the temperature scaling method and alternative calibration metrics to networks trained with one-hot encoding, cross-entropy loss, focal loss, and label smoothing. The findings suggest that temperature scaling of networks trained with focal loss and appropriate hyperparameters demonstrated strong performance in terms of both calibration and accuracy across both datasets.

This thesis investigates selective classification methods with asymmetrical misdiagnosis costs for skin lesion images. Selective classification is a decision-making framework that allows a model to reject images when it is uncertain or unconfident, which can reduce the risk of misdiagnosis and improve patient safety. However, most existing selective classification methods assume that all types of misclassification have equal costs, which is not realistic in medical image analysis. This thesis evaluates various methods of uncertainty estimation with neural networks and probability calibration. Additionally, a modification to SelectiveNet, called EC-SelectiveNet, is proposed, which discards the selection head during testing and relies on expected costs to make decisions. The results demonstrate the advantages of training for full coverage, even when operating at lower coverage, and show that EC-SelectiveNet outperforms other selective classification methods, in both symmetric and asymmetric cost settings.

The fifth contribution of this thesis is a study on dataset fine-tuning for skin lesion image datasets. Dataset fine-tuning is challenging for medical image analysis due to the heterogeneity and variability of data sources. This study utilises four diagnostic image datasets, including two locally sourced datasets from NHS Tayside and NHS Forth Valley and two publicly available datasets. The study emphasises the importance of assessing the generalisability of deep learning algorithms for macroscopic skin lesion images in real-world settings and highlights the potential benefits of utilising large public macroscopic datasets for pre-training and fine-tuning.

Chapter 1

Introduction

1.1 Research Motivations

In recent times, contemporary artificial intelligence (AI) techniques have exhibited remarkable results across various fields (Sarker, 2021). The upsurge in deep learning, a machine learning approach where a model with deep layers learns meaningful representations and task-specific outputs (such as classification tasks) jointly, is the primary reason for this progress (LeCun et al., 2015). By eliminating the need for complex feature engineering, this approach leads to enhanced model performance, with the learned representations being more task specific. In the medical domain, researchers have explored the potential of AI and deep learning algorithms to improve the performance of medical image analysis tasks, including classification, segmentation, and detection across different modalities. Integrating these techniques intelligently into clinical workflows can aid in enhancing efficiency by automating complex and time-consuming tasks requiring expert knowledge.

An example of AI integration in a clinical domain can be observed in the realm of digital pathology—an increasingly burgeoning discipline encompassing the application of digital tools and methodologies for the analysis and interpretation of medical images. AI has found practical utility within the digital pathology clinical pipeline by providing automated analyses, thereby alleviating the time-consuming burden for clinicians, and aiding in diagnostic decision-making processes. A notable challenge encountered by clinical pathologists pertains to the evaluation of entire slide images due to their sheer size, rendering a comprehensive assessment daunting. Nevertheless, AI algorithms demonstrate competence in analysing entire slides, identifying critical regions of interest, and furnishing supplementary information, such as nuclei counts by classification, thereby facilitating the diagnostic process (Dimitriou et al., 2019). Significant strides in this area encompass the development of Spatially Constrained Convolutional Neural Network for nucleus detection (Sirinukunwattana et al., 2016) and the Camelyon digital pathology challenge—an open science initiative aimed at developing algorithms for detecting cancer metastasis in whole slide images of lymph node sections (Bejnordi et al., 2017).

Artificial intelligence is rapidly becoming a realistic prospect in dermatology. Empirical investigations have demonstrated that AI algorithms exhibit diagnostic accuracy comparable to that of skin cancer experts when it comes to diagnosing skin lesions in dermoscopic images (Liopyris et al., 2022). Dermoscopy, as a method for capturing images of skin lesions, finds utility in the interface between primary and secondary care. However, reviewing these referred images poses a substantial time burden for dermatologists, making it an area where AI's implementation could significantly alleviate their workload, thereby allowing them to devote more attention to more complex cases. Nonetheless, the integration of AI in dermatology is a multifaceted endeavour due to the intricate nature of patient pathways within the domain. Hence, the comprehensive evaluation of AI interventions mandates clinically-led research studies. A notable illustration of ongoing work in the realm of AI for dermatology lies in the efforts of the International Skin Imaging Collaboration (ISIC) [1], an organization dedicated to collecting and disseminating images for clinical and computer vision research.

Although deep learning algorithms have demonstrated significant improvement in performance for medical image analysis tasks, they necessitate large, annotated datasets to train models and construct meaningful representations necessary for optimal performance. However, acquiring annotations for medical images can be an expensive process, requiring specialised training and a substantial amount of time, compounded by the need for multiple annotations due to inter-observer variability. This challenge has been acknowledged as a primary issue in utilising deep learning algorithms for medical image analysis tasks, as reported in a survey conducted by Litjens et al. (2017).

[1] International Skin Imaging Collaboration: `https://www.isic-archive.com/`

1.2 Cost of Annotations

The limited capacity for annotation in medical image analysis has prompted the exploration of methods designed to address this challenge. One such approach is active learning (Settles, 2009), a machine learning technique that entails iteratively selecting informative samples from a large unannotated dataset to be annotated by an expert, with the aim of enhancing the performance of a predictive model (Figure 1.1). The primary objective of active learning is to achieve high accuracy while utilising fewer annotated examples than conventional supervised learning techniques.

For instance, in the field of dermatology, active learning can be employed to identify representative samples from a vast pool of unannotated images, which can then be annotated by dermatologists. By adopting this strategy, the dermatologist can concentrate on the most informative images, instead of annotating images at random. Despite limitations in the amount of annotated data that can be generated, which is often due to financial or temporal constraints. Consequently, by selectively annotating the most informative images, active learning has the potential to enhance the performance of the resulting machine learning model, thereby reducing the need for additional annotated examples.

Active learning has demonstrated its efficacy in traditional machine learning environments (Settles, 2009), but its applicability in conjunction with deep learning algorithms is subject to limitations that can potentially impede its effectiveness. A particular issue that arises is a selection bias problem (Sener and Savarese, 2017), as active learning algorithms tend to prioritise the selection of the most informative samples. This can engender a biased dataset that is not reflective of the entire popula-

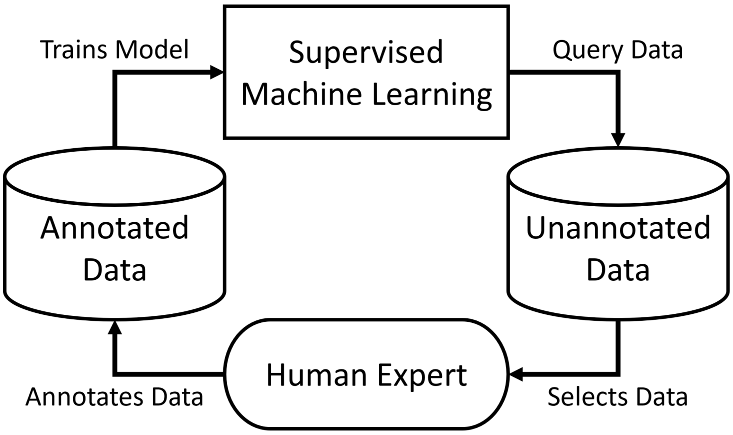


Figure 1.1: Pool-based active learning framework.

tion, ultimately resulting in a deep learning model that acquires feature representations that are not generalisable to unseen data. Consequently, suboptimal classification performance can occur.

Active learning has been shown to be effective in mitigating issues arising from the limited annotation budget; however, further improvements in performance can be achieved by maximising the utility of the unannotated pool. One approach to achieving this is through unsupervised representation learning, which is capable of learning representative features from unannotated images (Bengio, Courville and Vincent, 2013). The acquired representation can then be transferred to other models, thereby reducing the amount of annotated data required to produce a high-performing model. This can help to alleviate the burden on the active learning algorithm in balancing the selection of representative and informative samples. Unsupervised representation learning is typically achieved using either reconstruction or self-supervised methods.

A reconstruction method aims to learn by generating new examples, such as an autoencoder model that compresses its input data into a lower-dimensional latent space and then reconstructs the original data. The learned latent space can be reused as an encoder in a classification model, which can then be trained in a supervised fashion. In contrast, a self-supervised representation method uses a pretext task to learn representational tasks in a supervised fashion. One example of this is the RotNet method (Gidaris et al., 2018), which involves rotating images and having the model predict the rotation of the image. In order to predict the rotation of an image the model has to learn to represent localise salient objects in the images and their relationship to other objects in the image. These learned representations can then be subsequently be transferred to other tasks such as classification. By leveraging these unsupervised representation learning methods, it is possible to extract meaningful features from unannotated data, thereby reducing the burden of annotation and improving the performance of downstream machine learning tasks.

1.3 Triage Misdiagnosis

The limited annotation and the utilisation of techniques to enhance the performance of a specified task may compromise the model's robustness towards images that are outside its training data, such as new disease classifications or diverse image capture conditions. Consequently, the model must be trained to acknowledge the uncertainty of its predictions, which presents a challenge for deep learning algorithms due to their lack of interpretability (Zhang and Zhu, 2018) and inadequately calibrated predictive probabilities (Gal and Ghahramani, 2016). Therefore, it is imperative to investigate methods to better calibrate deep learning algorithms and adopt selective classification approaches to reject images that the model is ill-equipped to handle when making a prediction on new image.

Calibration denotes the systematic procedure of conforming the anticipated probabilities of a model with the authentic probabilities of the target variable. When applied in the realm of deep learning, this procedure necessitates adapting the model's output to match the actual distribution of outcomes in the population under consideration (Guo et al., 2017). Generally, three distinct methodologies are utilised to improve the calibration of deep learning algorithm outputs, namely model regularisation, post-hoc calibration, and Bayesian neural networks. Model regularisation involves the imposition of regularisation during the training phase, while post-hoc calibration entails fine-tuning the output probabilities after the model has undergone training. Additionally, Bayesian neural networks are acknowledged to be intrinsically superior for calibration purposes.

The use of asymmetrical costs is a modification that can be implemented to address the varying costs associated with misclassifications stemming from false positives and false negatives. False positives refer to cases where a healthy patient is incorrectly diagnosed as having a disease, while false negatives denote instances where a patient with a disease is erroneously identified as healthy. The consequences of these types of misclassifications can be vastly dissimilar. In the context of a skin lesion classification scenario, a false positive may lead to unnecessary and potentially harmful interventions such as biopsies, additional testing, and heightened anxiety. Conversely, a false negative could result in delayed treatment, missed diagnoses, and ultimately, a worsened patient outcome. The integration of selective classification alongside asymmetrical costs may yield selective classification systems that reject images that could lead to higher costs if misdiagnoses, thereby minimising the costs associated with such misclassifications.

Selective classification, or classification with a reject option (Chow, 1957), represents a machine learning approach that entails the assignment of one of several conceivable labels to an input image or region of interest, while also incorporating an extra option to reject the image (Figure 1.2).

The integration of a reject option in the classification process permits the system to circumvent erroneous diagnoses or recommendations when the input image's classification is uncertain. In lieu of making an incorrect prediction, the system can request supplementary information or refer the image to a human expert for further evaluation. One prevalent strategy for selective classification involves establishing a threshold on the confidence score generated by the classifier; if the confidence score falls below the threshold, the image is discarded.

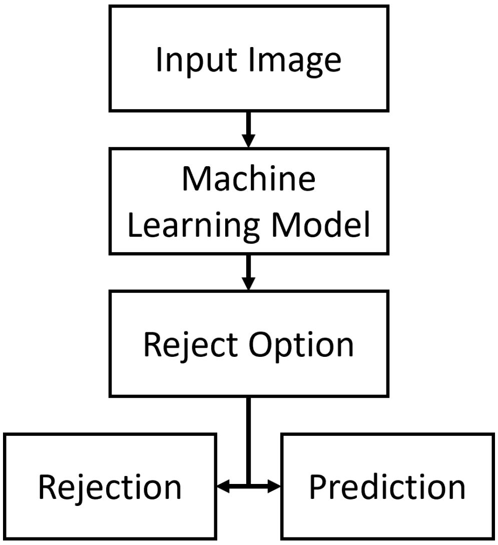


Figure 1.2: Selective Classification framework.

1.4 Dataset Fine-Tuning

The majority of research in the field of medical image analysis employs open-source datasets (Wu et al., 2022). Open-source datasets are publicly released datasets intended to facilitate further research. While such datasets have aided advancements in this area, the resulting machine learning models are often not amenable to generalisation across disparate datasets captured at different sites when fine-tuned (Chin et al., 2022). It is essential for machine learning models to generalise across medical data from distinct capture sites when fine-tuned to achieve broad applicability and effectiveness in diverse clinical contexts. Failure to do so may result in suboptimal performance and inaccuracies that can adversely impact clinical decision-making in medical applications where precision and accuracy are paramount. For example, if a model trained on imaging data from one hospital is applied to imaging data from a different hospital that employs distinct imaging technology, it may generate erroneous results that could lead to incorrect diagnoses and treatments.

1.5 Research Questions and Contributions

Drawing from the aforementioned motivations, this thesis centres on the following research questions:

- *How can a deep learning model be effectively trained to achieve optimal performance when faced with a scarcity of annotations, and a large corpus of unannotated data?*
- *To what extent can selective classification techniques be applied in order to mitigate the costs associated with asymmetrical misdiagnosis of skin lesion images?*

To provide solutions to the questions, this thesis presents the following contributions.

- An active learning framework specifically for histopathology patches, which serves to enhance annotator efficiency and optimise the volume of nuclei annotations (**Chapter 2**).
- Multi-Directional Contrastive Predictive Coding, a modification to a state-of-the-art unsupervised representation learning algorithm, specifically tailored for pathology images that lack a discernible directionality (**Chapter 3**).
- An empirical investigation of deep learning calibration techniques for both multi-class dermatology and binary histopathology patches, encompassing Bayesian neural networks, model regularisation, and post hoc methods (**Chapter 4**).
- A comparative analysis of selective classification techniques, in both binary and multi-class scenarios that encompasses Bayesian neural networks, calibrated neural networks, and specialised models designed for selective classification (**Chapter 5**).
- A study of deep learning algorithms, specifically their capacity to generalise across various dermatology datasets from distinct sources when fine-tuned, some of which were referred for dermatologist examination, and larger open-source dermatology datasets (**Chapter 6**).

1.6 Thesis Structure

This thesis is organised into separate chapters corresponding to the major contributions made. Each chapter includes an exposition of the underlying motivations, a survey of the pertinent literature, and an account of the conducted experiments. This section presents a succinct overview of the individual chapters:

Annotator Efficient Active Learning

With the increasing popularity of deep learning in medical image analysis and digital pathology (Tizhoosh and Pantanowitz, 2018), it has become increasingly important to develop methods that can reduce the need for costly data annotations. Active learning is a promising approach to minimising the amount of annotated data required to train machine learning models (Settles, 2009). However, the effectiveness of traditional active learning strategies with deep learning is limited (Wang et al., 2016). In patch-based machine learning systems, active learning methods typically request annotations for individual small patches, which can be laborious and expensive for annotators who must rely on visual context. To address this issue, Chapter 2 proposes an active learning framework that selects tiles for annotation that are composed of multiple patches, which is expected to increase annotation throughput (Carse and McKenna, 2019). This framework is evaluated with various query strategies on the task of nuclei classification, using convolutional neural networks (CNNs) trained on small patches containing single nuclei. Traditional query strategies performed worse than random sampling.

Unsupervised Representation Learning

Recent advancements in deep learning algorithms have had a significant impact on digital pathology tasks. However, a significant challenge in this field is the need for large amounts of annotated data. To overcome this issue, unsupervised learning techniques, particularly contrastive predictive coding (CPC) (van den Oord et al., 2018), have been proposed to leverage abundant but unannotated data for training classifiers. In Chapter 3, a modification to the CPC framework for use in digital pathology patch classification is purposed, which involves the use of an alternative mask to construct the latent context and a multi-directional PixelCNN autoregressor (van den Oord et al., 2016). Using the Path Camelyon histology patch dataset (Veeling et al., 2018), it is demonstrated that this proposed method can produce effective deep feature representations for improved classification accuracy in digital pathology when compared to the standard implementation of CPC (Carse, Carey and McKenna, 2021).

Predictive Probability Calibration

It is well established that deploying deep learning classifiers for medical image analysis tasks requires careful consideration of issues related to predictive calibration (Maron et al., 2019). Mis-calibration, defined as the discrepancy between predictive probability (confidence) and classification correctness (Guo et al., 2017), can significantly impact the ability to make cost-sensitive and selective decisions (Carse, Süveges, Hogg, Trucco, Proby, Fleming and McKenna, 2021). To understand the effectiveness of various calibration methods, Chapter 4 reports empirical study was conducted on two medical image datasets: one for multi-class dermatology classification and one for binary histopathology image classification. The chapter applied the temperature scaling method, in which the temperature parameter is optimised using various calibration measures instead of the standard negative log-likelihood, to networks trained with one-hot encoding and cross-entropy loss, as well as networks trained with focal loss and label smoothing. The results of these methods were compared to those obtained using two Bayesian neural network approaches. The findings suggest that while alternative calibration metrics may not offer significant advantages for tuning temperature, temperature scaling of networks trained with focal loss and appropriate hyperparameters demonstrated strong performance in terms of both calibration and accuracy across both datasets (Carse et al., 2022).

Asymmetrical Selective Classification

Skin lesion classifiers must enable decision-making that is sensitive to cost. Chapter 5 investigates techniques for selective, cost-sensitive classification in both binary triage and multi-class disease classification scenarios, using misclassification costs provided by clinical dermatologists based on healthcare economics. The chapter evaluates various methods of uncertainty estimation with neural networks and probability calibration. Additionally, a modification to SelectiveNet (Geifman and El-Yaniv, 2019), called EC-SelectiveNet (Carse, Süveges, Hogg, Trucco, Proby, Fleming and McKenna, 2021) is proposed, that discards the selection head during testing and relies on expected costs to make decisions. The results demonstrate the advantages of training for full coverage, even when operating at lower coverage, and show that EC-SelectiveNet outperforms standard CNNs using temperature scaling (Guo et al., 2017) or Bayesian neural networks using different measures of uncertainty, in both symmetric and asymmetric cost settings.

Evaluating Dataset Fine-Tuning

Chapter 6 examines the fine-tuned generalisability of deep neural network classifiers for macroscopic skin lesion images in the UK NHS. Although deep learning has shown promise in dermatology, its ability to accurately diagnose macroscopic skin disease images that lack dermoscopic information remains a significant challenge (Jones et al., 2022). To address this, four diagnostic image datasets were utilised, including two locally-sourced datasets and two publicly available datasets. Two types of neural network models were trained and evaluated on each dataset, with pre-training on the SD-260 (Yang et al., 2019) dataset followed by fine-tuning on the target domain data showing the most promising results. This chapter emphasises the importance of assessing the generalisability of deep learning algorithms when fine-tuned for macroscopic skin lesion images in real-world settings and highlights the potential benefits of utilising large public macroscopic datasets for pre-training and fine-tuning. Future research is necessary to evaluate the generalisability of these algorithms across different populations and acquisition settings when fine-tuned.

Conclusions and Discussions

This chapter briefly articulated and introduced the research conducted, elucidating its contributions and limitations. It highlighted the potential for cost-effective annotation and predictive triage diagnosis in the realm of medical image analysis, with particular emphasis on histopathology and dermatology. This thesis lays groundwork for future research to build upon. This chapter underscores the criticality of mitigating annotation costs and triage misdiagnose, to promote the widespread utilisation of medical image analysis systems in clinical settings.

Chapter 2

Annotator Efficient Active Learning

2.1 Introduction

2.1.1 Active Learning for Medical Image Analysis

Active learning is a type of machine learning that hypothesises that having a learning algorithm select the data that is used during training can reduce the amount of data needed for training (Settles, 2009). Active learning is used within modern applications to reduce the quantity of data that needs to be annotated by selecting unannotated data to be annotated and added to the dataset used to train the model. Limiting the amount of data annotations needed can reduce annotation costs (which can be expensive when dealing with specialised data such as histopathology) and computation costs as the models can be trained with fewer data. In a pool-based scenario, the learning algorithm has access to a large pool of unannotated data. Over multiple iterations, the learning algorithm selects the most beneficial data from the pool to be annotated and added the training dataset, as shown in Figure 1.1. One of the main advantages of pool-based active machine learning for medical image analysis is its ability to reduce the amount of human labour required. Medical image analysis often involves manual annotation, which can be time-consuming and labour-intensive. By using pool-based active learning, the burden of annotation is greatly reduced, as the algorithm can identify the most informative samples and prioritise them for labelling.

Active learning algorithms utilise query strategies to select data for annotation. While some popular query strategies, such as uncertainty sampling, have been demonstrated to be effective on deep learning algorithms (Gal et al., 2017), the unique feature-representation learning process of deep learning algorithms can present challenges. Specifically, the selection of only difficult examples for training can lead to a lower-quality model due to the resulting features not being representative of the entire data distribution. This issue is illustrated by Pop and Fulop (2018), who demonstrate the occurrence of mode collapse when using a Bayesian uncertainty query strategy to train a CNN. To address these challenges, batch-aware query strategies that make use of clustering methods have been shown to be effective in deep learning environments (Sener and Savarese, 2017, Zhdanov, 2019, Kirsch et al., 2019). These strategies optimise the selection of batches of images for annotation rather than individual data points.

2.1.2 Deep Active Learning for Digital Pathology

To save computation time, it is common practice in digital pathology to use patches from whole slide images when applying machine learning algorithms. These patches can be efficiently processed by deep learning algorithms like CNNs, and do not require the entire slide image to be annotated. However, using patch-based methods with patches for tasks such as nuclei detection and classification can be problematic when using active learning to query for annotation. This is because patches are more time-consuming and labour-intensive to annotate, and may lack sufficient spatial context for accurate annotation, even for expert pathologists.

To improve annotation efficiency and reduce costs, this chapter proposes a modified active learning framework to improve annotator throughput by selecting large tiles of whole-slide images made up of multiple nuclei patches to be annotated rather than annotating individual nuclei patches. This modified framework was tested using various active learning query strategies on a nuclei detection and classification task using the CRCHistoPhenotypes dataset (Sirinukunwattana et al., 2016).

This work was presented at the European Congress on Digital Pathology 2019 and published as part of its proceedings (Carse and McKenna, 2019).

2.2 Review of Active Learning for Medical Images

In recent years, active learning has undergone significant development in various areas, including query strategies for deep learning algorithms, techniques to reduce annotator workload, and application to medical image analysis tasks. This section provides an overview of key contributions in these domains that inform the present chapter's contribution.

2.2.1 Pool-Based Active Learning Query Strategies

In pool-based active learning, a large pool of unannotated data and a small set of labelled data are utilised (Settles, 2009). Query strategies are employed to identify the most useful unannotated data for annotation and incorporation into the learning algorithm (Figure 1.1). This approach is particularly relevant in the context of medical image analysis, given the abundance of such data that is often collected and stored, but only a limited portion of which has been thoroughly annotated for use in machine learning applications. This review focuses on query strategies for pool-based active learning in both traditional and deep machine learning algorithms.

Traditional Machine Learning Query Strategies

As the focus of this chapter is limited to deep learning algorithms, only a brief review of pool-based active learning query strategies for traditional machine learning approaches will be performed.

Uncertainty sampling is a query strategy in which the learning algorithm focuses on data points it is most uncertain about to improve model performance (Lewis, 1995). Uncertainty can be measured using techniques such as entropy or distance to the decision boundary. It allows the algorithm to selectively request labels for the most informative data points, leading to more efficient and effective learning.

Query by committee is a query strategy in which a committee of multiple classifiers make predictions on unlabelled data (Seung et al., 1992). If their predictions are diverse or conflicting, the learning algorithm may request a label for that data point. Query by committee can help reduce overfitting and improve generalisation.

Expected model change is a query strategy in which the learning algorithm estimates the change in overall performance after labelling a particular data point and prioritises data points with the greatest expected impact (Settles et al., 2007). This allows the algorithm to focus on data points most likely to improve performance.

Expected error reduction is a method in active learning in which the learning algorithm estimates the reduction in error rate after labelling a particular data point, and prioritises data points with the greatest expected impact (Roy and McCallum, 2001). This allows the algorithm to focus on data points most likely to improve performance.

Variance reduction is a method in active learning in which the learning algorithm prioritises data points that are expected to have the greatest impact on reducing the variance of the predictions (Cohn et al., 1996). This can be achieved by calculating the variance of the predictions for a particular data point.

Density-weighted methods for active learning involve selecting data samples based on the density of the samples in the feature space, to select samples that are underrepresented or less dense (Settles and Craven, 2008). These samples are likely to be more informative and valuable for the model to learn from, which can improve its performance and generalisation. There are several ways to implement density-weighted methods, including using a density estimate or weighting samples based on their informativeness and density.

Deep Learning Query Strategies

The application of active learning to deep learning algorithms has been met with various challenges (Ren et al., 2021). One such challenge is the increased data requirements of deep learning models, as they must concurrently learn both representative features and a classifier. Additionally, deep neural networks often experience issues related to confidence calibration, which can hinder the effectiveness of uncertainty-based active learning approaches due to the unreliable nature of softmax predictions as a measure of certainty as demonstrated in the work from Gal and Ghahramani (2016). Within the field of deep active learning, two main categories can be identified: scoring query strategies, which select data for annotation based on a particular metric, and batch-aware query strategies, which select the optimal batch of data for annotation.

Scoring Query Strategies

To address the need for more representative data annotations, Wang et al. (2016) proposed the Cost-Effective Active Learning (CEAL) algorithm. This algorithm is an extension of the standard uncertainty-based active learning method, which selects data for annotation based on uncertainty in the model's prediction (epistemic uncertainty). CEAL integrates a process called pseudo-labelling (Lee et al., 2013), in which data with low uncertainty used their predicted labels to augment the annotated data, resulting in a more diverse training set that includes both difficult, uncertain samples that can aid in classifier improvement, as well as certain samples that contribute to the development of a more generalised feature representation. The name CEAL reflects the cost-effective nature of this active learning approach, as it allows for the incorporation of new annotations without the associated labelling costs. The effectiveness of CEAL was evaluated in comparison to supervised learning, random active learning, and Triple Criteria Active Learning (TCAL) (Demir and Bruzzone, 2014) on the Cross-Age Celebrity Dataset (Chen et al., 2014). The results showed that CEAL achieved convergence with supervised learning using only 60% of the data, outperforming both TCAL and random query methods in terms of convergence speed.

Gal and Ghahramani (2016) demonstrated that deep learning models employing softmax activation functions are unable to capture model uncertainty. Softmax-based uncertainty estimate primarily captures aleatoric uncertainty, which is connected to data randomness. It does not, however, reflect epistemic uncertainty, which results from the model's lack of knowledge or ambiguity about the data distribution. When the model hasn't seen a variety of examples or is unsure about the underlying relationships in the data, epistemic uncertainty is significant. To address this limitation, they introduced a method for approximating Bayesian inference using dropout in deep CNNs. Specifically, they applied dropout to each weight layer in the CNN during both training and testing and used the sample variance of the resulting prediction feedforward, which was repeated B times, to estimate the model uncertainty. This process is depicted in Equation (2.1), where w represents the learned weights, x denotes the input, and $\widehat{y}(b)$ is the CNN output obtained with dropout patten b applied to each layer.

$$Var(f(x,w)) \approx \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{y}(b) - \frac{1}{B}\sum \widehat{y}(b)\right)^2 \qquad (2.1)$$

Gal et al. (2017) used Bayesian CNNs to evaluate a variety of active learning query strategies, including max entropy of the Bayesian samples (Shannon, 1948), variation ratios as a measure of uncertainty across the Bayesian samples (Freeman, 1965), mean standard deviation across the Bayesian samples (Kampffmeyer et al., 2016), Bayesian active learning disagreement (BALD) (Houlsby et al., 2011), and random sampling. The results indicated that variation ratios performed the best, followed closely by BALD and max entropy, while mean standard deviation performed similarly to the random baseline. However, the authors noted that the variation ratios method did not generalise well to more complex datasets, such as the ISIC skin lesion dataset (Gutman et al., 2016). In contrast, the BALD acquisition method demonstrated improved performance on these other datasets, and the use of Bayesian CNNs overall resulted in a significant improvement in performance compared to max entropy.

The Deep-Fool active learning method, proposed by Ducoffe and Precioso (2018), employs a strategy based on margin theory for sampling unannotated data that is close to decision boundaries. This approach, which has been previously discussed in the literature (Settles, 2009), leverages adversarial attacks to determine the distance of a given data point to a decision boundary by adding small perturbations to the input image and measuring the resulting change in prediction (Kurakin et al., 2018). The query strategy subsequently selects the data closest to the decision boundary to be labelled, adding the labelled data and its adversarial counterparts to the training set. According to the authors, this method exhibits competitive performance compared to core set active learning, while also being significantly less computationally complex and more efficient.

Batch-Aware Query Strategies

Sener and Savarese (2017) observed that when using CNNs, selecting a single piece of data can be detrimental to the training process due to its minimal statistical impact. They concluded that an active learning algorithm working with a CNN should choose an optimal batch of data for annotation. To achieve this, they treated the problem as a core-set selection problem, in which a small subset of data is selected to be representative of the entire dataset. This problem is equivalent to a k-center problem (Farahani and Hekmatfar, 2009), in which a set of n points is chosen to cover all remaining data points while minimising the radius from each data point, as illustrated in Figure 2.1.

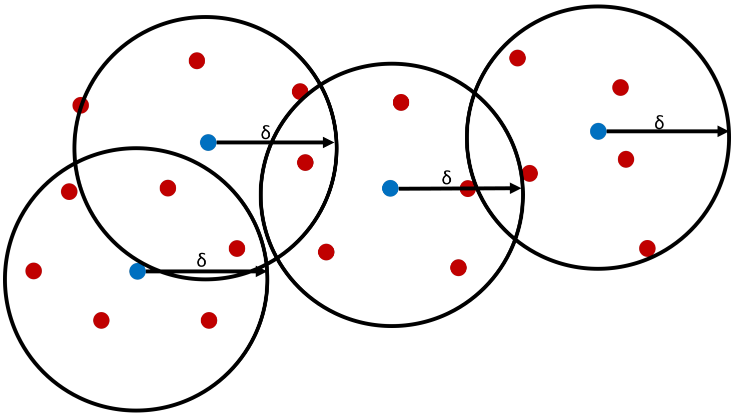


Figure 2.1: A visual illustration of the core-sets query strategy, in which four points are selected that cover all other data points and minimise δ (Sener and Savarese, 2017).

To compare their method to others, they conducted experiments with different selection methods, using Ladder Networks (Rasmus et al., 2015) as a method for weakly supervised learning, as these can be trained on unlabelled data at each iteration. They discovered that for all selection algorithms tried weakly supervised learning significantly improved model performance. Results on the CIFAR 10 and CIFAR 100 datasets (Krizhevsky et al., 2009) demonstrate that the core set method significantly outperformed the other methods evaluated.

Kirsch et al. (2019) introduced BatchBALD, a batch-aware extension of BALD (Gal et al., 2017), which selects data points that exhibit high mutual information between model parameters and model output. To achieve this, the authors modify the definition of mutual information to incorporate both the general uncertainty of the model and the expected uncertainty for a specific set of model parameters. Submodularity is then utilised to identify the optimal set of data points that maximise mutual information. The authors demonstrate that BatchBALD outperforms BALD on multiple datasets, including MNIST (LeCun et al., 1998), EMNIST (Cohen et al., 2017) and CINIC-10 (Darlow et al., 2018). However, the authors also acknowledge certain limitations of BatchBALD, including its reduced effectiveness on unbalanced datasets and the noise introduced using Monte Carlo dropout.

Zhdanov (2019) introduced Diverse mini-batch active learning (DBAL), which combines uncertainty and diversity sampling to identify an optimal batch of data. The data is first encoded and then clustered using K-means, with data points closest to the centre of each cluster being selected. To incorporate uncertainty, weighted K-means is utilised, where each data point is assigned a weight based on an uncertainty function (in this case, margin-based uncertainty, which the authors found to be particularly effective). To improve computational efficiency, they pre-filter the unlabelled data points by uncertainty and only cluster the remaining data points. The authors demonstrated that DBAL outperforms other methods when tested on multiple datasets, including the MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2009) datasets.

2.2.2 Application of Active Learning for Medicine

Active learning has the potential to significantly reduce annotation costs in medical image analysis tasks, which has very expensive annotation costs (Budd et al., 2021). This is demonstrated by the PanNuke dataset (Gamper et al., 2019, 2020), which was generated through a combination of machine learning algorithms trained on public datasets, and active learning methods to refine the annotations. Specifically, the authors used the algorithm to detect and classify nuclei in whole slide images, and then applied active learning techniques to measure the epistemic uncertainty and select the images for expert annotation. Through this approach, the authors were able to annotate 205,343 nuclei with 5 classification labels, with only 10% of the data requiring manual annotation from pathologists. This serves as a compelling example of the utility of active learning in digital pathology, enabling the efficient construction of large and detailed datasets.

The survey of deep active learning for medical image analysis by Budd et al. (2021), covers a significant body of research. This research encompasses a diverse range of medical image analysis tasks, including but not limited to MRI segmentation (Konyushkova et al., 2019, Zhao et al., 2019), skin lesion classification and segmentation (Shi et al., 2019, Gorriz et al., 2017), ultrasound classification Liu et al. (2020), and histopathology whole slide image segmentation (Folmsbee et al., 2021, Jin et al., 2021). Despite this extensive research on active learning, the majority of the literature assumes that all annotations are equally costly, with little effort being made to account for the potential variations in annotation costs.

2.2.3 Active Learning for Annotation Efficiency

Active learning is a widely researched methodology in the field of machine learning that emphasises the selective annotation of a subset of data, rather than annotating all available data, to improve the performance of machine learning algorithms while also reducing annotation costs. The traditional approach to active learning has been to assume that the costs of annotating all data are the same. However, recent research has indicated that this is not always the case in practice, with some data being more difficult to annotate than others (Settles et al., 2008). In order to address this issue, various methods have been proposed to increase annotation efficiency, particularly in the context of deep learning algorithms. One such method is the cost-effective active learning approach tailored to multi-class semantic segmentation, known as CEREALS, which was proposed by Mackowiak et al. (2018).

The Cost-Effective Region-based Active Learning for Semantic Segmentation (CEREALS) method reduces annotation costs by identifying informative image patches that have low annotation costs. The authors accomplished this by developing information maps and estimated annotation cost maps for each image, which were then fused to extract patches that maximise information while minimising annotation cost. The uncertainty for each pixel was calculated using vote entropy across Bayesian samples of the model, generated through the use of Monte Carlo dropout (Gal and Ghahramani, 2016). Based on how many clicks were necessary to annotate the image, the annotation costs were approximated. Although the approximation has limitations, the authors chose to use it because there were no other datasets with annotation costs available. Evaluation of the CEREALS method on the Cityscapes dataset (Cordts et al., 2016) showed that it achieved high mean Intersection over Union (mIoU) scores while also requiring fewer annotator clicks when compared to other active learning methods. Building upon this approach, Colling et al. (2020) proposed a similar method but modified the definition used to estimate the cost of annotating and demonstrated a reduced annotation cost compared to other algorithms that don't take into account variable annotation costs.

2.3 Annotator Efficient Active Learning for Histopathology

Annotation of medical images, particularly histopathology tasks, poses a significant challenge due to the cost and lack of expert annotators. For instance, tasks such as nuclei detection and classification necessitate annotators to meticulously select and categorise nuclei in whole slide images. A notable example of this can be observed in the annotation procedure for the CRCHistoPhenotypes dataset (Sirinukunwattana et al., 2016), in which annotators had to examine 500x500 pixel patches extracted from 100 H&E slides (0.55 μm/pixel, equivalent to 20x optical magnification) and assign each nucleus a classification of epithelial, inflammatory, fibroblast, or miscellaneous. The annotations were collected by experienced pathologists and graduate students supervised by the same pathologists, highlighting the costly nature of such annotation efforts, given the financial requirements for expert annotators. This illustrates the potential for applying active learning methods that incorporate considerations of annotation cost.

2.3.1 Tile-Based Active Learning

Patch-based approaches are widely used in digital pathology and medical image analysis. However, applying active learning to these methods can be laborious, particularly in systems that utilise small patches. These patches can be challenging to annotate in isolation, even when provided with spatial visual context. To address this, this chapter proposes a tile-based approach that requests annotations over larger tiles containing multiple small patches. Working with larger tiles reduces the burden on the annotator and can increase the annotation collection throughput. This modification enables a learning algorithm to be trained on small patches, while treating the data as larger tiles only when querying the unannotated data to be annotated (Figure 2.2).

The proposed query strategy involves a simple modification to an existing query strategy, as outlined in Algorithm 1 and visualised in Figure 2.3. In this algorithm, S represents an existing query strategy, which is called at the end of each active learning iteration once a model has been trained on the current annotated data. The algorithm extracts all patches from each unannotated tile and generates predictions for each patch, which are then averaged to obtain a prediction for the overall tile.

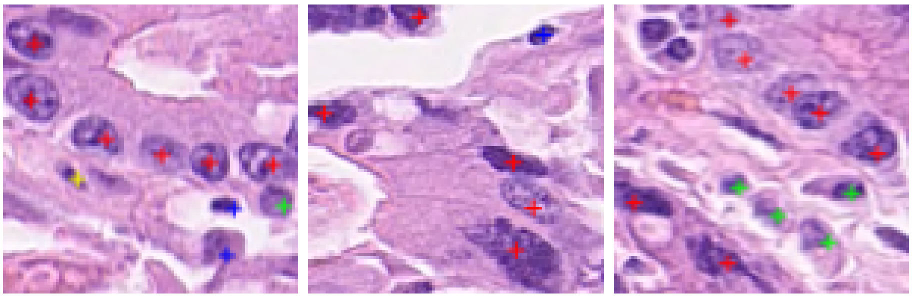


Figure 2.2: Example tiles that contain annotated nuclei to be extracted as patches and utilised for the purpose of training the model.

These tile-level predictions can then be utilised within an active learning query strategy, such as entropy uncertainty sampling, to sample tiles based on uncertainty values. This approach can also be applied to more complex query strategies, such as core-set sampling, by solving the K-centre problem for the tile predictions rather than for individual data point feature representations.

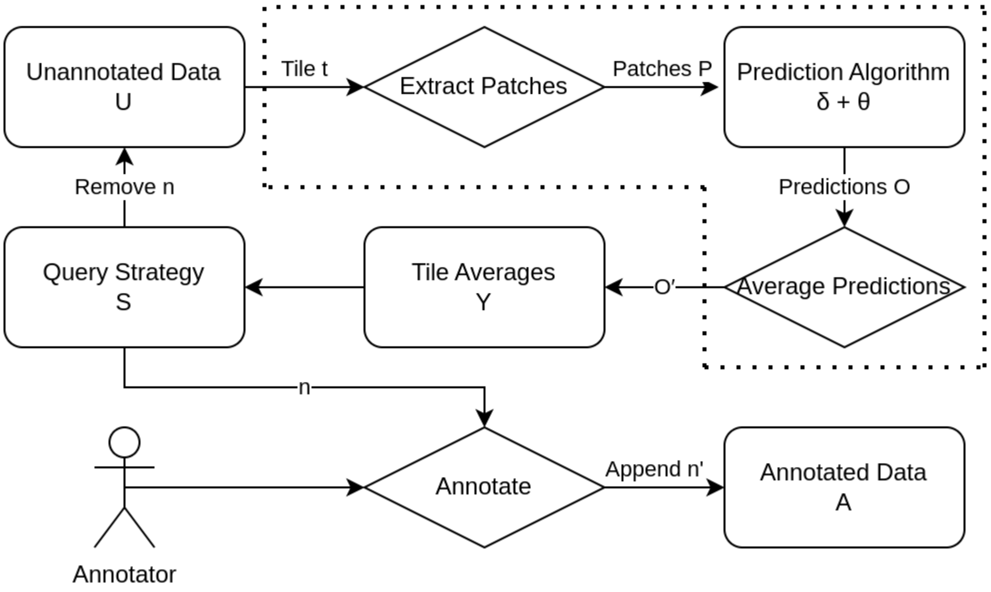


Figure 2.3: Block diagram of Algorithm 1.

Algorithm 1: Tile-based active learning

```
Input  : $\theta$ : trained model,
         $\delta$ : prediction algorithm,
         $S$ : active learning query strategy,
         $U$ : set of unannotated data,
         $A$ : set of annotated data,
         $Y$ : empty set of tile averages
Output: $U'$ : updated set of unannotated data,
         $A'$ : updated set of annotated data
TileQueryStrategy $\theta, \delta, S, U, A, Y$
    foreach tile $r$ in $U$ do
        $P \leftarrow$ ExtractPatches($r$)          extract patches from tile
        $O \leftarrow \delta(\theta, P)$             predictions on extracted patches
        $O' \leftarrow$ Average($O$)                average the patch predictions
        $Y \leftarrow$ Append($Y, O'$)              append tile average to array
    end
    $n \leftarrow S(Y)$                                  selects tile from list
    $U' \leftarrow$ Remove($U, n$)    removes selection from unannotated set
    $n' \leftarrow$ Annotate($n$)           annotator annotates the selected tile
    $A' \leftarrow$ Append($A, n'$)    appends the selection to annotated set
    return $U', A'$
```

2.4 Active Learning Experiments

This section describes the datasets, training parameters, experimental setup, and results for the experiments on nuclei classification using tile-based active learning on whole slide image patches. The code and complete results for these experiments are available in the project GitHub repository[1].

2.4.1 Dataset

The publicly available CRCHistoPhenotypes dataset (Sirinukunwattana et al., 2016) was used to evaluate the proposed tile-based active learning approach, as it consists of a large number of annotated nuclei from large patches extracted from whole slide images, making it well-suited for this type of active learning. The dataset, which has also been utilised in multiple nuclei classification and detection studies, as well as active

[1] GitHub Repository: `github.com/jmcjacob/Patch-Active-Learning-Pathology`

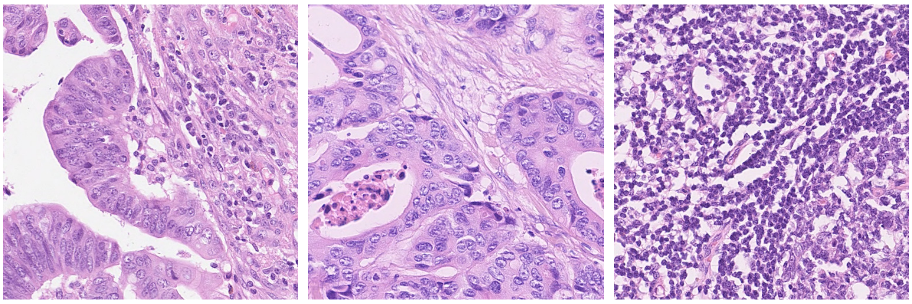


Figure 2.4: Three example images from the CRCHistoPhenotypes dataset Sirinukunwattana et al. (2016) are shown, each containing multiple nuclei that will be extracted into patches and augmented.

learning experiments (Shao et al., 2018), comprises 22,444 annotated nuclei from 100 500x500 pixel non-overlapping patches extracted from 10 H&E whole-slide images of colorectal adenocarcinomas from 9 different patients. Each nucleus is annotated with its coordinates and corresponding classification (epithelial, inflammatory, fibroblast, or miscellaneous). The dataset includes 7,722 epithelial, 5,712 fibroblast, 6,971 inflammatory, and 2,039 miscellaneous annotated nuclei. 2,500 images were generated by dividing the patches into 100x100 pixel tiles, from which each nucleus was extracted into a 30x30 patch for use in training the CNN model. During training, data augmentation was employed by randomly applying Gaussian blurring and horizontal and vertical flipping.

2.4.2 Training Parameters

This experiment employed a CNN inspired by the architecture used in the nuclei classification benchmark for the CRCHistoPhenotypes dataset (Sirinukunwattana et al., 2016). The network consisted of two convolutional layers, the first with 36 4x4 filters and the second with 48 3x3 filters, each followed by a 2x2 max pooling layer. The convolutional layers were followed by two fully connected layers with 1200 and 512 neurons, respectively. This architecture is summarised in Table 2.1. Each hidden layer utilised Rectified Linear Unit (ReLU) activation functions, while the fully connected layers employed dropout regularisation (Srivastava et al., 2014) with a drop rate of 0.5. Dropout was utilised during training to enable the use of Monte Carlo sampling after training.

During the active learning process, the training environment is continuously evolving as the training data is incrementally expanded. To accommodate this dynamic situation, the Adadelta adaptive gradient

Table 2.1: The Convolutional Neural Network architecture employed for nuclei classification in the tile-based active learning experiments.

Type	Filter Dimensions	Input/Output Dimensions
Input		30 x 30 x 3
Convolutional	4 x 4 x 1 x 36	26 x 26 x 36
Max Pooling	2 x 2	12 x 12 x 36
Convolutional	3 x 3 x 36 x 48	10 x 10 x 48
Max Pooling	2 x 2	5 x 5 x 48
Fully Connected	5 x 5 x 48 x 1200	1 x 1200
Fully Connected	1 x 1 x 512 x 512	1 x 512
Output	1 x 1 x 512 x 4	1 x 4

descent algorithm (Zeiler, 2012) was chosen, as it does not require manual tuning of the learning rate as it adjusts automatically to the gradients of the model. To prevent overfitting to the constantly changing training data, which may be limited in size, early stopping was employed. The approach proposed by Prechelt (2012) compares the generalisation loss and training progression until a specified threshold of $\frac{GL(t)}{P_k(t)} > \alpha$ is reached. The generalisation loss (Equation 2.2) is calculated by comparing the validation loss at each epoch $L_{val}(t)$ to the minimum validation loss across all epochs, while the training progression (Equation 2.3) is calculated by analysing the training losses $L_{tr}(t)$ over a batch of recent epochs of size k.

$$GL(t) = 100 \cdot \left(\frac{L_{va}(t)}{\underset{t' \leq t}{min} L_{va}(t')} - 1 \right) \tag{2.2}$$

$$P_k(t) = 1000 \cdot \left(\frac{\sum_{t'=t-k+1}^{t} L_{tr}(t')}{k \cdot min_{t'=t-k+1}^{t} L_{tr}(t')} - 1 \right) \tag{2.3}$$

2.4.3 Experiment Setup

A series of experiments were conducted to evaluate the effectiveness of the proposed tile-based active learning approach in conjunction with various query strategies. The experiments exclusively employed a tile-based active learning framework due to the inherent challenge of directly comparing it with the traditional patch-based active learning. The query strategies utilised in these two approaches are not directly comparable, necessitating a more comprehensive experimental design, which would be better suited for future research endeavours. The query strategies evaluated included several basic methods, which served as baselines, as well as deep learning-specific query algorithms. The baseline strategies included random querying, least confident uncertainty, margin uncertainty, and entropy uncertainty sampling. The deep learning-specific query strategies tested were K-Centre sampling (employing greedy approximation), core-set sampling (Sener and Savarese, 2017), and BALD using Monte Carlo dropout (Gal et al., 2017). These methods were selected due to their state-of-the-art status in the field of deep active learning.

In each experiment, all available data was initially treated as unannotated, and then two randomly selected tiles were annotated to form an initial training dataset. During each active learning iteration, two additional tiles were chosen from the pool of unannotated tiles and added to the training dataset. A randomly initialised model (using uniform Xavier initialisation, as proposed by Glorot and Bengio (2010)) was then trained on the updated dataset. This process was repeated for 50 active learning iterations, resulting in a final training set of 102 tiles out of 2,500 in each experiment. To account for random variations, each experimental condition was run five times, with different seeds used to generate random elements such as model weight initialisation and initial annotated patches.

2.4.4 Results

The performance of the various query strategies was assessed by evaluating the trained models on a fixed test set after each active learning iteration. Table 2.2 presents the mean test accuracy and cross-entropy loss over five runs, after 50 iterations, for each of the query strategies. As shown, only the K-Centre sampling approach yielded higher average accuracy than a random sampling baseline. The core-set sampling strategy produced results that were similar to those of the random sampling approach, while the other query strategies all performed worse.

Figures 2.5, 2.6, and 2.7 illustrate the test accuracy, mean class accuracy, and cross-entropy loss for models trained with annotated data selected using each of the query strategies after each active learning iteration, averaged over five runs. The figures also include the results for a model trained using all 1487 annotated training tiles; a fully supervised CNN trained on the entire annotated dataset achieved an accuracy of

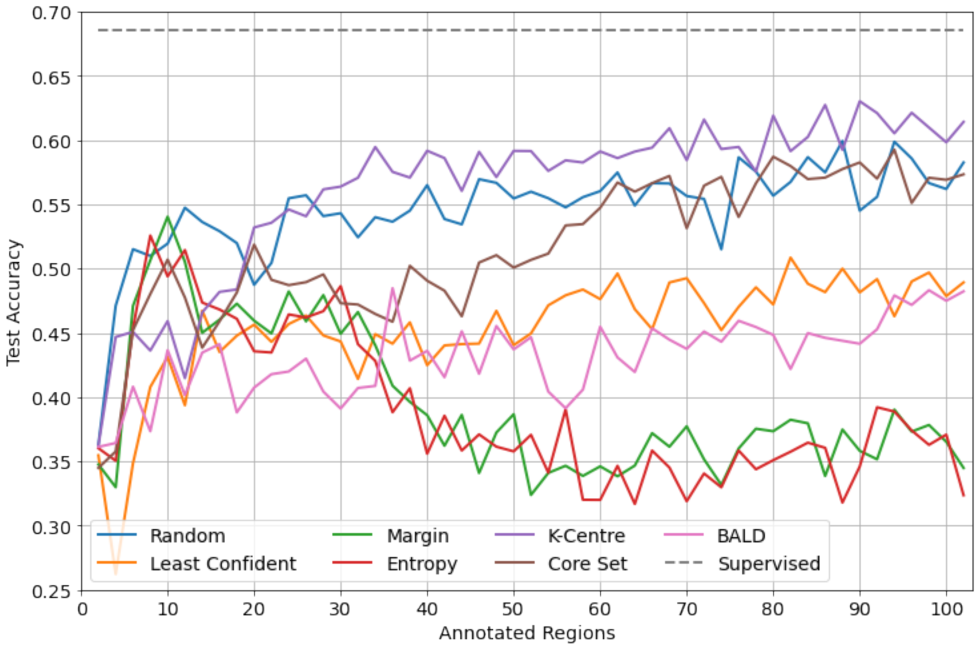


Figure 2.5: Average test accuracy of trained models using different amounts of annotated tiles selected through various query strategies.

68.53% and a cross-entropy loss of 1.111. In contrast, the model trained with the K-Centre query strategy achieved an accuracy of 61.41% and a cross-entropy loss of 1.137, using only 7% of the annotated data.

2.5 Conclusion

This chapter presents a method for mitigating the annotation burden in patch-based nuclei classification systems using deep active learning. The results reported in Section 2.4.4 indicate that traditional active learning approaches are less effective when applied to deep learning models, and that specialised active learning techniques for deep learning also fail to outperform random sampling baselines. This phenomenon, which has

Table 2.2: Test results for each query strategy after 50 active iterations. For Accuracy and Mean Class Accuracy higher is better and Loss lower is better.

Query Strategy	Random	Least Confident	Margin	Entropy	K-Centre	Core-Set	BALD
Accuracy	58.25	48.92	45.84	32.37	**61.41**	57.33	48.23
Mean Class Accuracy	53.50	47.36	42.40	41.07	**54.39**	52.50	46.14
Loss	1.154	1.243	1.268	1.39	**1.123**	1.157	1.247

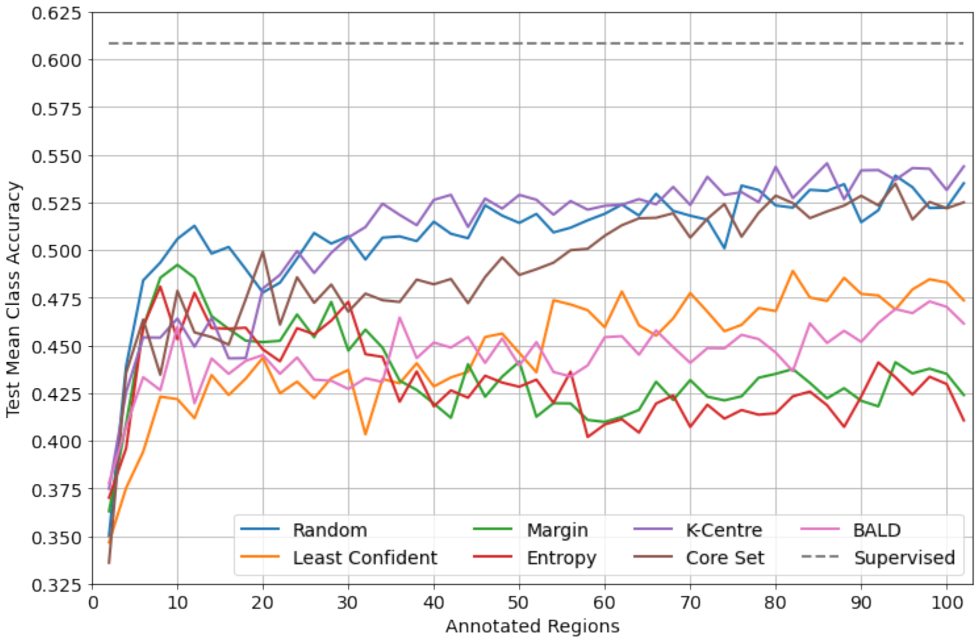


Figure 2.6: Average test mean class accuracy of trained models using different amounts of annotated tiles selected through various query strategies.

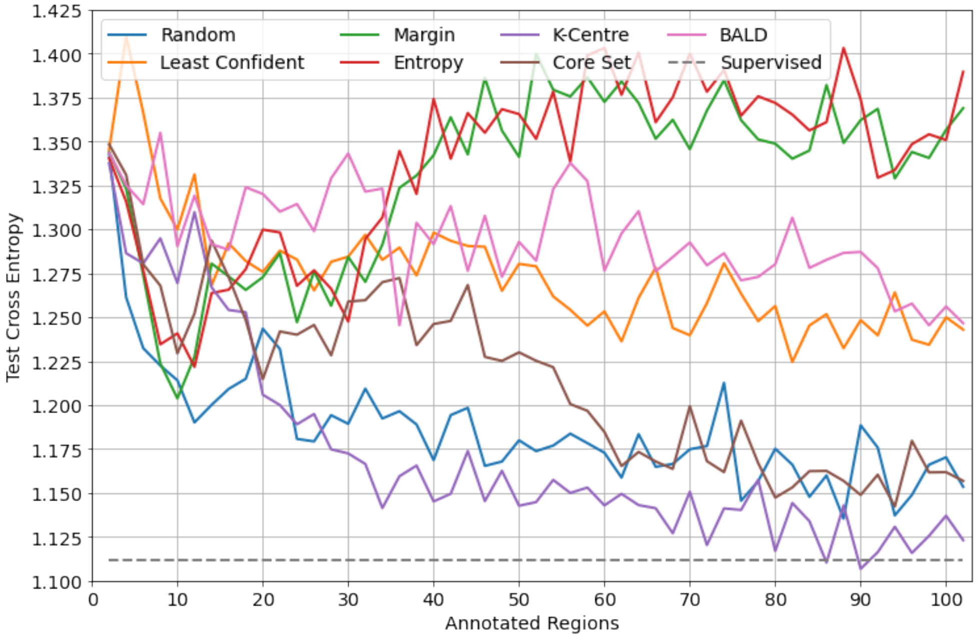


Figure 2.7: Average test loss of trained models using different amounts of annotated tiles selected through various query strategies.

been previously noted in the literature on active learning for deep learning (Ren et al., 2021), highlights the need for more robust active learning methods in this domain.

The phenomenon under consideration can be elucidated as follows: traditional query strategies, which prioritise sampling challenging examples, have resulted in the final annotated dataset employed for training lacking representativeness with respect to the overall dataset distribution. Consequently, for a deep learning model, it becomes essential to train on a dataset that is not only representative but also informative. This is because deep learning models simultaneously learn representative features and classifiers. On the other hand, conventional machine learning methods would gain greater advantage from adopting such a strategy when employing static hand-crafted features.

The most effective query strategies observed in this chapter were the k-centre and core set approaches. These strategies were designed to ensure that the selected batches of samples maintain both representativeness and informativeness. Such an approach is particularly well-suited for neural networks, as it facilitates the acquisition of representative features along with a discriminative classifier. However, it is noteworthy that even though these intelligent query strategies demonstrated superior performance, the baseline approach, employing random sampling for querying, yielded comparable results. This finding suggests that the limited size of the dataset negatively impacts the overall model performance, despite the utilisation of sophisticated query strategies.

To address this limitation and enhance the model's performance, two potential solutions can be considered. First, incorporating a supplemental strategy like CEAL (Wang et al., 2016). Alternatively, leveraging unsupervised representative learning prior to active learning can be advantageous. This entails training the model on the unannotated dataset to learn feature representations, which can then be utilised during the active learning phase, potentially improving the model's ability to generalise and perform better in a data-limited scenario.

Reducing the cost of annotating data is crucial for enabling the development of deep learning systems for digital pathology and medical image analysis, particularly for organisations with limited resources. While active learning holds promise as a means of addressing this challenge, further research is required to achieve meaningful improvements on tasks such as those presented in this chapter. This has motivated the investigation of unsupervised learning techniques as a complementary approach for leveraging unannotated data, potentially in conjunction with active learning.

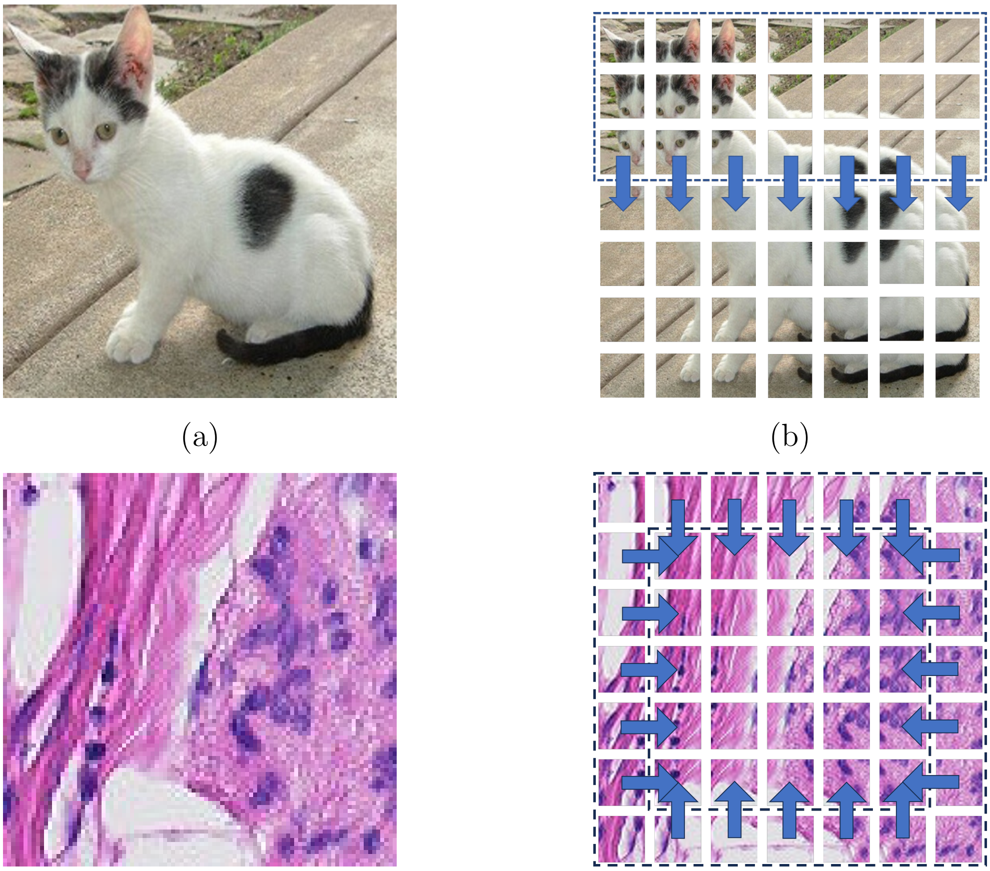
Chapter 3

Unsupervised Representation Learning

3.1 Introduction

The application of modern deep learning algorithms has demonstrated significant improvements in digital pathology tasks such as nuclei detection and disease classification (Litjens et al., 2017), as previously discussed in Section 2.1.1. The ability to jointly learn deep representations and discriminative classifiers or regressors through end-to-end training allows for feature representations that are specifically tailored to a given task. However, this approach necessitates a significant amount of annotated data for adequate generalisation, which poses a major challenge for digital pathology (Madabhushi and Lee, 2016) and other medical image analysis domains. In an effort to address this challenge, we previously explored the use of active learning (Chapter 2), however, the limitations of this approach were highlighted in Section 2.5 and subsequently led to a focus on unsupervised representation learning as a means to extract information from unannotated images. Unsupervised representation learning can be utilised to improve generalisation, decrease data dimensionality, improve computational performance, and initialise deep supervised learning models when access to annotated data is limited (Bengio, Courville and Vincent, 2013).

One approach to reducing the need for large, annotated datasets is through the use of unsupervised representation learning and transfer learning. This is accomplished by using the weights of a deep encoder, trained on a large pool of unannotated data, to initialise another model (Weiss et al., 2016). Contrastive predictive coding (CPC) is a



(a)

(b)

(c)

(d)

Figure 3.1: (a) An example image from the ImageNet dataset (Deng et al., 2009). (b) Extracted overlapping patches with those used to produce context and autoregressor direction highlighted. (c) An example image from the Patch Camelyon dataset (Veeling et al., 2018). (d) Extracted overlapping patches with those used to produce context and autoregressor direction highlighted.

state-of-the-art method for unsupervised representation learning (van den Oord et al., 2018). It involves training an autoregressive model to predict future data representations in a sequence, using a loss function with noise-contrastive estimation (NCE) and importance sampling components, to preserve the density ratio between each sample and its representation. Although originally developed for sequential data, CPC has been adapted for images by splitting each image into overlapping patches and using an encoder to produce a matrix of feature representations. A mask is then applied to the matrix so that an autoregressive model can only see a subset of the feature representations in order to predict the representations of the masked patches from the context available to it. This framework has been applied successfully to object detection and Imagenet classification tasks with modifications to model capacity, layer normalisation, prediction directions, and patch-based augmentations (Henaff, 2020). However, in previous implementations, the autoregressive model's predictions were made in multiple directions individually, which can be inefficient when dealing with images, such as in digital pathology, where the orientation of the image is arbitrary and does not carry useful information. Figure 3.1(b) illustrates an example of this framework applied to an image.

The current chapter builds upon the idea that unsupervised representation learning can be utilised to learn deep representations, which can then be used in conjunction with transfer learning to train a discriminative classifier with limited annotated data (as depicted in Figure 3.2). By implementing this approach, the need for complex deep learning-specific active learning query strategies is mitigated and instead allows for a focus on uncertainty-based querying. In order to realise this approach, a state-of-the-art unsupervised representation learning algorithm for digital pathology images was required. To this end, a multi-directional CPC

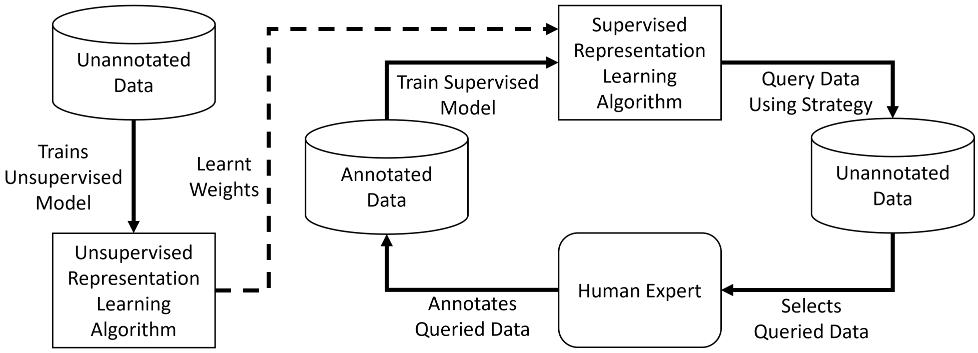

Unannotated Data
Trains Unsupervised Model
Unsupervised Representation Learning Algorithm
Learnt Weights
Supervised Representation Learning Algorithm
Train Supervised Model
Query Data Using Strategy
Annotated Data
Unannotated Data
Human Expert
Annotates Queried Data
Selects Queried Data

Figure 3.2: Proposed active learning framework with learnt representations from unsupervised representation learning on unannotated data.

extension was proposed, which includes an alternative mask for building latent context and a new extension to the autoregressive model PixelCNN (van den Oord et al., 2016) for multi-directional predictions (as depicted in Figure 3.1(d)). The effectiveness of this modification was demonstrated using the PatchCamelyon dataset (Veeling et al., 2018), derived from the Camelyon16 dataset (Litjens et al., 2018), where it was shown that classification can be performed with less annotated data when utilising representations learned in this way.

This work was presented at the IEEE International Symposium on Biomedical Imaging 2021 and published as part of its proceedings (Carse, Carey and McKenna, 2021).

3.2 Unsupervised Representation Learning for Computer Vision

This is a brief review of deep unsupervised representation methods for computer vision applications covering generative/reconstructive and self-supervised methods. In the field of deep learning, methods have shown exceptional performance in tasks involving abundant annotated data. However, their performance is known to suffer in scenarios where the supervision is limited. One solution to this issue is to utilise unsupervised learning techniques to learn highly structured data representations, which can lead to more data-efficient models (Lake et al., 2015).

Deep learning models typically consist of layers that are used for specific tasks, such as classification or regression, and others that are used to encode the data into feature representations, known as encoders. Unsupervised representation learning in deep learning focuses on learning the parameters of the encoder. Once trained, the encoder can be used for transfer learning and applied to tasks such as classification, object detection, or segmentation (Weiss et al., 2016).

In the context of computer vision, there are two main approaches to deep representation learning: generative methods and self-supervised learning (Bengio, Courville and Vincent, 2013). Generative methods aim to learn representative feature encodings by attempting to reconstruct images. On the other hand, self-supervised learning involves training models in a supervised manner using auto-generated annotations for classification or regression tasks.

3.2.1 Generative and Reconstructive Methods

Unsupervised representations can be learned through reconstructive methods by encoding input data into a lower-dimensional latent space and subsequently reconstructing the original data from the latent representation. The optimisation of encoder and decoder parameters is achieved through the use of a reconstruction error as a learning signal. In contrast, generative models are capable of generating new data samples that resemble the training data. Specifically, within the realm of medical image analysis, generative models have been employed for image-to-image translation tasks such as converting images from one modality to another (Kaji and Kida, 2019). The surge in popularity of generative models in medical image analysis can be attributed to their ability to learn useful representations from vast quantities of unannotated data, which can then be leveraged to enhance performance on tasks such as segmentation and classification (Yi et al., 2019).

Autoencoders

Autoencoders are a class of reconstructive model that employ an encoder to reduce the dimensionality of the input data and a decoder to reconstruct the original data from the reduced representation (Kramer, 1991). Transfer learning can be achieved by utilising the parameters learned by the encoder and training a new classifier with the encoded representations. One of the most widely used types of autoencoders is the undercomplete autoencoder, which is trained to minimise the reconstruction error by learning a compressed feature representation (Goodfellow et al., 2016). Other variations of autoencoders include sparse autoencoders, which are designed to learn sparse representations that have been shown to improve performance when used for transfer learning (Makhzani and Frey, 2013). Denoising autoencoders, on the other hand, are trained by corrupting the input data with noise and the goal of the network is to reconstruct the original data without the added noise. This denoising training process forces the encoder and decoder to implicitly learn the underlying structure of the data, which can be beneficial for transfer learning (Bengio, Yao, Alain and Vincent, 2013).

The final category of autoencoder is the regularised autoencoder, which modifies the traditional autoencoder architecture in order to enhance the model's ability to learn more informative representations and capture relevant information. One of the most widely employed forms of regularised autoencoder is the variational autoencoder, as proposed by Kingma and Welling (2013). This variant of autoencoder ensures that the latent space possesses desirable properties that enable a generative process by encoding an input image as a distribution over the latent space, as opposed to encoding it into a single point in the latent space. A variational autoencoder is trained by sampling from the encoded latent distribution and decoding it into an output image. The loss function for this model is based on the reconstruction error of the output image and the Kulback-Leibler divergence (Kullback and Leibler, 1951) as the regularisation term.

Most recent work with autoencoders in medical image analysis has been using variational autoencoders as they have shown a high level or performance and have been used for multiple tasks (Wei and Mahmood, 2020). Akrami et al. (2020) used a combination of variational autoencoders and transfer learning to build unsupervised lesion detection models for MRI brain scans images and showed their robustness when working with training and testing datasets with different parameters (Thiagarajan et al., 2020).

Generative Adversarial Networks

Generative adversarial networks (Goodfellow et al., 2014) are a class of deep learning models that are designed to generate new samples of data that resemble existing samples from a given dataset. They consist of two main components: a generator network, which produces new samples, and a discriminator network, which attempts to distinguish the generated samples from the real samples. The two networks are trained in an adversarial manner, with the generator attempting to produce samples that the discriminator cannot distinguish from real samples, and the discriminator attempting to correctly identify the generated samples. However, generative adversarial networks are known to be difficult to train, one common problem when training generative adversarial networks is mode collapse, where the generator produces a limited set of outputs that fail to capture the full diversity of the real data distribution. Several techniques, such as Wasserstein generative adversarial networks (Arjovsky et al., 2017) and gradient penalty (Gulrajani et al., 2017) have been proposed to stabilise the training process.

In the context of representation learning, generative adversarial networks can be used to learn a compact, low-dimensional representation of the data that captures the underlying structure of the dataset. This can be achieved by training the generator to produce samples that are similar to the real samples, but in a lower-dimensional space. The generator can then be used as a feature extractor, mapping the real samples to their corresponding low-dimensional representations. Additionally, the discriminator can be used as a classifier, allowing the learned representations to be used for downstream tasks such as classification (Srivastav et al., 2021) or clustering (Mukherjee et al., 2019).

A variant of generative adversarial networks is BigBiGAN (Donahue and Simonyan, 2019), which is designed to generate high-quality images from a compact, low-dimensional representation of the data. It is based on the idea of BiGAN (Donahue et al., 2016) and the primary difference between BigBiGAN and the original BiGAN is the scale of the model. BigBiGAN is trained on a large dataset such as ImageNet, which contains millions of images, whereas BiGAN is trained on smaller datasets. The increased size of the dataset allows BigBiGAN to learn a more powerful and expressive representation of the data. BigBiGAN consists of three main components: an encoder network, a generator network, and a discriminator network. The encoder network maps the input data to a low-dimensional representation, which is then used as input to the generator network to produce a reconstructed sample.

3.2.2 Self-supervised Methods

Self-supervised learning for representation learning is a subset of machine learning that aims to learn meaningful representations of data without the need for explicit annotations. A prevalent method for self-supervised representation learning in medical imaging is to employ a pretext task, a task that is simple to solve through the utilisation of the features learned by the model, but also serves as a means to learn relevant features for the ultimate task of interest. The utilisation of self-supervised pre-training techniques is rapidly gaining popularity in medical image analysis, as it enables the learning of useful features from large amounts of unannotated data, which can then be applied to enhance performance in tasks such as segmentation and classification (Shurrab and Duwairi, 2022).

RotNet

RotNet, as proposed by Gidaris et al. (2018), is a novel unsupervised representation learning technique that utilises rotation as a self-supervised task for learning useful representations of images. The fundamental concept behind RotNet is that by training a neural network to predict the rotation of an image, the model can learn features that can be utilised in other tasks such as image classification. The RotNet model comprises of a convolutional neural network that takes an image as input and predicts the angle of rotation. The output of the last convolutional layer, prior to the fully connected layers, is used as the representation of the image. This technique enables the model to train on a large dataset of unannotated images, which can then be utilised for other tasks such as image classification as demonstrated by Zhou et al. (2021).

Deep Clustering

DeepCluster, as proposed by Caron et al. (2018), is a technique for unsupervised representation learning that combines the utilisation of clustering algorithms as a form of self-supervision to train deep neural networks. The method starts by initialising the weights of a deep neural network randomly and subsequently utilising the network to extract features from a dataset of unannotated images. These features are then employed to cluster the images into different groups using a clustering algorithm, such as k-means. Upon completion of the clustering process, the annotations generated by the clustering algorithm are used as pseudo-annotations for the images. The neural network is then fine-tuned using these pseudo-annotations to enhance the representations of the images. This process is repeated multiple times, with the neural network being fine-tuned using the updated pseudo-annotations generated by the clustering algorithm.

Non-Parametric Instance-Level Discrimination

Non-parametric instance-level discrimination (Wu et al., 2018) is an unsupervised representation learning approach that prioritises the learning of feature representations capable of capturing similarity among instances, rather than similarity between classes. To accomplish this, the method frames the learning task as a non-parametric classification problem at the instance level, in which discrimination is focused on individual instances. To train the model, the authors utilised the NCE loss function (Gutmann and Hyvärinen, 2010) to address computational challenges associated with the large number of instances. Furthermore, they employed a proximal regularisation method (Parikh et al., 2014) to promote smoothness during the training process. The effectiveness of this approach was demonstrated by achieving state-of-the-art performance compared to other recent methods across different CNN models, using considerably less data. Specifically, the method achieved 10% higher top-5 accuracy on ImageNet (Deng et al., 2009) with only 1% of the data.

Local Aggregation

In their work, Zhuang et al. (2019) proposed a method for unsupervised representation learning using local non-parametric aggregation in the latent feature space. The proposed method is based on training an encoder to produce latent features, and encouraging similar data instances to move together and dissimilar instances to separate in the latent feature space. To achieve this, the method employs a clustering technique. Specifically, multiple passes of k-means clustering are performed for each latent encoding to determine its close neighbours and background neighbours. The loss function used to train the model is based on the negative log-likelihood of an encoding being recognised as a close neighbour, given that it is recognised as a background neighbour. The effectiveness of the proposed method is demonstrated through experiments on the Imagenet dataset (Deng et al., 2009). The results show that the model trained with this method exhibits substantial ability to recognise high-level visual context without the need for any dataset annotations. This highlights the potential of the proposed method for unsupervised representation learning in computer vision tasks.

Momentum Contrast

He et al. (2020) proposed a method for unsupervised representation learning called momentum contrast. This approach leverages dynamic dictionaries and employs a contrastive loss framework to yield notable advancements in unsupervised learning within the domain of computer vision. Momentum contrast's use of dynamic dictionary in contrast to static dictionaries continually evolves to encompass key representations drawn from a queue of data samples. This adaptability enables the method to effectively navigate the intricate expanse of high-dimensional visual space, capturing intricate relationships between features and patterns of the underlying data distribution. A contrastive loss function is used during training that represents the alignment between an image representation and key representation from the dynamic dictionary. By minimising this distance, the encoder is induced to acquire discriminative features that can be used to obtain superior performance in downstream tasks. This is supported by the experimental results presented in the work, which show that the learned feature representations can be used to pre-train encoders for a variety of tasks such as classification, detection, and segmentation. This highlights the potential of the momentum contrast method for unsupervised representation learning in computer vision tasks.

Pretext-Invariant Representations

Misra and Maaten (2020) illuminates the drawbacks intrinsic to the utilisation of pretext tasks for self-supervised representation learning. Usually pretext tasks are instructed to solve a specific task that lacks correlation with the downstream tasks and instead of improving performance for downstream tasks may lead to overfitting and constrained generalisation. To address this issue, the authors proposed a novel approach called pretext-invariant representation learning. This method involves applying a transformation, such as a rotation, to an input image, and then encoding both the original and transformed images using a shared encoder. The final embeddings are generated using separate prediction heads. To further improve the robustness of the learned representations, the authors employed a noise contrastive estimator (Gutmann and Hyvärinen, 2010) as the loss function, which aims to reduce the similarity between the original and transformed image's feature representation, while maximising the similarity between the transformed image and randomly sampled negative images.

In their experiments the authors used a ResNet (He et al., 2016) architecture to function as the encoder for representation generation. To substantiate the effectiveness of their approach, the authors subject their method to a comparative evaluation against a jigsaw pretext task (Noroozi and Favaro, 2016). The jigsaw pretext task entails the arrangement of shuffled patches within an image, with its objective being the capture of spatial relationships and contextual cues. The pretext-invariant representation learning technique surpasses the performance of this pretext task, attaining state-of-the-art outcomes.

Deep InfoMax

Deep infomax, first proposed in (Hjelm et al., 2018), is a self-supervised deep learning approach for learning compact and informative representations of data through the optimisation of mutual information between the data and the learned representations. Mutual information is a measure of the dependence between two random variables and quantifies the amount of information that one variable contains about the other. In the case of deep infomax, the data is considered as one random variable and the learned representation as the other. The mutual information between the data and the learned representation is estimated using a variational lower bound.

An extension of this approach, Augmented Multiscale Deep InfoMax (AMDIM) (Bachman et al., 2019), aims to learn hierarchical representations of data by utilising a multiscale architecture. This architecture is composed of multiple sub-networks, each responsible for learning representations at a different scale. The decoder network also comprises of multiple sub-networks, each responsible for reconstructing the data from the representations at a corresponding scale. Like deep infomax, the mutual information between the data and the learned representations is estimated using a variational lower bound. The key difference between deep infomax and AMDIM is that the latter utilises multiple scales to learn hierarchical representations, while the former only uses one scale. The authors of the paper have demonstrated state-of-the-art results using transfer learning to train classifiers.

3.2.3 Unsupervised Representation Learning for Medical Image Analysis

In the field of digital pathology, transfer learning is a widely used technique for various tasks (Srinidhi et al., 2021), as it has been shown to accelerate the convergence of deep learning models (Bayramoglu and Heikkilä, 2016). Studies have also demonstrated the effectiveness of utilising representation learning to initialise a CNN in situations where the learning task may be challenging due to a lack of annotated data (Hou et al., 2016). This approach can also be extended to active learning, where limited annotations make it difficult to learn features effectively (Carse and McKenna, 2019).

Instead of relying on features trained on general computer vision data, some researchers have explored the use of more general histology features for cell-level tasks (Hu et al., 2018). One method for achieving this is by training a unified generative adversarial network with a modified loss function for cell-level representation learning. Another approach is to use multi-scale convolutional sparse coding, which aims to jointly learn features at different scales with enforced scale-specificity (Chang et al., 2017).

Many of the methods under consideration depend on utilising an image's inherent orientation. However, this approach can prove challenging in instances where an image lacks a natural orientation, such as with histopathology patches or dermoscopic skin lesions. Consequently, there is a need to modify existing methodologies to better accommodate these types of medical images.

3.3 Multi-Directional Contrastive Predictive Coding

3.3.1 Contrastive Predictive Coding

CPC is an unsupervised method of feature representation that effectively extracts information from sequential data (van den Oord et al., 2018). The fundamental principle of CPC is to learn 'slow representations' that accurately capture the input data distribution over an extended period, rather than focusing on low-level, local representations. This is achieved by encoding a target variable, denoted by x, and a context variable, denoted by c, into compact distribution vector representations that maximally preserve information, as shown in Equation (3.1). By maximising $I(x;c)$ between the encoded representations, the underlying latent variables between inputs can be extracted.

$$I(x;c) = \sum_{x,c} p(x,c) log \frac{p(x|c)}{x} \tag{3.1}$$

A CPC model comprises two essential components, namely an encoder and an autoregressive model. Specifically, the encoder function, denoted by g_{enc}, is responsible for encoding each element x_t of the input sequence X into latent representations at time step t, which are represented as $z_t = g_{enc}(x_t)$. On the other hand, the autoregressive model function, g_{ar}, summarises the elements of the latent representation sequence up to a including t, $z_{\leq t}$, into a latent context representation, which is represented as $c_t = g_{ar}(z_{\leq t})$. Rather than utilising the autoregressive model to predict future samples, a density ratio is modelled to maintain the mutual information between x_{t+k} and c_t, as elucidated in Equation (3.2). By utilising density ratios, the model avoids the need to learn from the high-dimensional input distribution.

$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})} \tag{3.2}$$

The direct evaluation of $p(x)$ or $p(x|c)$ is not feasible, and therefore, it is necessary to approximate these values using sampling techniques such as NCE (Gutmann and Hyvärinen, 2010) or importance sampling. The joint training of encoder and autoregressive models using NCE with importance sampling can be facilitated by the InfoNCE loss function (Equation (3.3)) (van den Oord et al., 2018), which was introduced by a previous study. Specifically, the InfoNCE loss function involves optimising the density ratio that maintains the mutual information between the context vector c_t and future observations, given a set $X = \{x_1, \ldots, x_N\}$ consisting of one positive sample from $p(x_{t+k}|c_t)$ and N negative samples from the proposed distribution $p(x_{t+k})$. The InfoNCE loss is defined as the cross-entropy of classifying the positive sample correctly.

$$L_N = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right] \tag{3.3}$$

3.3.2 Contrastive Predictive Coding for Computer Vision

CPC is a method originally proposed for sequential data, and its application to computer vision involves first dividing an input image into overlapping patches. Each patch is then encoded, and an autoregressive model is employed to generate a context vector from the patch representations at the top of the image (as depicted in Figure 3.3, where the top 3 rows of patches were utilised (van den Oord et al., 2018)). This approach treats each column of the image as a sequence, with the context vector from the top of the image used to model the density ratio with patch representations below. This technique has been demonstrated to achieve data-efficient results on high-level computer vision classification datasets, such as Imagenet (Deng et al., 2009).

3.3.3 Pixel Convolutional Neural Network

PixelRNNs and PixelCNNs, as described by van den Oord et al. (2016), represent two classes of generative neural networks that excel in creating images through a sequential prediction process. These models leverage their capacity for sequential generation to create images by progressively

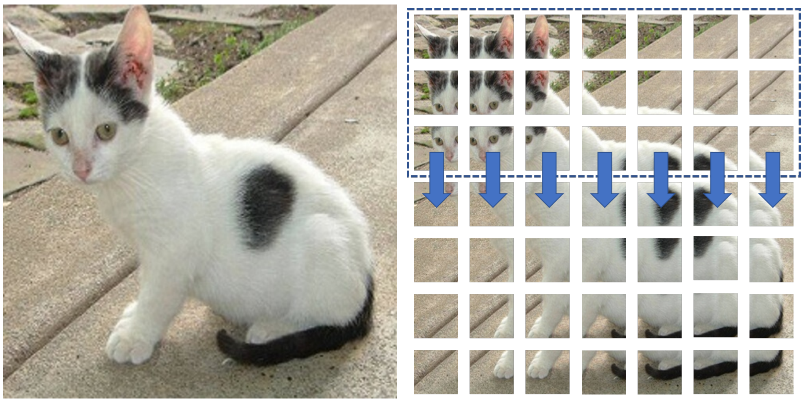


Figure 3.3: Example of how CPC works by dividing input images into overlapping patches and the columns of patches are then treated as sequences and tasked with predicted the patch representations with the top three rows being using as the context vector.

predicting individual pixels across two spatial dimensions. In essence, they build images one pixel at a time, capturing the intricate interdependencies present in the pixel data. Both PixelRNNs and PixelCNNs are designed with a primary objective: to model the discrete probability distribution of raw pixel values within images. This characteristic makes them adept at comprehensively capturing the underlying statistical properties of image datasets. By considering the raw pixel values directly, these models are able to capture both low-level details and high-level structures that constitute the visual content of images.

PixelCNNs, in particular, are notable for their utilisation of autoregressive connections in the image generation process. This distinctive feature allows PixelCNNs to meticulously model images on a pixel-by-pixel basis, enabling the decomposition of the complex joint distribution of image data into a product of conditional distributions. This decomposition is pivotal in achieving the generation of images that closely resemble the training data. The autoregressive approach ensures that each pixel's generation depends only on previously generated pixels, effectively capturing intricate patterns and textures characteristic of the training images.

An aspect where PixelCNNs hold a advantage over PixelRNNs lies in training efficiency. This efficiency stems from the inherent parallelisability of convolutional operations, which are central to the PixelCNN architecture. Convolutional layers facilitate simultaneous processing of image regions, leading to substantially faster training times when dealing with sizeable image datasets. This efficiency contributes to the popularity and efficacy of PixelCNNs in the realm of generative image modelling.

3.3.4 Multi-Directional Contrastive Predictive Coding

The approach of treating columns of the representation matrix as individual sequences (as described in Section 3.3.2), can negatively impact performance when working with images where image orientation is irrelevant, such as histology patches or dermoscopic skin lesion images. While the orientation of certain histology whole slide images can be biologically meaningful, the orientation of image patches, such as those shown in Figure 3.1(c) (sentinel lymph node), is not important. In such cases, the autoregressive model can struggle to predict patch representation from the provided context as the vertical image axis is arbitrary, unlike Imagenet where it correlates with the direction of gravity acting upon the image content.

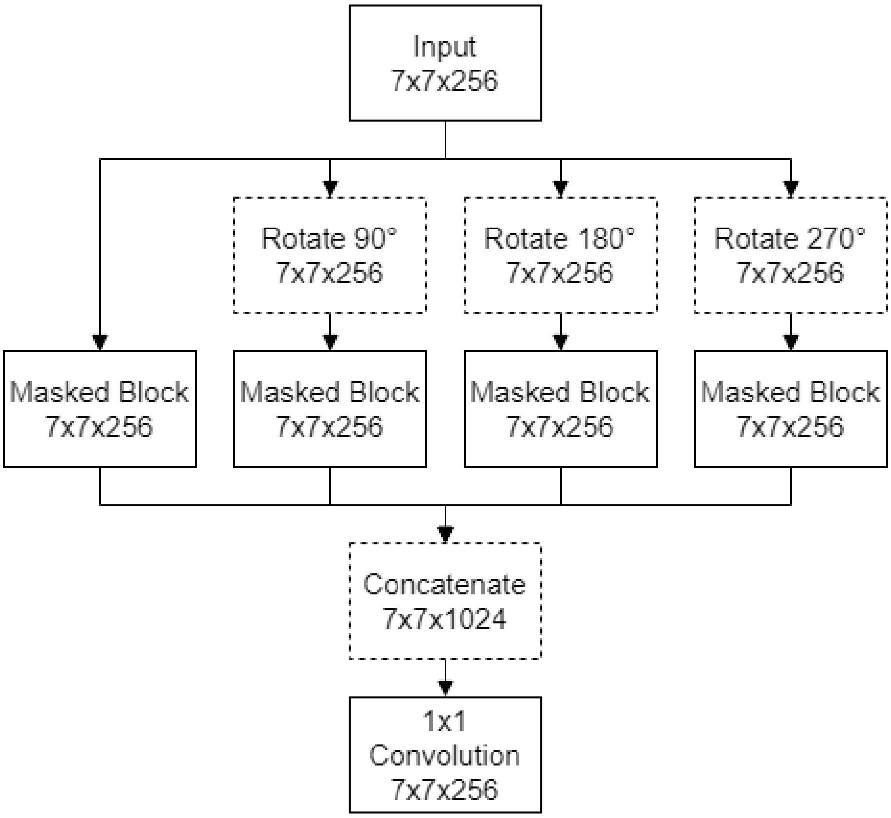


Figure 3.4: Architecture of a multi-directional masked block, wherein four distinct rotational transformations of the initial input are generated. These transformed inputs are fed into the PixelCNN masked blocks, resulting in the derivation of four outputs. These output are concatenated, which is succeeded by the employment of a 1x1 convolutional operation.

To address this limitation, this work proposes two modifications inspired by image in-filling: an alternative latent mask for producing a context vector, and modification to PixelCNN (van den Oord et al., 2016) for multi-directional context building. The proposed multi-directional CPC utilises these two modifications to more effectively learn representations from images where image rotation is uninformative.

The modified version of the PixelCNN is used as the autoregressive model in a multi-directional CPC model. This modification replaces each masked convolutional block of the PixelCNN architecture with a multi-directional masked blocks (Figure 3.4). Each multi-directional masked block takes a single input image and by rotating 0°, 90°, 180°and 270°, produces four versions of the original input. A masked block, as described in van den Oord et al. (2016), is then applied to each of them. The four outputs from the masked blocks are then concatenated and put through a final 1x1 convolutional layer for dimensionality reduction.

This multi-directional autoregressive model is used to learn a latent context from multiple directions at the same time. To take advantage of this, an alternative latent mask inspired by in-filling is introduced. With this mask (illustrated in Figure 3.1(d), the autoregressive model only has access to the patch representations around the perimeter of the patch representations. This means that images, where rotation is unimportant, can be better represented than with features learned using a single-directional CPC.

3.4 Unsupervised Representation Learning Experiments

This section presents a description of the datasets, training parameters, experimental setup, and results from experiments with the proposed multi-directional contrastive predictive coding approach on digital pathology whole-slide patches. The methodology, experimental results and detailed information on the datasets used in this chapter are provided in a transparent and reproducible manner. The codes and full results used in this section can be accessed on the project's GitHub repository[1] for further validation, replication and testing by the research community.

[1] GitHub Repository: `github.com/UoD-CVIP/Multi_Directional_CPC_Histology`

3.4.1 Dataset

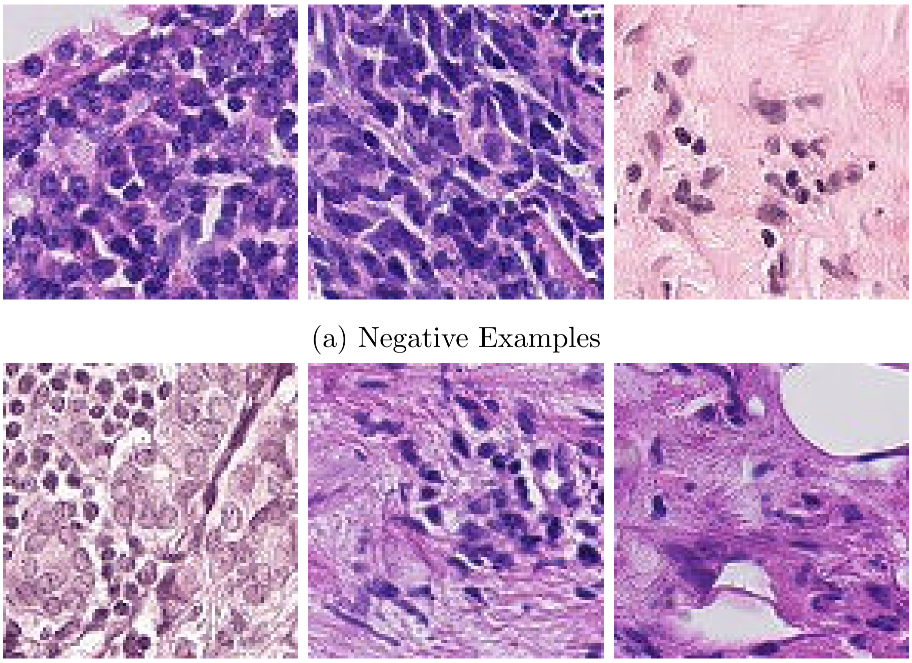
The publicly available open-source dataset Patch Cameleyon from Veeling et al. (2018) was chosen for its suitability to evaluate the proposed method. This dataset was selected due to its large number of non-overlapping whole-slide image patches that possess no inherent directionality, making it a suitable benchmark for evaluating rotation-invariant representations. It was previously used to evaluate Rotation Equivariant CNNs (Veeling et al., 2018). The patches were extracted from 400 whole-slide image scans of sentinel lymph node sections from the Camelyon16 dataset (Litjens et al., 2018). These whole-slide images were collected from two centres in the Netherlands (Radboud University Medical Center[2] and University Medical Center Utrecht[3]) and digitised using an objective of 40x magnification, resulting in a pixel resolution of 0.243 microns. Each patch has been annotated with a binary annotation indicating the presence or absence of metastatic tissue, by determining if the centred 32x32 pixels of the patch contains at least one pixel of tumour. The dataset is balanced between the two binary classification. The dataset contains a total of 327,680 patches which were split into training and testing sets, with a ratio of 90:10. To improve generalisation, data augmentation was applied during training, by randomly rotating, flipping vertically, and flipping horizontally each patch during sampling.

3.4.2 Experiment Setup

In order to evaluate the proposed multi-directional contrastive predictive coding approach, an ablation study was conducted using different combinations of the single or multi-directional autoregressive models, and top-down or in-filling style latent masks. To evaluate the learned representations from the CPC models, the trained encoders were used to initialise the encoder weights of 9 CNN classifiers. These CNN classifiers were then trained on a smaller, annotated subset of the training data, which was varied in size from 10 to 100,000 patches in logarithmic scale. This was repeated 3 times, each time with a different subsample of the training data to validate the robustness of the results. The test set was held static across all experiments.

[2] Radboud University Medical Center: `www.radboudumc.nl`

[3] University Medical Center Utrecht: `www.umcutrecht.nl`



(a) Negative Examples

(b) Positive Examples

Figure 3.5: Example images from the Patch Cameleyon dataset (Veeling et al., 2018, Litjens et al., 2018).

3.4.3 Training Parameters

The CPC models were trained using a method that involved splitting each input image into overlapping 24x24 patches that overlapped their neighbours by 12 pixels. A ResNeXt architecture with 101 layers was utilised as the encoder, followed by an additional convolutional layer to produce a 128-dimensional feature vector for each 24x24 patch in the image. The autoregressive model of the CPC was composed of 6 masked convolutional blocks to produce the context vector and predict the masked feature vectors. Additionally, 16 randomly selected images were used as negative samples for the CPC InfoNCE loss function. These parameters were taken from the original implementation of CPC for computer vision tasks (van den Oord et al., 2018).

The Adam optimisation algorithm was utilised to train the CPC and CNN models, with an initial learning rate of 1e-4. The Adam optimiser (Kingma and Ba, 2014) is based on adaptive estimation of first and second-order moments in the parameter gradients to adjust the learning rate during training. The CPC model was trained for 10 epochs with a batch size of 64 and the CNN models were trained for 50 epochs with a batch size of 258. 20% of the available training data was used as the validation set to prevent overfitting, and early stopping was implemented by saving the model when the loss was lowest on the validation set.

The CPC models took an average of 33 hours to train using a single Nvidia GeForce RTX 2080 Ti using 16-bit precision. The training loss over the epochs (Figure 3.6) suggests that the use of a multi-directional autoregressive model was more efficient at reducing the InfoNCE loss than a single-directional top-down autoregressive model. The in-filling style latent mask in combination with the multi-directional autoregressive model stabilised the CPC training process.

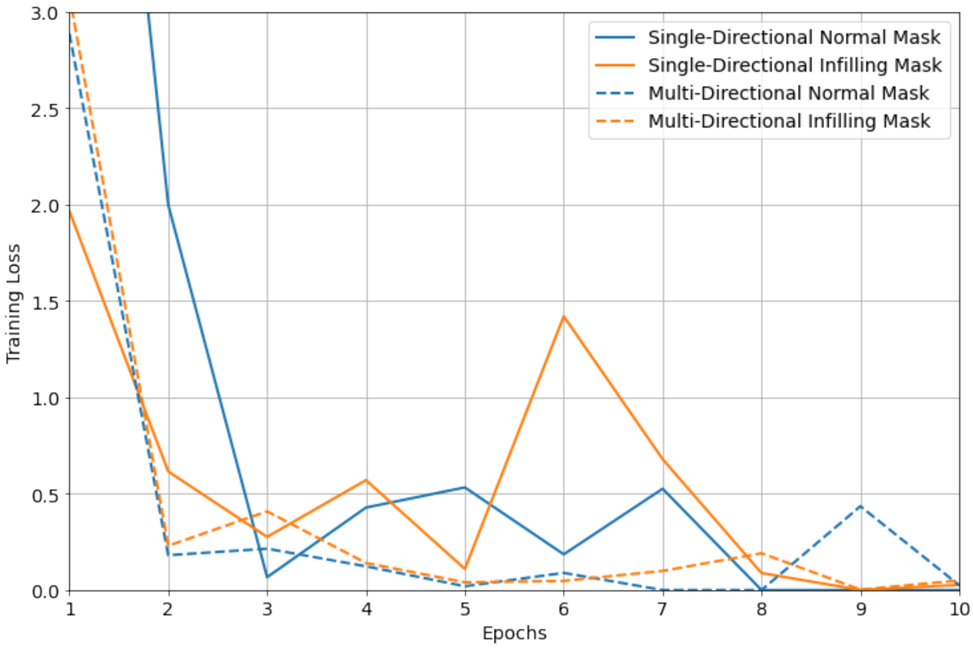


Figure 3.6: Training loss each epoch with the CPC Models.

3.4.4 Results

The research conducted a comprehensive evaluation of CNN classifiers' performance using a held-out testing dataset consisting of 32,768 images. These CNN classifiers were initialised with different sets of weights and biases transferred from various CPC encoders. To ensure robustness of the results, the experimentation process was repeated three times, with 100 bootstraps sampled from the metrics from these runs. This technique helps to capture the inherent variability in model performance due to different initialisations and data splits. The key metrics used to gauge the effectiveness of the different models were the mean accuracies and their associated 95% confidence intervals. These metrics were presented in both a tabular format (Table 3.1) and a graphical format (Figure 3.7). The mean accuracies provide a general overview of how well the models

Table 3.1: Mean test accuracies of the CNN classifiers with different pretraining (95% confidence intervals in parentheses).

Training Examples	No Pretraining	Single Directional Normal Mask	Single Directional Infilling Mask	Multi Directional Normal Mask	Multi Directional Infilling Mask
10	0.500 (0.0000)	0.558 (0.0042)	0.557 (0.0083)	0.547 (0.0062)	**0.590 (0.0094)**
32	0.563 (0.0106)	0.563 (0.0044)	**0.603 (0.0040)**	0.541 (0.0076)	0.580 (0.0077)
100	0.653 (0.0103)	0.592 (0.0016)	0.629 (0.0080)	0.598 (0.0071)	**0.657 (0.0082)**
316	0.692 (0.0058)	0.589 (0.0095)	**0.740 (0.0013)**	0.715 (0.0037)	0.709 (0.0057)
1000	0.725 (0.0013)	0.758 (0.0012)	0.739 (0.0015)	0.760 (0.0014)	**0.773 (0.0005)**
3162	0.750 (0.0002)	0.775 (0.0027)	0.774 (0.0011)	0.784 (0.0006)	**0.785 (0.0014)**
10000	0.786 (0.0005)	0.787 (0.0008)	0.777 (0.0012)	**0.792 (0.0033)**	0.781 (0.0017)
31624	0.781 (0.0011)	0.774 (0.0003)	0.770 (0.0008)	**0.785 (0.0005)**	0.784 (0.0008)
100000	0.775 (0.0006)	0.774 (0.0011)	**0.786 (0.0003)**	0.784 (0.0006)	0.784 (0.0009)

performed, while the confidence intervals offer insights into the stability and variability of the predictions. A baseline approach was included that involved training CNN without any pretraining. It's important to note that these baseline results cannot be directly compare to the outcomes presented in Chapter 2 since the experiments presented there focuses on different distinct tasks with other datasets.

The results revealed several patterns, including the CNN that were trained using a small amount of annotated data showed higher accuracies when initialised with parameters from CPC models that used the multi-directional PixelCNN autoregressor. This implies that the CPC-based pretraining is particularly advantageous when dealing with limited labelled data when access to annotated data is limited. The standard CPC approach faced challenges in learning representations suitable for transfer learning and instead lead to instances where the accuracy of the model initialised with CPC parameters was worse when compared to a CNN initialised randomly. As the number of annotated images increased, the advantage of using CPC pretraining diminished. This implies that the benefits of transfer learning from CPC encoders become less prominent when a larger annotated dataset is available, and in some cases, randomly initialised weights might be sufficient or even outperform CPC-based initialisations.

Table 3.1 depicts the ablation results obtained by employing two components: the use of an infilling style mask and the incorporation of a multi-directional PixelCNN autoregressive model. The finding emphasised that the infilling style mask played a crucial role, as the CNN models initialised with CPC models trained with infilling style masks consis-

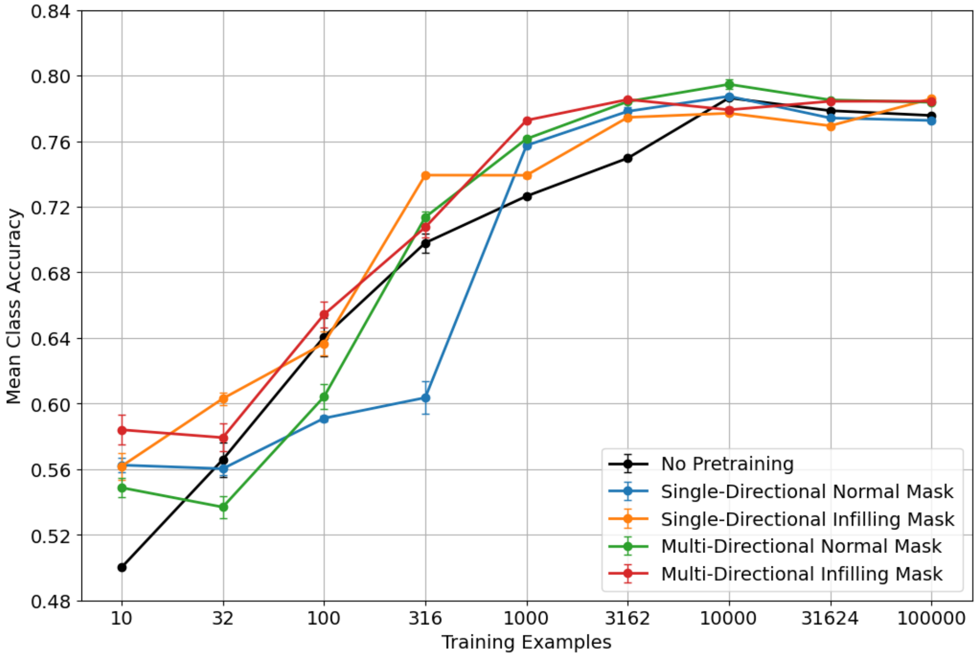


Figure 3.7: Mean class testing accuracies of the CNN classifiers trained using varying amounts of training examples, with 95% confidence intervals shown using error bars.

tently achieved the best results with limited annotated data. However, the CPC model combining a multi-directional PixelCNN autoregressive model with a normal single directional mask demonstrated optimal performance when dealing with larger subsets of the dataset. The confidence intervals reported in Table 3.1 provided insights into the stability of the methods' predictions. They showed that the model predictions were relatively stable, with slightly more variability arising from the additional components of the multi-directional CPC approach. This added variability was balanced by an increase in the mean performance, suggesting a slight trade-off between stability and performance enhancement.

3.5 Conclusion

The findings presented in Section 3.4.4 shed light on the performance of the original CPC implementation as applied to patch-based digital pathology tasks. The research conducted in this chapter, which builds upon the work of van den Oord et al. (2018), uncovers that the initial version of CPC does not exhibit desirable outcomes within the context medical image analysis. However, in response to this limitation, a novel approach involving multi-directional modifications to the CPC framework was proposed and examined.

This innovative adaptation led to significant improvements in results, as evidenced by enhanced classification accuracies, especially in scenarios where access to annotated data was constrained. Although there was a slight trade-off between stability and accuracy was observed in the experiments. This shows that the potential of the proposed multi-directional CPC modifications could be a valuable tool in scenarios when dealing with rotation invariant medical images and annotated data is scarce.

This chapter brings to the forefront the notion that methods built upon the presumption of inherent image directionality may not be optimal when applied to images devoid of such directional cues, as is often the case in certain biomedical imaging settings. General purpose algorithms should be evaluated on diverse datasets instead of the standard computer vision datasets such as CIFAR (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009) that are rooted in image directionality. The implications of the research findings are not confined solely to patch-based digital pathology tasks. Rather, they propose a broader application of the multi-directional CPC methodology to a diverse array of visual tasks in which the concept of image orientation holds little relevance. This encompasses various scenarios within the realm of biomedical imaging, as well as potentially extending to other domains where image directionality is not a crucial factor. In essence, the chapter advocates for the integration of this innovative approach as a versatile solution for tasks that involve images characterized by their orientation-agnostic nature.

Chapter 4

Predictive Probability Calibration

4.1 Introduction

Deep learning algorithms have been widely adopted for medical image analysis tasks and have even outperformed medical specialists in certain contexts, such as the binary classification of dermoscopic melanoma and nevi images (Maron et al., 2019). However, to ensure reliable translation of these classifiers to clinical settings, further improvements are necessary. As discussed in previous chapters 2 and 3, a known challenge with deep learning algorithms is poor calibration, which often leads to overconfident predictions. Mis-calibration, or the deviation between confidence and correctness, can impair the model's ability to accurately identify uncertainty in its predictions. To support cost-sensitive and selective clinical decision-making (Chapter 5) and avoid adverse outcomes, well-calibrated probabilistic outputs are helpful. Calibrated predictions are also relevant for applications such as active learning (Chapter 2), reinforcement learning (Dai et al., 2020), and out-of-distribution detection (Ulmer et al., 2020).

There have been numerous reports in the literature on techniques aimed at improving the calibration of deep learning classifiers in medical image classification. However, the effectiveness of these methods can be variable and there is a lack of clear guidance on which approach is most appropriate for a particular task or dataset. This chapter reports an empirical study which contributes to the existing body of evidence in this area by evaluating the performance of various calibration methods on medical image classification datasets from dermatology and histopathology.

Temperature scaling (Guo et al., 2017) is a widely used technique for calibrating modern neural networks, due in part to its post-hoc nature, ease of implementation, and demonstrated effectiveness. This method involves scaling output logits using a temperature parameter that is optimised on a validation set. The original implementation of temperature scaling utilised negative log-likelihood to optimise the temperature parameter. However, some researchers have proposed the use of other calibration metrics for this purpose, suggesting that this may lead to improved calibration (Mukhoti et al., 2020, Frenkel and Goldberger, 2021) but omitting to provide empirical evidence to support. This chapter aims to evaluate this claim empirically, using density-based and maximum calibration error estimators as metrics for assessing the performance of temperature scaling with various optimisation approaches.

Section 4.2 describes various measures of calibration which have been proposed and studied in the literature. This chapter investigates the impact of using different metrics to optimise temperature in the context of deep learning classifiers for medical image analysis. The analysis includes networks trained using both traditional cross-entropy loss with one-hot encoded target labels, as well as more recently proposed methods such as focal loss and label smoothing.

In this chapter, the effectiveness of various calibration methods is also compared to that of two Bayesian neural network approaches: one based on Bayes-by-Backprop and the other utilising a Laplace approximation method. These methods are described in detail in Section 4.3. The experimental design and details of the chapter are provided in Section 4.4. The results of the analysis, which includes evaluations on the ISIC 2019 multi-class dermatology dataset (Codella et al., 2018, Combalia et al., 2019, Tschandl et al., 2018) and the large Patch-Camelyon binary histopathology dataset, are presented in Section 4.4.4. These results provide insight into the relative performance of the different methods in terms of both calibration and accuracy on two distinct medical image classification tasks.

This work was presented at the Uncertainty for Safe Utilization of Machine Learning in Medical Imaging 2022 (UNSURE) workshop hosted at Medical Image Computing and Computing Assisted Intervention (MICCAI) in Singapore and published as part of its proceedings (Carse et al., 2022).

4.2 Measures of Calibration

Reliability diagrams are graphical representations that compare the predicted probabilities from a model to the frequencies of the outcome variable. The ordinate of the reliability diagram depicts the empirical frequency of the outcome variable within a specific predicted probability bin, while the abscissa represents the predicted probability values for each bin. The optimal reliability diagram will exhibit a diagonal line, signifying that the predicted probabilities correspond precisely with the empirical frequencies. To construct a reliability diagram, The predicted probabilities are partitions into a set of equally sized probability bins. For instance, if there are 10 predicted probabilities ranging from 0 to 1 inclusive, these probabilities may be subdivided into 10 bins representing distinct ranges of predicted probabilities. For each bin the corresponding empirical frequency of occurrence each probability is plotted. Figure 4.1 illustrates an example of a reliability diagram.

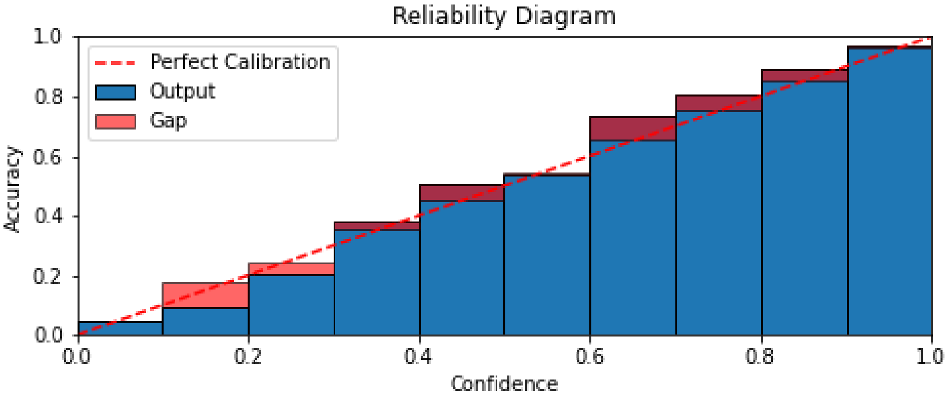


Figure 4.1: Example reliability diagram.

There are several metrics that can be used to assess the calibration of a machine learning model. One such measure is the expected calibration error (ECE) (Guo et al., 2017), which quantifies over n samples the discrepancy between predictive confidence ($\hat{p}_i$) and classification accuracy ($1(\hat{y}_i = y_i)$) over M number of bins B (Equation (4.1)). A commonly used estimator for ECE involves dividing the range of predicted probabilities into a set of equally spaced bins and computing the weighted average of the absolute differences between the accuracy of the predictions in each bin and the mean of the probabilities in that bin. Another metric, known as the maximum calibration error (MCE), is obtained by taking the maximum of the error across all bins (Equation (4.2)). This measure can be particularly useful in high-stakes situations where the worst-case calibration is of particular concern.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \left(\frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \right) - \left(\frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \right) \right| \tag{4.1}$$

$$MCE = \max_{m} \left| \left(\frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \right) - \left(\frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \right) \right| \tag{4.2}$$

Histogram-based estimation of ECE, which involves dividing the range of predicted probabilities into equally spaced bins, has been widely used to evaluate model calibration (Müller et al., 2019). However, this approach has been criticised for its inherent bias and statistical inefficiency, leading to the development of methods that adapt the number and sizes of bins (Roelofs et al., 2022). An alternative approach is to use continuous density estimators, such as kernel density estimation (KDE) (Parzen, 1962), to estimate the densities of accuracy and confidence in place of histograms. This method, known as KDE-ECE (Zhang et al., 2020), has been shown to offer improved data efficiency compared to histogram-based approaches. In this chapter, a triweight kernel was used when calculating KDE-ECE.

The negative log-likelihood loss function is a common choice for optimising deep neural network classifiers because it assesses the capacity of a probabilistic model to accurately predict the true conditional distribution. Beyond its use as an optimisation objective, the negative log-likelihood can also serve as a measure of a model's calibration.

4.3 Review of Calibration Methods

There are several approaches that have been proposed to improve model calibration. These can be broadly grouped into three main categories: regularisation methods, post-processing methods, and methods that inherently account for model uncertainty, such as Bayesian neural networks (Gawlikowski et al., 2021). Regularisation methods aim to improve calibration by adding constraints or regularises to the model during training, while post-processing methods involve adjusting the model's predictions after training has been completed. Bayesian neural networks, on the other hand, are designed to explicitly incorporate uncertainty into the model, enabling more robust and well-calibrated predictions.

4.3.1 Model Regularisation For Calibration

Model regularisation aims to improve the generalisation and calibration of a machine learning model by modifying the objective function used to optimise the model or by altering the training data in a way that encourages the model to be more robust. Dataset regularisation methods, such as data augmentation (Hendrycks et al., 2019) or the inclusion of out-of-distribution data in the training set (Hendrycks et al., 2018), are widely used to achieve this goal. In this chapter, data augmentation is employed in all experiments, and the effects of two additional regularisation techniques, label smoothing and focal loss, on model calibration are also investigated.

Neural network classifiers are typically trained using one-hot label encoding, in which the objective is to minimise the expected cross-entropy between the target outputs t_k and the network outputs y_k, where $t_k = 1$ for the true class and $t_k = 0$ for all other classes. An alternative approach, known as label smoothing, involves modifying the target class distribution by minimising the expected cross-entropy with modified targets $\hat{t} = (1 - \alpha)t + \frac{\alpha}{C}$, where C is the number of classes and α is a free parameter (Szegedy et al., 2016). The parameter α controls the degree of smoothing, with $\alpha = 1$ resulting in a uniform distribution and $\alpha = 0$ corresponding to one-hot encoding. Label smoothing has been shown to improve calibration and robustness to out-of-distribution data in medical image analysis tasks (Islam and Glocker, 2021) and is popular due to its ease of implementation and minimal computational overhead.

Focal loss is a loss function that was originally designed to improve the performance of object detection by encouraging the model to focus more on samples with lower confidence (Lin et al., 2017). It has been shown to improve model calibration in some cases (Mukhoti et al., 2020). Focal loss weights the predictions based on their confidence, with the goal of forcing the model to learn more from examples that are close to the decision boundary. This can reduce overconfidence, improve calibration, and lead to better performance on unbalanced datasets. Focal loss can be expressed as $FL(y) = -\alpha(1 - y)^{\gamma} \log(y)$, where the factor $(1 - y)^{\gamma}$ is included with the cross-entropy loss to weight the predictions. The hyperparameters $\gamma > 0$ and $\alpha \in [0, 1]$ control the weighting of less confident examples and improve numerical stability, respectively.

4.3.2 Post-Hoc Calibration

Temperature scaling is a widely used post-hoc calibration method for modern neural networks, including in medical image analysis applications (Guo et al., 2017, Liang et al., 2020). It is a single-parameter variant of Platt logistic scaling (Platt, 1999) that applies a learned temperature parameter, $T > 0$, to rescale the output logits z of a neural network before applying a softmax activation function to obtain probabilistic predictions $\hat{y}$ (see Equation (4.3)). When $T = 1$, the standard softmax activation is recovered. As the temperature parameter is used to scale all of the logits, the output $\hat{y}$ has a monotonic relationship with the unscaled output, meaning that classification accuracy is unaffected by temperature scaling.

$$\hat{y} = \frac{e^{z/T}}{\sum_{j=1}^{J} e^{z_j/T}} \tag{4.3}$$

The original implementation of temperature scaling optimises the temperature parameter T by minimising the negative log-likelihood of the predictions (expected cross-entropy) on a validation set. In the experiments described in Section 4.4, alternative measures for optimising T are considered. It is hypothesised that optimising T using a calibration measure on a validation set will result in improved test calibration when evaluated using that same measure. Further details on the experimental design and results are provided in Section 4.4.

4.3.3 Bayesian Approximation

Bayesian neural networks are a type of machine learning model that infer distributions over their weight parameters, as opposed to the traditional approach of obtaining point estimates. This allows for the use of Monte Carlo sampling to approximate predictive distributions and compute estimates of predictive means and uncertainty measures, such as variance. Bayesian neural networks have been applied in medical applications for the purpose of improving calibration and estimating uncertainty (Kwon et al., 2020).

Bayes-by-Backprop (Blundell et al., 2015) is a method for training Bayesian neural networks that combines the use of backpropagation to calculate gradients with variational inference to approximate the posterior distribution $q(w|\theta)$ over the model's weights. The parameters of this distribution, θ, are determined by minimising the KL divergence between the variational posterior and the true posterior, which is estimated using Monte Carlo sampling of the evidence lower bound (ELBO) as shown in Equation (4.4). Here D is the dataset and N is the number of Monte Carlo samples.

$$\mathrm{ELBO}(D,\theta) \approx \sum_{n=1}^{N} \log q(w^i|\theta) - \log p(w^i) - \log p(D|w^i) \tag{4.4}$$

To train a Bayesian neural network, the ELBO is typically combined with the cross-entropy loss to form a composite loss function. When doing so, the ELBO is weighted by a factor of $\pi_m = \frac{2^{M-m}}{2^{M}-1}$, where M is the total number of batches and m is the current batch. This weighting scheme gives more influence on the Bayesian complexity term during the early stages of training, while allowing the model to learn more from the data as training progresses.

Laplace approximation (MacKay, 1992) is a method to produce Bayesian neural networks from neural networks by approximating the posterior distribution over a model's parameters by fitting a Gaussian distribution with a mean equal to the maximum a posteriori (MAP) of the parameter, and variance equal to the observed fisher information. The curvature of this Gaussian is estimated using approximations to the Hessian matrix at the maximum (Botev et al., 2017). Laplace approximation can be applied post-hoc to a trained neural network, allowing for the sampling of probabilistic predictions at low computational cost compared to other methods such as Bayes-by-Backprop (Daxberger et al., 2021). This makes it a popular choice for Bayesian inference in practice.

4.4 Calibration Experiments

This section presents the details of the datasets, training parameters, experimental setup, and results. The code and complete results for these experiments can be found in the project GitHub repository[1].

4.4.1 Datasets

Two datasets were used: the ISIC 2019 challenge dataset (Codella et al., 2018, Combalia et al., 2019, Tschandl et al., 2018) and the Patch Camelyon (PCam) dataset (Veeling et al., 2018). The ISIC 2019 dataset con-

[1] GitHub Repository: `github.com/UoD-CVIP/Medical_Calibration`

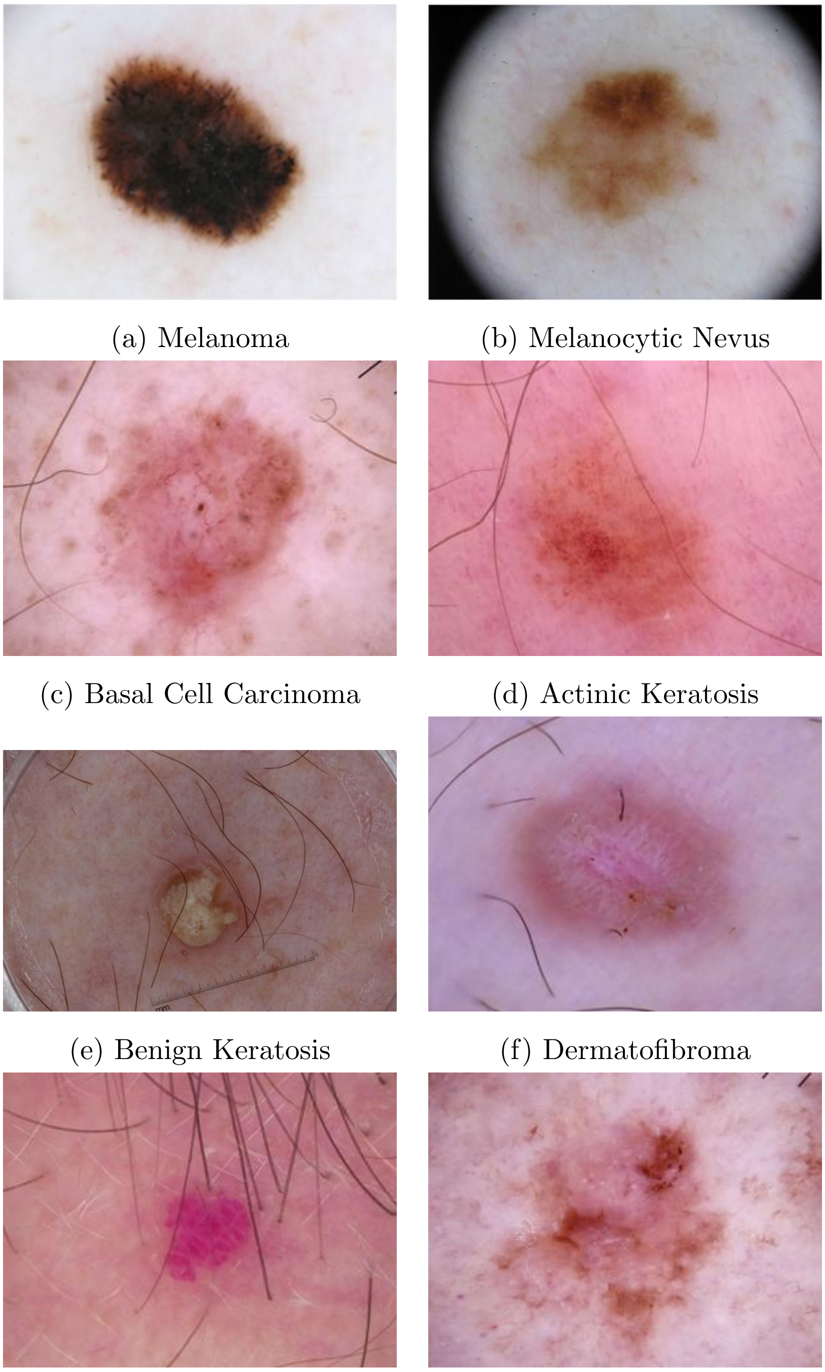
sists of 25,331 dermoscopic skin lesion images belonging to eight diagnostic classes: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma (example images of each classification in Figure 4.2). The PCam dataset consists of 327,680 96x96 pixel image patches extracted from whole-slide images of H&E-stained lymph node sections from the Camelyon16 dataset (Bejnordi et al., 2017) (example images in Figure 3.5). The datasets were split into training, validation, and testing sets with proportions 6:2:2. As the ISIC 2019 images vary in size each image was pre-processed by cropping their width to be equal to their height and resizing to 256 × 256 pixels. Both datasets were augmented through normalisation of each image channel, random horizontal and vertical flipping, and random rotation by multiples of 90°.

4.4.2 Experiment Setup

Seven different types of CNN classifiers were trained on both the ISIC 2019 and the PCam datasets. In the case of ISIC 2019, each classifier was trained three times using different random seeds, which altered the data splits and weights initialisations for the training, validation, and testing sets. The first classifier, referred to as the baseline model, was trained using a standard cross-entropy function with one-hot label encoding. The next two classifiers used label smoothing with cross-entropy, with alpha values of 0.1 and 0.2, respectively. An additional two classifiers were trained using focal loss, with gamma values of 2.0 and 5.0, respectively. Temperature scaling was applied to each of these model types after training, with the temperature parameter optimised using a Limited-memory BFGS optimiser (Liu and Nocedal, 1989) for various measures of calibration, including the negative log-likelihood, KDE-ECE, MCE, and combinations of these three measures. For comparison, two types of Bayesian neural networks were also trained, using Bayes-by-Backprop and Laplace approximation, respectively.

4.4.3 Training Parameters

An EfficientNet encoder with a compound coefficient of 7, pre-trained on ImageNet, was utilised as the CNN model in this chapter. The encoder was followed by a fully connected hidden layer with a width of 512 neurons before the output layer. For Bayesian convolutional neural networks, the final hidden and output fully-connected layers were replaced with



(a) Melanoma

(b) Melanocytic Nevus

(c) Basal Cell Carcinoma

(d) Actinic Keratosis

(e) Benign Keratosis

(f) Dermatofibroma

(g) Vascular Lesion

(h) Squamous Cell Carcinoma

Figure 4.2: Example images from the ISIC Challenge 2019 dataset (Codella et al., 2018, Combalia et al., 2019, Tschandl et al., 2018).

Bayesian fully-connected layers that learn distributions for the weights and biases to be sampled from. Cyclical learning rate scheduling was employed, with scheduling between 10^{-4} and 10^{-1}. The batch size for the ISIC 2019 data was 16, and the batch size for the PCam data was

64. Bayes-by-backprop used a weighted loss function that combined the expected lower bound from 10 forward-propagation passes and the cross-entropy of the 10 predictions. The Laplace approximation was optimised post-hoc by fitting it to a trained convolutional neural network model on the output layer of the model using a full Hessian structure. Each model was trained for 40 epochs, and the model with the best validation loss was selected for evaluation.

4.4.4 Results

The results for the ISIC 2019 and PCam datasets are presented in Table 4.1. The table is divided into sections showing the results for each trained model with the temperature set to one, followed by the results with the temperature optimised for various measures of calibration. It is important to note that temperature scaling does not affect accuracy. The results for the ISIC 2019 dataset are reported as the mean and standard deviation, estimated from three runs.

On the multi-class skin lesion classification task, temperature scaling consistently improved calibration. The largest improvement in mean KDE-ECE when using the baseline CNN model was from 0.046 to 0.012; other temperature scaling measures yielded similar results. Label smoothing achieved better accuracy, but the calibration was inferior to temperature scaling with one-hot labels. When temperature scaling was added to label smoothing, it tended to improve calibration, but it was not as effective as temperature scaling without label smoothing in this regard. Optimising for MCE was ineffective.

Focal loss with a value of $\gamma = 2.0$ and temperature optimised for KDE-ECE achieved calibration and accuracy that were competitive with, or perhaps slightly better than, the cross-entropy model with temperature scaling. Both of the focal loss models showed behaviour similar to that reported in Mukhoti et al. (2020), in that temperature optimisation for KDE-ECE resulted in better calibration with a significant impact on KDE-ECE compared to temperature optimisation for NLL.

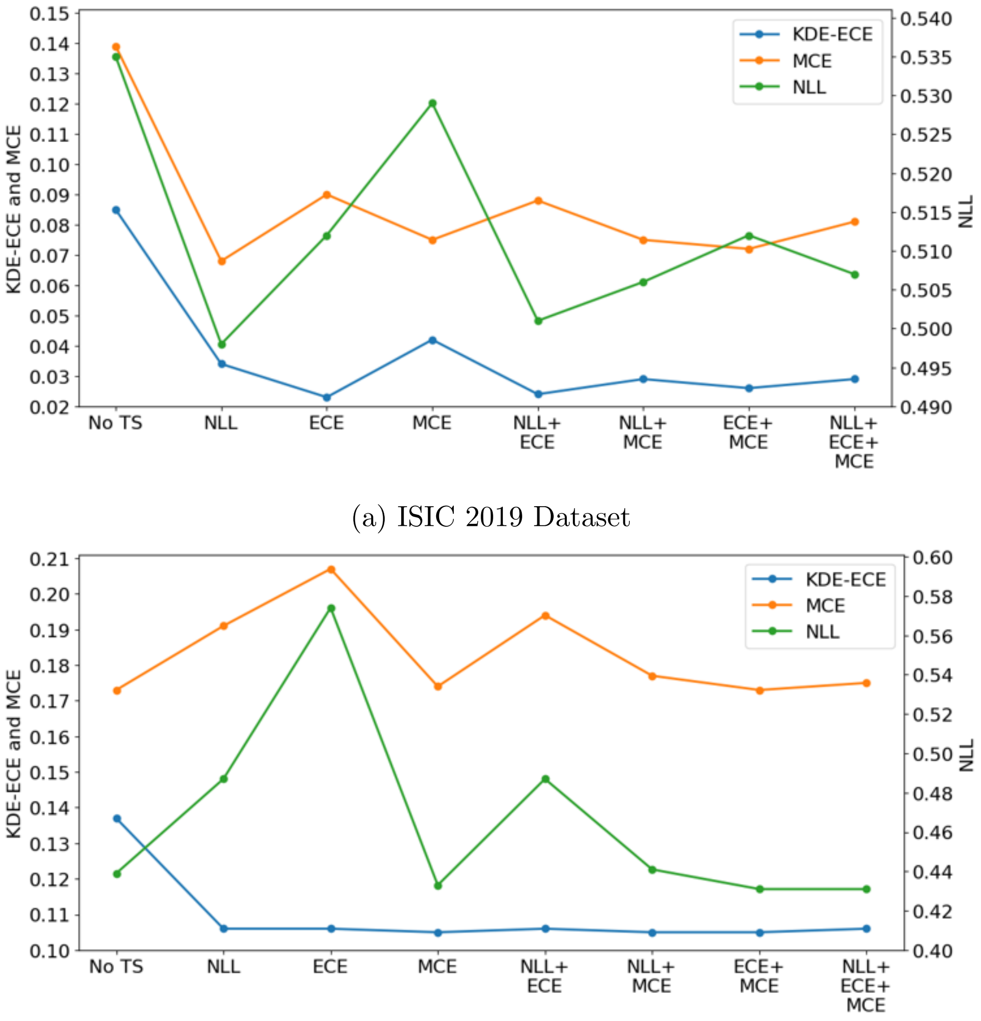
On the binary classification task using the PCam dataset, a different behaviour was observed when using temperature scaling. The baseline CNN model did not benefit from temperature scaling in terms of calibration, regardless of the calibration measure used to optimise the temperature.

Calibration using label smoothing appears to be better than that

Table 4.1: Calibration and accuracy results for ISIC 2019 and PCam datasets. ISIC 2019 results are means and standard deviations over three iterations. Each section reports results from a single model type; TS denotes temperature scaling. For KED-ECE, MCE and NLL lower is better. For ACC higher is better.

	ISIC 2019				PCam			
	KDE-ECE	MCE	NLL	ACC	KDE-ECE	MCE	NLL	ACC
Baseline CNN	0.046 ± 0.017	0.112 ± 0.042	0.514 ± 0.015	0.836 ± 0.012	0.123	0.187	0.543	0.848
TS nll	0.015 ± 0.002	0.037 ± 0.014	0.471 ± 0.017	0.836 ± 0.012	0.123	0.187	0.543	0.848
TS ece	0.017 ± 0.007	0.038 ± 0.010	0.472 ± 0.017	0.836 ± 0.012	0.123	0.204	0.585	0.848
TS mce	0.016 ± 0.004	0.042 ± 0.009	0.471 ± 0.018	0.836 ± 0.012	0.123	0.197	0.567	0.848
TS nll+ece	0.012 ± 0.002	0.043 ± 0.012	0.471 ± 0.017	0.836 ± 0.012	0.122	0.210	0.603	0.848
TS nll+mce	0.018 ± 0.002	0.038 ± 0.010	0.472 ± 0.018	0.836 ± 0.012	0.123	0.210	0.605	0.848
TS ece+mce	0.014 ± 0.021	0.041 ± 0.014	0.472 ± 0.017	0.836 ± 0.012	0.123	0.193	0.557	0.848
TS nll+ece+mce	0.015 ± 0.003	0.042 ± 0.011	0.472 ± 0.017	0.836 ± 0.012	0.123	0.187	0.543	0.848
LS $\alpha = 0.1$	0.039 ± 0.001	0.065 ± 0.001	0.495 ± 0.011	0.855 ± 0.005	0.116	0.146	0.398	0.848
TS nll	0.028 ± 0.003	0.087 ± 0.021	0.483 ± 0.013	0.855 ± 0.005	0.123	0.244	0.536	0.848
TS ece	0.029 ± 0.003	0.134 ± 0.023	0.504 ± 0.013	0.855 ± 0.005	0.124	0.346	1.112	0.848
TS mce	0.053 ± 0.016	0.070 ± 0.016	0.505 ± 0.004	0.855 ± 0.005	0.121	0.198	0.423	0.848
TS nll+ece	0.028 ± 0.003	0.086 ± 0.020	0.048 ± 0.013	0.855 ± 0.005	0.123	0.237	0.551	0.848
TS nll+mce	0.033 ± 0.006	0.071 ± 0.016	0.490 ± 0.015	0.855 ± 0.005	0.121	0.198	0.423	0.848
TS ece+mce	0.032 ± 0.006	0.072 ± 0.016	0.489 ± 0.016	0.855 ± 0.005	0.121	0.198	0.423	0.848
TS nll+ece+mce	0.031 ± 0.004	0.070 ± 0.010	0.488 ± 0.014	0.855 ± 0.005	0.121	0.199	0.424	0.848
LS $\alpha = 0.2$	0.105 ± 0.005	0.173 ± 0.050	0.562 ± 0.015	0.862 ± 0.009	0.097	0.142	0.370	0.852
TS nll	0.037 ± 0.001	0.145 ± 0.006	0.503 ± 0.013	0.862 ± 0.009	0.118	0.307	0.601	0.852
TS ece	0.036 ± 0.001	0.134 ± 0.023	0.504 ± 0.013	0.862 ± 0.009	0.113	0.226	0.396	0.852
TS mce	0.095 ± 0.018	0.129 ± 0.032	0.553 ± 0.015	0.862 ± 0.009	0.113	0.226	0.396	0.852
TS nll+ece	0.036 ± 0.001	0.148 ± 0.009	0.503 ± 0.013	0.862 ± 0.009	0.117	0.284	0.508	0.852
TS nll+mce	0.051 ± 0.006	0.099 ± 0.006	0.514 ± 0.012	0.862 ± 0.009	0.113	0.226	0.396	0.852
TS ece+mce	0.047 ± 0.003	0.106 ± 0.019	0.511 ± 0.011	0.862 ± 0.009	0.113	0.226	0.396	0.852
TS nll+ece+mce	0.046 ± 0.007	0.113 ± 0.012	0.510 ± 0.010	0.862 ± 0.009	0.114	0.243	0.409	0.852
FL $\gamma = 2.0$	0.057 ± 0.020	0.097 ± 0.027	0.491 ± 0.013	0.840 ± 0.004	0.122	0.101	0.356	0.854
TS nll	0.031 ± 0.004	0.078 ± 0.018	0.484 ± 0.007	0.840 ± 0.004	0.100	0.155	0.371	0.854
TS ece	0.011 ± 0.003	0.061 ± 0.021	0.492 ± 0.004	0.840 ± 0.004	0.101	0.180	0.388	0.854
TS mce	0.014 ± 0.002	0.062 ± 0.021	0.497 ± 0.010	0.840 ± 0.004	0.101	0.180	0.392	0.854
TS nll+ece	0.014 ± 0.003	0.062 ± 0.018	0.489 ± 0.007	0.840 ± 0.004	0.100	0.170	0.385	0.854
TS nll+mce	0.013 ± 0.003	0.063 ± 0.018	0.490 ± 0.008	0.840 ± 0.004	0.101	0.180	0.393	0.854
TS ece+mce	0.012 ± 0.003	0.062 ± 0.021	0.494 ± 0.008	0.840 ± 0.004	0.101	0.179	0.391	0.854
TS nll+ece+mce	0.026 ± 0.020	0.085 ± 0.034	0.497 ± 0.013	0.840 ± 0.004	0.101	0.179	0.392	0.854
FL $\gamma = 5.0$	0.180 ± 0.007	0.250 ± 0.010	0.615 ± 0.011	0.823 ± 0.002	0.229	0.289	0.530	0.835
TS nll	0.061 ± 0.010	0.123 ± 0.022	0.551 ± 0.011	0.823 ± 0.002	0.068	0.063	0.382	0.835
TS ece	0.024 ± 0.005	0.084 ± 0.026	0.589 ± 0.016	0.823 ± 0.002	0.070	0.077	0.391	0.835
TS mce	0.031 ± 0.009	0.074 ± 0.007	0.621 ± 0.046	0.823 ± 0.002	0.069	0.069	0.387	0.835
TS nll+ece	0.032 ± 0.008	0.101 ± 0.026	0.561 ± 0.011	0.823 ± 0.002	0.069	0.071	0.387	0.835
TS nll+mce	0.031 ± 0.009	0.102 ± 0.024	0.562 ± 0.010	0.823 ± 0.002	0.069	0.069	0.387	0.835
TS ece+mce	0.025 ± 0.005	0.079 ± 0.012	0.593 ± 0.029	0.823 ± 0.002	0.069	0.069	0.387	0.835
TS nll+ece+mce	0.028 ± 0.007	0.094 ± 0.025	0.569 ± 0.009	0.823 ± 0.002	0.069	0.069	0.387	0.835
Bayes-by-B'prop	0.118 ± 0.006	0.260 ± 0.021	0.886 ± 0.062	0.795 ± 0.069	0.115	0.208	0.551	0.857
Laplace approx.	0.041 ± 0.016	0.101 ± 0.037	0.507 ± 0.010	0.837 ± 0.012	0.122	0.210	0.603	0.848

obtained by the baseline model without temperature scaling. Adding temperature scaling to label smoothing appears to slightly worsen the calibration performance on this dataset. It can be observed that the



(a) ISIC 2019 Dataset

(b) Patch Camelyon Dataset

Figure 4.3: Calibration metrics of models calibrated using temperature scaling optimised on different calibration metrics (y axis). (a) ISIC 2019 results (b) Patch Camelyon results.

calibration measures as functions of temperature had shallow minima, and the optimised temperature values obtained using the validation set differed from those that would have been optimal for the test set.

For the LS model with $\alpha = 0.2$, the model is encouraged to assign a prediction of 0.8 when it is confident about the prediction. While the accuracy of the model is around 80%, the predictions of the model will be better calibrated.

Focal loss with $\gamma = 2.0$ had higher accuracy and similar calibration to the baseline model without temperature scaling. However, unlike the baseline, both focal loss models benefited from temperature scaling in terms of calibration. Focal loss with $\gamma = 5.0$ and temperature scaling achieved the best calibration of any of the models on the PCam dataset. The measure used to optimise the temperature made little or no difference in the results.

Bayes-by-Backprop performed relatively poorly on the ISIC 2019 dataset in terms of both calibration and accuracy. This could be due to the additional complexity involved with training a multi-class Bayesian neural network using back-propagation. On the other hand, Bayes-by-Backprop achieved the highest accuracy on the PCam dataset and had a lower KDE-ECE than the baseline network. Nevertheless, temperature scaling of networks trained with focal loss achieved comparable accuracy and better calibration in terms of both KDE-ECE and MCE.

The Laplace approximation outperformed Bayes-by-Backprop on the ISIC 2019 dataset and is also computationally less expensive. However, it was not competitive with temperature scaling in terms of calibration. On the larger PCam dataset, the Laplace approximation did not provide an advantage over the baseline model.

4.5 Conclusion

This chapter centres on the investigation and enhancement of deep neural network calibration for medical image classification. The primary objective of this chapter was to contribute to the existing body of knowledge by assessing the efficacy of diverse calibration methods across various tasks and datasets. The findings highlight the importance of carefully selecting a calibration approach that is appropriate for the task at hand.

The investigation encompassed two discrete datasets: the ISIC2019 dataset and the PCam dataset. Notably, the calibration enhancements rendered by temperature scaling consistently manifested within the confines of the ISIC2019 dataset. This augmentation was particularly conspicuous in instances where networks were trained employing either cross-entropy loss or focal loss. In stark contrast, the application of temperature scaling to networks trained with cross-entropy on the PCam dataset did not yield calibration improvements. However, a constructive calibration effect became evident upon the incorporation of temperature scaling in conjunction with focal loss.

Although an assumption may be held that the introduction of calibration measures to adjust temperature using validation data would be invariably improve calibration performance, the empirical results of this chapter contradicted this (as seen in Figure 4.3). The influence of optimisation measures on calibration performance was often marginal. Although the amalgamation of focal loss and temperature scaling surfaced as a robust strategy for cultivating desirable calibration outcomes. This could be attributed to the intricate interplay between the focused loss's concentration on intricate training examples and temperature scaling's capacity to exploit this concentration, thereby yielding heightened performance in binary classification instances. It is vital to highlight that this pattern may be more useful for binary classification, as multi-class scenarios increase the prediction space's complexity.

Subsequent study could explore deeper into the implications of hyperparameters, such as the focal loss parameter γ, in this context. Such research could lead to a better understanding of the significance of these hyperparameters in affecting calibration performance. Further to this, an avenue for inquiry lies in the exploration of alternative calibration methods. Ongoing developmental efforts are predominantly directed towards post-hoc calibration methods, favoured for their practicality and ease of implementation. For example, the potential advantages of employing an ensemble of temperature scaling techniques for the calibration of predictions related to specific classes (Zhang et al., 2020). Future work could investigate, calibration in proximity to decision boundaries presents opportunity for research to improve calibration outcomes.

Chapter 5

Asymmetrical Selective Classification

5.1 Introduction

5.1.1 Robust Selective Classification for Skin Lesions

The utilisation of automated image analysis for the assessment of skin lesions holds great promise in enhancing diagnostic accuracy and streamlining clinical workflows within the field of dermatology. By employing lesion classifiers that generate class probability distributions, it becomes possible to estimate the associated costs of clinical decisions, such as referral recommendations, thereby facilitating informed decision making.

It is important to note that the costs resulting from misclassification are typically asymmetric, with a greater impact associated with falsely categorising a malignant lesion as benign compared to the reverse. To achieve optimal decision making, it is crucial that the predicted class probabilities are properly calibrated. Additionally, a clinically practical system should possess the ability to determine its own level of training, which is integral to both robustness and clinical viability. Furthermore, the classifiers should exhibit selectivity, declining to analyse images that fall outside of their capabilities, particularly relevant for lesion types that may not be adequately represented within the training data.

In this chapter, methods for cost-sensitive and selective classification of skin lesions using binary and multi-class deep classification models are investigated. An experimental design that includes both binary (malignant vs benign) and multi-class classification tasks. The images utilised in this chapter were sourced from the ISIC 2019 dataset (Codella et al., 2018, Combalia et al., 2019, Tschandl et al., 2018).

To add selectivity to a machine learning model, selective classification can be utilised, where the classifier has the option to reject an image if it does not meet certain criteria. The goal of selective classification is to reduce the number of incorrect classifications and decrease the occurred costs of the predictions. In selective classification, a classifier decides whether to accept or reject an input then if accepted makes a prediction on the input. Sometimes the prediction is made first and is used to inform the rejection decision, for example thresholding the prediction confidence. The threshold can be set based on the desired false positive rate or other performance metrics. If the confidence score of the classifier is below the threshold, the image is rejected and not assigned a class label.

Asymmetric misclassification costs are important in the machine learning arena because they provide a framework for dealing with the issues that arise from uneven consequences of classification failures. The repercussions of false negatives and false positives are intrinsically unequal in the context of medical imaging. In the diagnosis of skin lesions, for example, failing to detect a malignant lesion (false negative) has considerably more serious repercussions than incorrectly recognising a benign lesion as malignant (false positive). The former can result in delayed crucial care, potentially jeopardising the patient's health or life, whilst the latter can result in unnecessary stress and follow-up operations, albeit with less severe consequences.

Using Asymmetrical Misclassification Costs entails attributing higher costs to false negatives than false positives; this paradigm acknowledges the increased importance of reducing severe errors. This strategic cost allocation directs the classifier's decision-making process. The increased cost of false negatives motivates the classifier to favour increased sensitivity - the ability to correctly identify true positives. As a result, the classifier is purposefully built to be more careful and meticulous during classification, with a strong emphasis on minimising false negatives. While attempting to improve sensitivity over specificity, the model may demonstrate a tendency to overestimate its predictive powers. This can show as overconfidence in its forecasts, resulting in a distorted probability distribution. The model's proclivity to assign high probability to its predictions may jeopardise calibration, resulting in a misalignment between projected and actual outcomes.

5.1.2 Summary of Work

The experiments in this chapter focus on the use of empirical coverage and selective costs to evaluate the performance of selective classification methods in skin lesion analysis. The significance of considering the asymmetry of misclassification costs in both binary (benign vs malignant) and multi-class disease classification scenarios is emphasised. An extensive evaluation of various selective classification methods, including predictive probability calibration, uncertainty estimation, and selective classification models, is carried out. A novel selective classification model, Expected Cost SelectiveNet (EC-SelectiveNet), is introduced and analysed. EC-SelectiveNet is based on the SelectiveNet model (Geifman and El-Yaniv, 2019) and makes selection decisions based on expected costs, rather than on the image rejection rate. EC-SelectiveNet discards the additional heads used in SelectiveNet (selection and auxiliary heads) and relies solely on the expected costs for image selection.

An earlier version of this work was presented at the Uncertainty for Safe Utilization of Machine Learning in Medical Imaging 2021 (UNSURE) workshop hosted at Medical Image Computing and Computing Assisted Intervention (MICCAI) in Strasbourg, France and published as part of its proceedings (Carse, Süveges, Hogg, Trucco, Proby, Fleming and McKenna, 2021).

5.2 Literature Review

Selective classification was initially introduced by Chow (1957), who explored the concept of a rejection option. Subsequently, this notion was further characterised as a risk-coverage trade-off in the literature (El-Yaniv et al., 2010). Various authors have endeavoured to construct algorithms that can optimally achieve the best trade-offs. The majority of the research in this area has focused on traditional machine learning methods, such as support vector machines and nearest neighbours (Hellman, 1970, Fumera and Roli, 2002, Wiener and El-Yaniv, 2015). More recently, Cortes et al. (2016) proposed a method for jointly learning prediction and selection functions instead of relying on conventional confidence-based rejection. The authors demonstrated that their approach yielded promising outcomes when compared to other selective classification experiments without having to rely on methods that produce noise-free confidence predictions.

Geifman and El-Yaniv (2017) were the pioneers of applying selective classification to deep learning algorithms by proposing a rejection mechanism from the model and an automatic threshold selection method to achieve the desired risk. They utilised the reject options of either the softmax response (maximum softmax prediction) or Monte Carlo dropout. Subsequently, Geifman and El-Yaniv (2019) introduced SelectiveNet, a deep learning model that can jointly learn the prediction and selection functions, trained for a specific target coverage. The authors asserted that their model's selective classification performance outperformed the methods against which they compared it, namely, softmax response and Monte Carlo dropout.

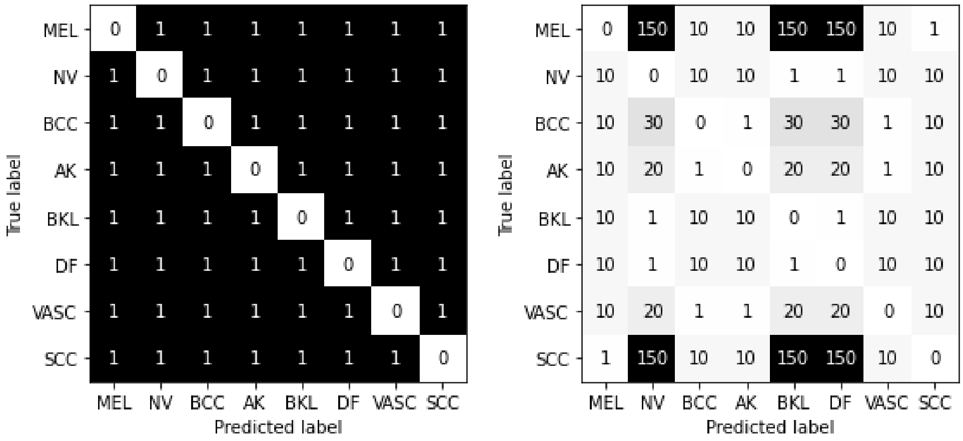
The utilisation of predictive probability outputs from neural networks for selective classification can pose a challenge due to the weak calibration of these probabilities, which arises from the softmax function. As such, it is imperative to calibrate the predictions prior to their use in selective classification, as delineated in Chapter 4. The uncertainty inherent in calibrated predictions can serve as a selection criterion by rejecting samples exhibiting high levels of uncertainty. Bayesian neural networks offer a means of quantifying uncertainty in neural networks and can be trained via a range of methods, including Monte Carlo dropout techniques (Gal and Ghahramani, 2016), backpropagation with weights treated as random variables (Blundell et al., 2015), and fitting a Gaussian distribution to the weights for posterior probabilities (MacKay, 1992). By sampling Bayesian neural networks, uncertainty can be measured from the samples using various methods (Gal et al., 2016).

5.3 Asymmetrical Selective Classification

The process of selective classification involves two key components: the selection function and the prediction function. The selection function, denoted as $\sigma(x)$, determines whether or not an image x should be classified. If an image is rejected, then $\sigma(x) = 0$, and if it is selected, then $\sigma(x) = 1$. The empirical coverage, $\phi(\sigma|S)$, is defined as the proportion of images selected for classification, calculated as the mean of the selection function over the images in the data set S. The prediction function, $P(x)$, is used to make a classification decision for each selected image, and each decision incurs a cost. The average cost over the selected images is referred to as the empirical selective cost.

The mis-classification costs can be specified in a matrix C, where C_{jk} is the cost of assigning class k when the true class is j. These costs are specific to the deployment setting and are influenced by various factors such as health economics, quality of life considerations, and available treatments. In many reported experiments on dermatology image classification, a symmetric cost matrix is used, i.e., $C = \mathbf{1} - I$, where $\mathbf{1}$ is a matrix of ones and I is the identity matrix. However, this assumption of symmetry is unrealistic, and in many medical classification tasks, the costs are highly asymmetric.

For instance, in the binary classification of malignant (class 1) and benign (class 0) lesions, the cost matrix may reflect that mis-classifying a malignant lesion as benign is much more costly than the reverse mis-classification. In this scenario we might have, $C_{1,0} = 10.0$, $C_{0,1} = 1.0$, $C_{1,1} = 0.0$, and $C_{0,0} = 0.0$ for example. For multiple lesion classes, the cost matrix may be more complex and should be decided in consultation with relevant stakeholders such as general practitioners, patient representative groups, and health economists. It is important to note that the values used for asymmetric costs should vary depending on the specific clinical setting and should be determined through discussions with relevant experts.



True label \ Predicted label	MEL	NV	BCC	AK	BKL	DF	VASC	SCC
MEL	0	1	1	1	1	1	1	1
NV	1	0	1	1	1	1	1	1
BCC	1	1	0	1	1	1	1	1
AK	1	1	1	0	1	1	1	1
BKL	1	1	1	1	0	1	1	1
DF	1	1	1	1	1	0	1	1
VASC	1	1	1	1	1	1	0	1
SCC	1	1	1	1	1	1	1	0

(a)

True label \ Predicted label	MEL	NV	BCC	AK	BKL	DF	VASC	SCC
MEL	0	150	10	10	150	150	10	1
NV	10	0	10	10	1	1	10	10
BCC	10	30	0	1	30	30	1	10
AK	10	20	1	0	20	20	1	10
BKL	10	1	10	10	0	1	10	10
DF	10	1	10	10	1	0	10	10
VASC	10	20	1	1	20	20	0	10
SCC	1	150	10	10	150	150	10	0

(b)

Figure 5.1: Cost matrices for the classes of the ISIC 2019 dataset. (a) A symmetrical cost matrix, in which the costs of misclassification are equivalent. (b) An asymmetrical cost matrix, where the costs of misclassification are differentiated based on various factors.

Optimising classifiers for a particular cost matrix may seem like a viable solution, however, it is not recommended due to the potential changes in the cost matrix after implementation. This would necessitate the frequent retraining of the classifiers, which can become a tedious and time-consuming process.

Given a trained classifier that outputs a calibrated posterior distribution $P(x, \theta)$ over classes T for an image x using the model parameters θ, the expected costs of classification can be utilised to make a decision on the image's classification (Ferrer, 2022). In the binary classification scenario of benign ($t = 0$) and malignant ($t = 1$) classes, the expected cost of a benign classification is defined as $R_0 = C_{10}P(t = 1|x)$, while the expected cost of a malignant classification is expressed as $R_1 = C_{01}P(t = 0|x)$. The image x is classified as malignant if $R_1 < R_0$, otherwise it is classified as benign. In the case of multiple classes, the expected cost for each class is calculated as the sum of costs incurred for each class assuming it to be the true class, and the class $\hat{j}$ with the minimum expected cost is selected as the final decision (Equation (5.1)).

$$\hat{j} = \arg\min_{j} \sum_{t=1}^{T} C_{tj} P(t|x, \theta) \tag{5.1}$$

5.4 Selective Classification Methods

5.4.1 Predictive Probabilities

The softmax response selection function $\sigma_{SR}(x)$ is computed by determining the maximum value of the prediction function $P(x, \theta)$ with a symmetrical cost matrix. This assumes that the neural network model utilised has employed a softmax activation function to generate predictive probabilities. While this methodology is straightforward to implement and intuitive, it has limitations as the output probabilities from a softmax activation are not properly calibrated and do not reflect the uncertainty of the model, as demonstrated by Gal and Ghahramani (2016).

$$\sigma_{SR}(x) = \max_{t} P(t|x, \theta) \tag{5.2}$$

In scenarios where costs are asymmetrical, the expected costs of a classification decision, as denoted by Equation (5.1), can be utilised to formulate a selective classification decision function $\sigma_{EC}(x)$ through the employment of the cost matrix C.

$$\sigma_{EC}(x) = \min_{j} \sum_{t=1}^{T} C_{tj} P(t|x, \theta) \tag{5.3}$$

Temperature scaling is a method utilised for calibrating the output probabilities of a neural network model (as discussed in Section 4.3.2). This method was chosen due to its demonstrated effectiveness in calibrating predictive probabilities for medical images as well as its ease of implementation. The temperature scaling technique involves scaling the output logits of a neural network with a temperature value, which is optimised on a validation set during the model's training process. The temperature value is determined from the training epoch that exhibits the lowest validation loss. This technique can be incorporated into a selection function by modifying the prediction function $P(x, \theta)$ to include dividing the logits by a temperature value before applying a softmax activation (Equation (4.3)) resulting in $P_{TS}(x)$. Equation (5.4) shows how temperature scaled probabilities can be used for selective classification.

$$\sigma_{TS}(x) = \max_t P_{TS}(t|x, \theta) \tag{5.4}$$

5.4.2 Bayesian Uncertainty

Bayesian neural networks represent a promising approach for improving the accuracy of probabilistic predictions and for more effectively estimating uncertainty. This is achieved through the representation of model parameters using distributions that can then be sampled from using forward propagation of the network.

In this chapter, Bayesian neural networks are evaluated using two methods, namely Bayes-by-Backprop (Blundell et al., 2015) and Laplace Approximation (MacKay, 1992), for training the network. The resulting Bayesian neural network can be sampled M times and the average of the predictions (Equation (5.5)) can be used to produce more calibrated probabilities, as demonstrated by (Jospin et al., 2022). The uncertainty in the Bayesian neural network can be estimated through the variance of the predictive samples (Equation (5.6)) and can be employed as a method of selection.

$$\sigma_{AVG}(x) = \max_t \frac{1}{M} \sum_{m=1}^{M} P(t|x, \theta_m) \tag{5.5}$$

$$\begin{aligned} \sigma_{VAR}(x) &= \frac{\sum_{m=1}^{M} (P(t|x, \theta_m) - \mu)^2}{M - 1} \\ \mu &= \frac{1}{M} \sum_{m=1}^{M} P(t|x, \theta_m) \end{aligned} \tag{5.6}$$

Multiple methods exist for estimating the predictive uncertainty of Bayesian neural networks, including variation ratios (Freeman, 1965), which measure the spread of the distribution of sample predictions around the mode (Equation (5.7)).

$$\begin{gathered}\sigma_{VR}(x) = 1 - \frac{\sum_{m=1}^{M} \mathbb{1}(\arg\max_t P(t|x,\theta_m) = \hat{t})}{M} \\ \hat{t} = \arg\max_t \sum_{m=1}^{M} \mathbb{1}(\arg\max_c P(c|x,\theta_m) = t)\end{gathered} \tag{5.7}$$

Predictive entropy (Shannon, 1948) captures the average information content of the distribution of sample predictions (Equation (5.8)).

$$\sigma_{PE}(X) = -\sum_{t}^{T} = 1 \left(\frac{1}{M} \sum_{m=1}^{M} P(t|x,\theta_m) \right) \log \left(\frac{1}{M} \sum_{m=1}^{M} P(t|x,\theta_m) \right) \tag{5.8}$$

Mutual information (Houlsby et al., 2011) quantifies the relationship between the predictive samples and the posterior distribution over the parameters of the model (Equation (5.9)).

$$\sigma_{MI}(x) = \sigma_{PE} + \frac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{T} P(t|x,\theta_m) \log P(t|x,\theta_m) \tag{5.9}$$

5.4.3 SelectiveNet

In the context of neural networks or Bayesian neural networks, data representations optimised for classification have been widely studied. However, Geifman and El-Yaniv (2019) posit that data representations can also be optimised for scenarios where a portion of the data is anticipated to be rejected. To address this issue, they introduce SelectiveNet, a modified training approach for neural networks that enables end-to-end optimisation for a specific target coverage (the probability mass of the non-rejected images).

This is achieved by adding two heads to the model's encoder, in addition to the predictive head (denoted as $P(x)$). These heads consist of a selective head (denoted as $G(x)$) that outputs a selection score and an auxiliary head (denoted as $A(x)$) that provides predictions used within the loss function. The overall loss function used to optimise the entire model is based on selective risk and balances the predictive and selective heads against the auxiliary head to ensure that robust features for classification are learned while still optimising for target coverage. The SelectiveNet loss function (Equation (5.10)) is a combination of two functions ($L_{p,g}$ and L_a), weighted by a hyperparameter α to control the relative importance of coverage optimisation.

$$L = \alpha L_{p,g} + (1 - \alpha) L_a \tag{5.10}$$

The first term uses both the predictive and selective heads (Equation (5.11)) and combines cross-entropy loss l with coverage ϕ (Equation (5.12)). For selective classification, the output of the selective head (Equation (5.13)) is utilised. The hyperparameter k represents the target coverage for the model, while λ regulates the significance of this target coverage. On the other hand, the auxiliary head uses a standard cross-entropy loss (L_a) to encourage the model to learn robust features from the training data.

$$L_{p,g} = \frac{\sum_{n=1}^{N} l(p(x^n, \theta), y) g(x^n, \theta)}{\phi} + \lambda \cdot \max(0, k - \phi)^2 \tag{5.11}$$

$$\phi = \frac{1}{N} \sum_{n=1}^{N} g(x^n, \theta) \tag{5.12}$$

$$\sigma_{SN}(x) = G(x) \tag{5.13}$$

5.4.4 Expected Cost SelectiveNet

Expected costs serve as a method for selection in both the CNN and the SelectiveNet model, as evidenced by Equation (5.1). A new approach to selection is proposed, referred to as Expected Cost SelectiveNet, which is based on expected costs computed from the predictive head, instead of the selective head output utilised in SelectiveNet.

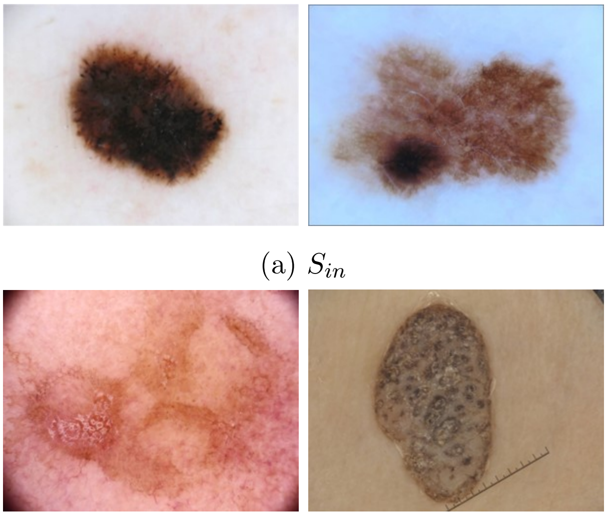
Despite the fact that SelectiveNet directly outputs a selection score, the proposed EC-SelectiveNet method utilises the expected costs computed from the predictive head for selection. The selective head is used during training to guide representation learning but, in contrast to the approach presented in Geifman and El-Yaniv (2019), both the selective head and auxiliary head are discarded at test time.

5.5 Binary Classification Experiments

This section details the datasets, training parameters, experimental setup and results for the experiments with binary asymmetric selective classification for skin lesion triage. The code and full results used within this section can be found on the project GitHub repository [1].

5.5.1 Dataset Processing

The ISIC Challenge 2019 (Codella et al., 2018, Combalia et al., 2019, Tschandl et al., 2018) was employed in this chapter and consists of a total of 25,331 images spanning eight distinct classes, including melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma. For the purposes of the experiments, two datasets were compiled from the ISIC 2019 data, referred to as S_{in} and $S_{unknown}$.



(a) S_{in}

(b) $S_{unknown}$

Figure 5.2: Example images from the test data sets S_{in} and $S_{unknown}$.

S_{in}: These data encompassed the melanoma, melanocytic nevus, and basal cell carcinoma (BCC) images from the ISIC 2019 dataset, which were assigned to two classes for classification: malignant (melanoma, BCC) and benign (melanocytic nevus). The S_{in} dataset was split into three subsets for training, validation, and testing, containing 12432, 3316, and 4972 images, respectively.

$S_{unknown}$: These data consisted of 4,360 images from classes that were not present in S_{in}, including benign keratosis, dermatofibroma, actinic keratosis, and squamous cell carcinoma, and were assigned to either the malignant or benign class. The $S_{unknown}$ dataset was not utilised for training, but instead was employed to test the performance of selective

[1] GitHub Repository: `github.com/UoD-CVIP/Selective_Dermatology`

classification on images from disease types not represented in the training data.

The combination of the S_{in} and $S_{unknown}$ test sets is referred to as the $S_{combined}$ dataset. Figure 5.2 provides illustrative examples from the ISIC 2019 dataset. In the current chapter, a random split strategy was employed to divide the dataset into three distinct sets: training, validation, and testing. To ensure comparability across all images, normalisation was performed utilising the standard deviation and mean calculation for each colour channel of the images. Subsequently, each image underwent resizing to 256x256, and during the training phase, data augmentation was carried out by implementing randomised horizontal and vertical flips and rotations of multiples of 90°.

5.5.2 Experiment Setup

Eight models were trained on the training split of S_{in}, utilising the validation set to identify the optimal model from the training epochs. The performance of each selection method was then evaluated by using selection methods with the appropriate model. A detailed overview of the experimental setup is presented in Table 5.1. The evaluation of the selection methods on the trained models was carried out utilising a symmetrical cost matrix that is commonly used in such evaluations, where the cost of false positives and false negatives was set to 1.0. To investigate the effect of adjusting the level of asymmetry, the cost of false positives was set to 10.0 and 50.0 while keeping the cost of false negatives fixed at 1.0. The evaluation of the selection methods was conducted on three different datasets: S_{in} was used to evaluate the in-distribution performance, $S_{unknown}$ was used to assess the generalisation performance on unknown types of skin lesions, and $S_{combines}$ used to evaluate the joint performance on a test set containing in-distribution and unknown types of skin lesions.

5.5.3 Training Parameters and Model Architecture

The implementation of the conventional convolutional neural network comprises an EfficientNet (Tan and Le, 2019) encoder with a compound coefficient of 7. This encoder is followed by an average pooling operation that reduces the width and height by a factor of 8, thereby compressing the encoding size from 163,840 to 2560. Subsequently, the architecture includes a hidden layer equipped with 512 neurons and a rectified linear

Table 5.1: Binary experiments; trained models and selection methods to be evaluated.

Model	Selection Method
CNN	Softmax Response Temperature Scaled Softmax Response
SelectiveNet $k = 0.7$ SelectiveNet $k = 0.75$ SelectiveNet $k = 0.8$ SelectiveNet $k = 0.85$ SelectiveNet $k = 0.9$ SelectiveNet $k = 0.95$ SelectiveNet $k = 1.0$	SelectiveNet EC-SelectiveNet Temperature Scaled EC-SelectiveNet
Monte Carlo Dropout	Average Softmax Response Average Prediction Variance

unit activation function, which is followed by a final output layer with 2 output neurons and a softmax activation function. To mitigate overfitting, dropout regularisation is applied with a drop chance of 0.5 before and after the hidden layer. In total, the architecture encompasses 275 layers (convolutional and fully connected) with 65,099,224 total parameters. Additionally, the use of dropout during training enables the simulation of a Bayesian neural network through Monte Carlo dropout (Gal and Ghahramani, 2016) by sampling the trained model with the same dropout rates.

The SelectiveNet model architecture (Geifman and El-Yaniv, 2019), is constructed upon an EfficientNet encoder, followed by average pooling and a fully connected hidden layer containing 512 neurons. The prediction head of the model comprises a single fully connected output layer with 2 neurons and a softmax activation function. The selection head is composed of an additional hidden layer with 512 neurons and a softmax activation function, which is followed by an output layer with a single neuron and a sigmoid activation function. The auxiliary head is similarly structured to the classification head. These additional components result in an increased number of layers in the SelectiveNet architecture, bringing the total to 277 layers, with 65,364,449 parameters in total.

The training configurations were standardised across all experiments. The models were trained utilising 16-bit precision to compute gradients, and Stochastic Gradient Descent was employed to optimise the model parameters. The optimisation process employed a triangular cyclical scheduler (Smith, 2017) that cyclically adjusted the learning rate between 0.00001 and 0.1 and the momentum between 0.8 and 0.9 every 2000 training steps. The experiments were performed using mini-batches of 8 images each.

During the training of the conventional convolutional neural network, the cross-entropy loss function was utilised to evaluate the model's performance. To mitigate overfitting, dropout regularisation was applied with a drop rate of 0.5 before both the hidden and output layers. By consistently applying the same dropout pattern during both training and evaluation, the convolutional neural network can be treated as a Bayesian Neural Network and sampled multiple times through Monte Carlo dropout (Gal and Ghahramani, 2016).

The SelectiveNet model's performance is evaluated using a loss function that integrates the outputs of the three heads, predictive (p), selective (g), and auxiliary (h). The predictive and selective heads are utilised to calculate a portion of the loss $L_{p,g}$, which optimises the model for a specific target coverage. The auxiliary head, in contrast, calculates cross-entropy loss L_h, and the two components are weighted together using the parameter α (as specified in Equation 5.10). In the present set of experiments, the parameter α has been set to 0.5 and multiple target coverages have been explored. The remaining hyperparameters associated with the SelectiveNet loss function have been set to the values recommended by the authors in (Geifman and El-Yaniv, 2019).

5.5.4 Results

The findings presented in this section are not exhaustive, and complete cost coverage curves for all models and methods of selective classification are available in Appendix B.

SelectiveNet: Effect of Target Coverage

In order to examine the impact of the SelectiveNet target-coverage parameter, t, on the selection decisions made by the SelectiveNet selection head, the cost-coverage curves were plotted for various values of t, ranging from 0.7 to 1.0 with increments of 0.05, as depicted in Figure 5.3. The curves were computed for S_{in}, $S_{unknown}$, and $S_{combined}$. According to the design objectives of the target coverage parameter, a lower value of t would result in a more effective model at lower coverages. The findings, however, show that training with a $t = 1.0$ value resulted in the lowest test cost on S_{in} for coverage values as low as 0.2. As expected, the expenses incurred on $S_{unknown}$ were larger, and the curves did not show a clear ordering. Nonetheless, the model trained with $t = 1.0$ revealed a significant cost reduction as coverage declined. This could be because when the model is trained with a $t = 1.0$, the resulting model is closer to the best performing standard model with softmax response.

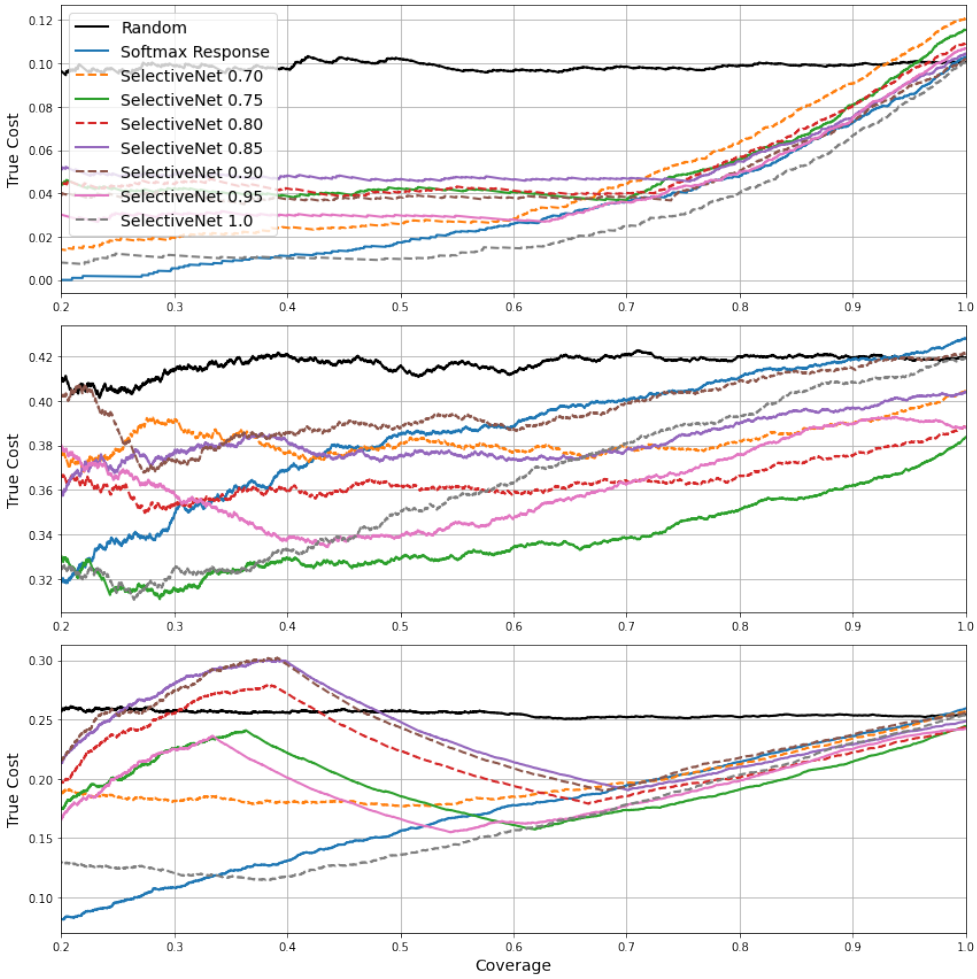


Figure 5.3: Cost-coverage curves for SelectiveNets trained with different target coverages. From top to bottom: S_{in}, $S_{unknown}$ and $S_{combined}$.

Does SelectiveNet Training Help?

The extent to which the target coverage t is imposed is regulated by the weighting parameter λ. Despite being set to target full coverage ($t = 1.0$), the model may, in exceptional circumstances during training, compromise coverage for cost. This could be because when the SelectiveNet model is trained with $t = 1.0$, penalties in the loss function promote coverage to reach 1.0 (see Equation 5.11) Even with these constraints in place, the model can still compromise coverage for cost if the benefit surpasses the penalty. As a result, the results obtained by SelectiveNet with $t = 1.0$ may differ from those obtained through training a network without selective and auxiliary heads. These networks were trained using cross-entropy loss and only retained the softmax predictive head, making selection decisions at test time based on the maximum softmax output. The corresponding cost-coverage curve is plotted in Figure 5.3 (labelled "softmax"). The results indicate that SelectiveNet trained with a target coverage of 1.0 performed better than a standard convolutional neural network with a softmax response for any coverage value above 0.4.

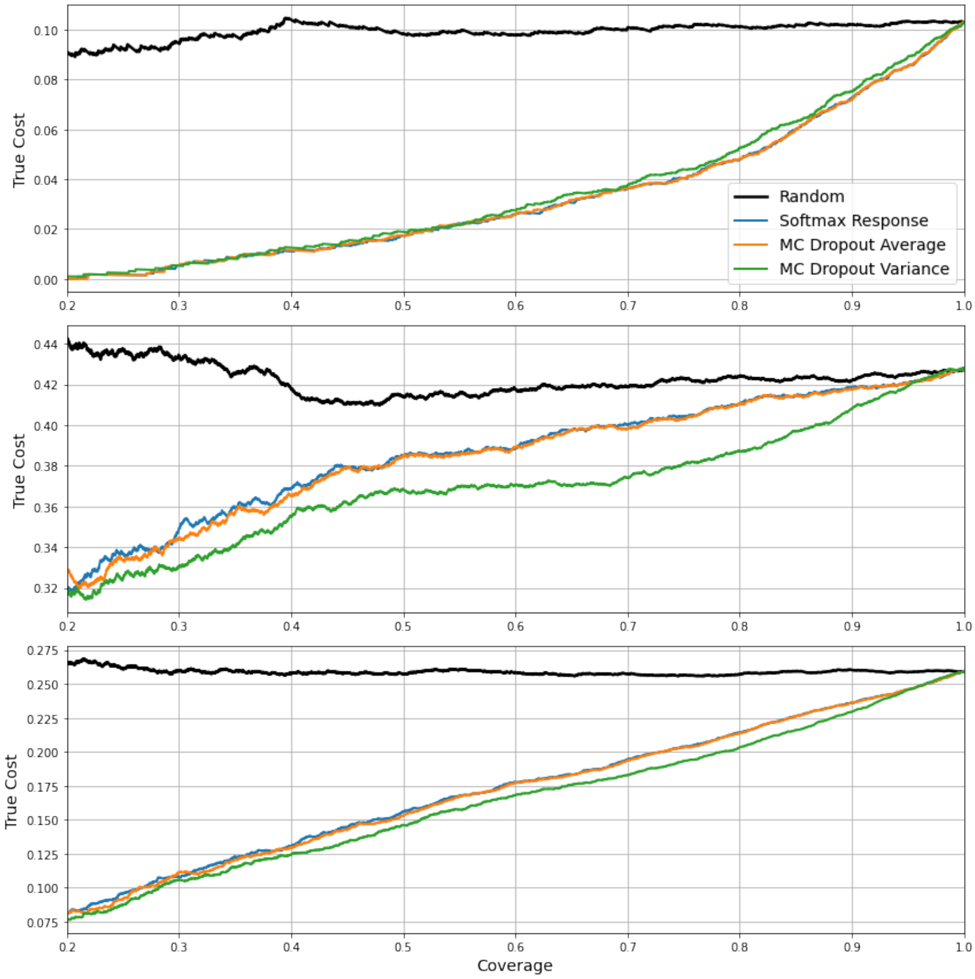


Figure 5.4: Cost-coverage curves using MC-Dropout. From top to bottom: S_{in}, $S_{unknown}$ and $S_{combined}$.

MC-Dropout, Temperature Scaling, and EC-SelectiveNet

The impact of MC-Dropout on selective classification was analysed by using the mean and variance of the Monte Carlo iterations as selection scores. Figure 5.4 compares the resulting cost-coverage curves to those obtained using a network without dropout at test time (labelled "softmax response"). The results reveal that for the S_{in} data, utilising the average of Monte Carlo samples had minimal impact, whereas the variance of the Monte Carlo samples performed slightly worse than simply relying on the maximum softmax response. Conversely, significant cost savings were achieved by using the variance of the Monte Carlo samples on the $S_{unknown}$ data, where model uncertainty is expected to be high.

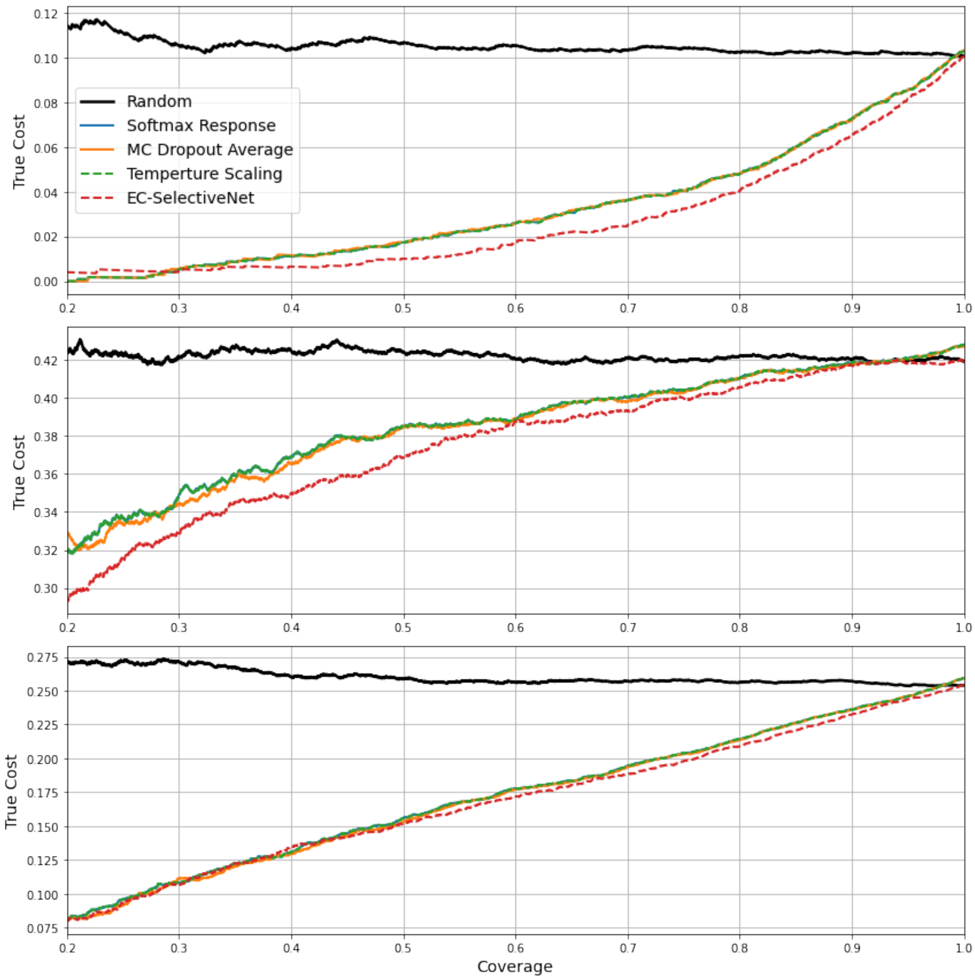


Figure 5.5: Cost-coverage curves for different selective classification methods. From top to bottom: S_{in}, $S_{unknown}$ and $S_{combined}$.

The effect of temperature scaling on a softmax network was analysed and the results are shown in Figure 5.5. The softmax network was trained using cross-entropy loss and temperature scaling was applied to improve calibration. However, the results indicated that temperature scaling had a minimal effect on the cost-coverage curves. While temperature scaling improves calibration, it does not capture enough information to provide a reliable failure warning, which can then be employed in rejection, as observed in other works (Jaeger et al., 2022). Furthermore, Figure 5.5 displays the results obtained using EC-SelectiveNet, in which the selection head was omitted during testing. The results demonstrate that EC-SelectiveNet achieved a noticeable improvement on both the S_{in} and $S_{unknown}$ datasets when compared to training a standard convolutional neural network model without the auxiliary heads.

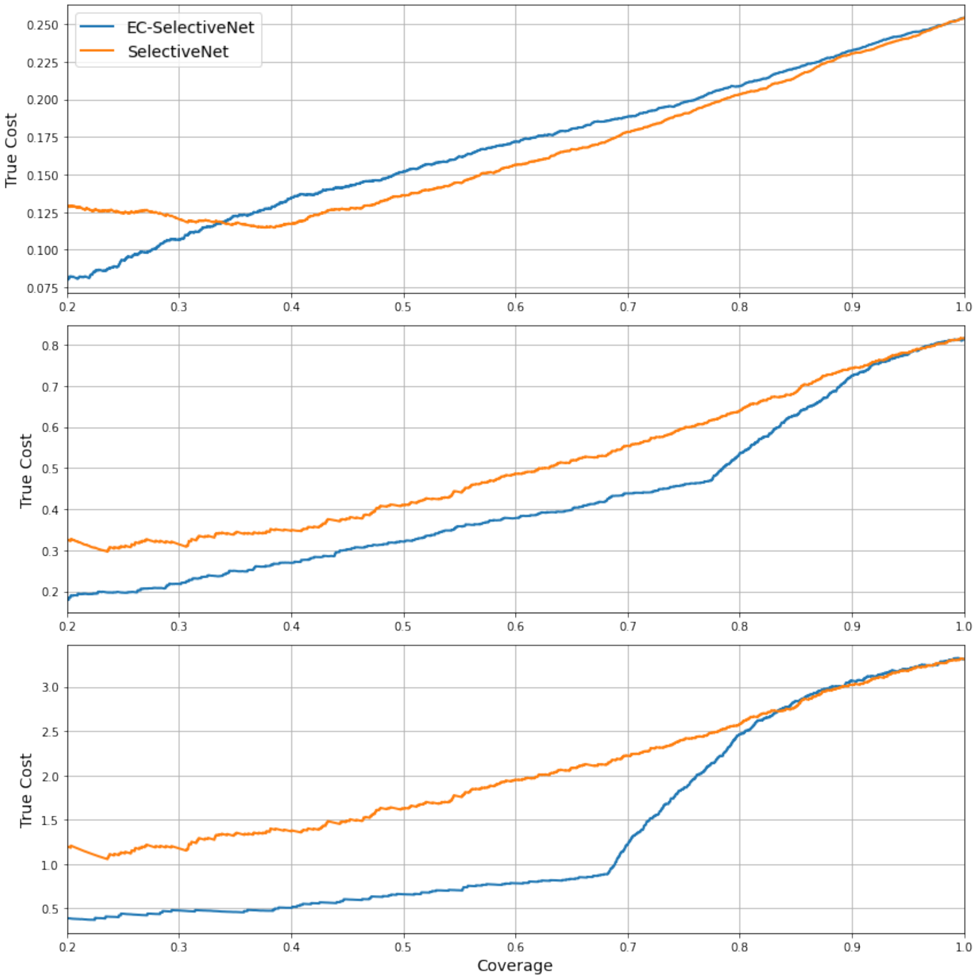


Figure 5.6: Cost-coverage curves for SelectiveNet and EC-SelectiveNet. From top to bottom: $C_{1,0} = 1$, $C_{1,0} = 10$ and $C_{1,0} = 50$.

Asymmetric Costs

The results of comparing SelectiveNet with EC-SelectiveNet with a target coverage of $t = 1.0$ are depicted in Figure 5.6. In symmetric cost scenarios, the performance of both methods was comparable, with SelectiveNet exhibiting a slight advantage in terms of cost, yielding a reduction of approximately 0.015 at intermediate coverage levels. However, when the cost matrix was asymmetric, EC-SelectiveNet demonstrated significant cost reductions of approximately 0.1 at all coverages below approximately 0.8.

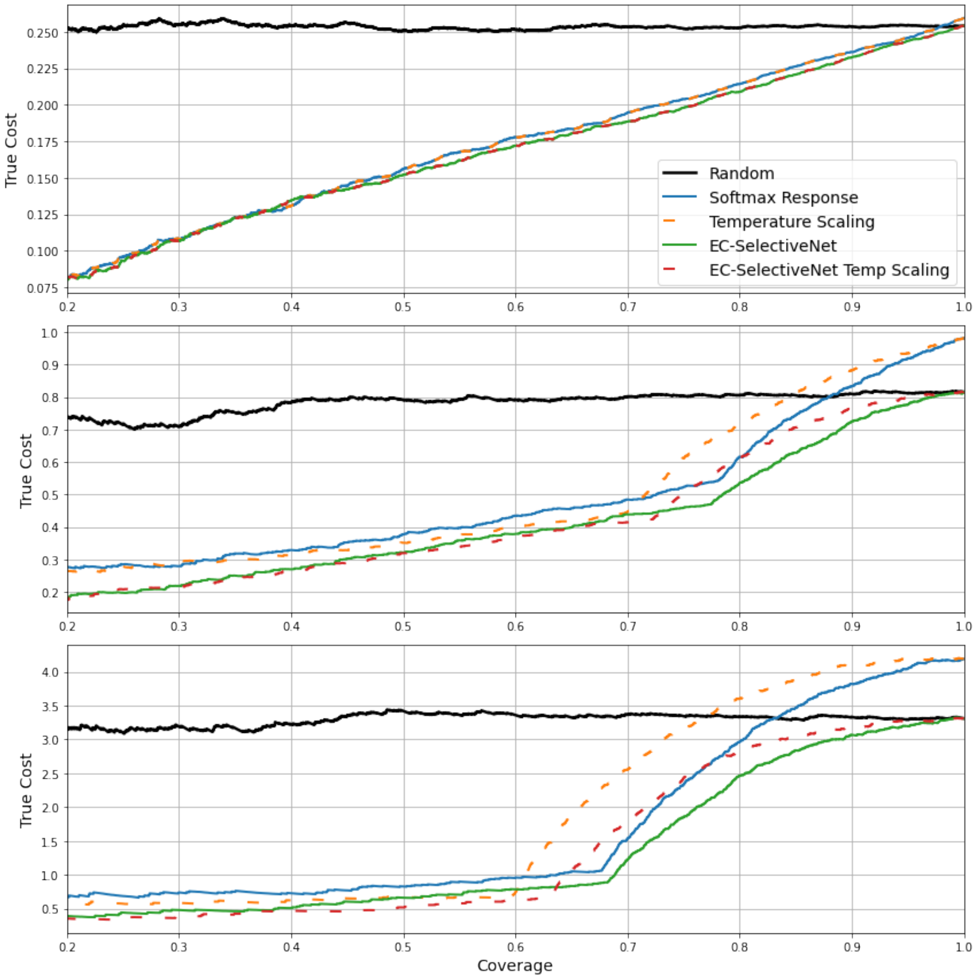


Figure 5.7: Cost-coverage curves for cross-entropy training and EC-SelectiveNet combined with temperature scaling. From top to bottom: $C_{1,0} = 1$, $C_{1,0} = 10$ and $C_{1,0} = 50$.

The effect of temperature scaling on selective classification is presented in Figure 5.7. The methodologies of both the softmax response and temperature scaling selection are founded on the principle of expected costs. The impact of temperature scaling was found to be minimal in the context of symmetrical costs, as discussed above. In the asymmetrical cost matrix scenario, a slight effect on selective classification was observed. This effect was consistent regardless of whether EC-SelectiveNet ($t = 1.0$) or a convolutional neural network trained with cross-entropy loss was employed. As depicted in Figure 5.7, the application of temperature scaling resulted in an increase in costs at high coverage levels and a reduction in costs at low coverage levels. This could be because the temperature parameter used for temperature scaling was designed for a symmetrical environment, and its application in asymmetrical contexts causes problems that worsen as the asymmetry increases. The figure also highlights the superiority of EC-SelectiveNet in comparison to temperature scaling.

5.6 Multi class Experiments

This section details the datasets, training parameters, experimental setup and results for the experiments with multi-class asymmetric selective classification for skin lesion classification. The code and full results used within this section can be found on the project GitHub repository [2].

5.6.1 Dataset Processing

In this chapter, the ISIC Challenge 2019 dataset (Codella et al., 2018, Combalia et al., 2019, Tschandl et al., 2018) was employed. The data was processed using a similar methodology to the one described in Section 5.5.1. A random splitting method was utilised to divide the dataset into three subsets: training, validation, and testing, in a 60:20:20 ratio. Prior to the training process, the images underwent normalisation, which involved computing the standard deviation and mean across each colour channel. Then, the images were square cropped by evenly trimming the horizontal sides and resized to 256x256. During the training phase, data augmentation was applied by randomly augmenting the images at each epoch. The augmentations consisted of 90-degree rotations and horizontal and vertical flips.

[2] GitHub Repository: github.com/UoD-CVIP/Asymetric_Selective_Dermatology

5.6.2 Experiment Setup

In the multi-class set of experiments, all models and methods were subjected to repetition thrice, with the final results being an average of the outcomes from each run. In order to ensure a comprehensive evaluation, the training, validation, and testing splits were randomised for each of the three repetitions. A total of eleven models were trained, utilising the validation set to determine the most suitable model. Table 5.2 shows the different trained models and the selection methods evaluated on each model.

Table 5.2: Multi-class experiments; trained models and selection methods to be evaluated.

Model	Selection Method
CNN	Softmax Response Expected Costs Temperature Scaled Softmax Response Temperature Scaled Expected Costs
SelectiveNet $k = 0.7$ SelectiveNet $k = 0.75$ SelectiveNet $k = 0.8$ SelectiveNet $k = 0.85$ SelectiveNet $k = 0.9$ SelectiveNet $k = 0.95$ SelectiveNet $k = 1.0$	SelectiveNet Softmax Response Temperature Scaled Softmax Response EC-SelectiveNet Temperature Scaled Expected Costs
Monte Carlo Dropout Bayes By Backprop Laplace Approximation	Average Softmax Response Average Expected Costs Bayesian Sample Agreement Average Variance Predicted Class Variance Predictive Entropy Variational Ratio Mutual Information

The efficacy of the selection methods employed with each of the models was evaluated using both a typical symmetrical cost matrix (Figure 5.1a) and an asymmetrical cost matrix (Figure 5.1b). The asymmetrical cost matrix was developed by a consultant-level dermatologist, taking into account the clinical costs associated with misdiagnosis such as death or unnecessary treatment. The values assigned to the asymmetric costs are context-dependent and should be arrived at through engagement with relevant experts, taking into consideration the specific clinical setting.

5.6.3 Training Parameters and Model Architecture

The architecture of the CNN and SelectiveNet in this chapter was consistent with the binary experiments outlined in Section 5.5.3. The number of output neurons in the output layer of the CNN and the predictive and auxiliary heads of the SelectiveNet architecture, was increased to 8, resulting in a total of 65,102,296 weights for the CNN model and 65,370,593 weights for the SelectiveNet model.

Additionally, the Bayes-by-Backprop model (Blundell et al., 2015) utilised in this chapter was implemented as a CNN, with the final two fully connected layers replaced by fully connected Bayesian layers. In these layers, each weight is represented by a normal distribution with a mean randomly selected from a range of 0 to 0.1 and a standard deviation sampled from a range of -7 to 0.1. The weight priors $P(\theta)$ was modelled using a scale mixture of two Gaussian distributions (Equation (5.14)) where J represented the number of weights in a model with standard deviation values (σ_1 and σ_2) of 0.1 and 0.4 and a π weighting the two Gaussian distributions of 0.5.

$$P(\theta) = \prod_{j=1}^{J} \pi M(\theta_j|0, \sigma_1) + (1 - \pi) M(\theta_j|0, \sigma_2) \tag{5.14}$$

The CNN and SelectiveNet models were trained with the same weights as those presented in the binary experiments described in Section 5.5.3. The Bayesian Neural Network model, which was trained using the Bayes-by-Backprop method (Blundell et al., 2015), employed variational inference to approximate the posterior distribution over weights, $q(\theta)$. The weights for the distribution, θ, were determined by minimising the KL-Divergence between the variational posterior and the true posterior. The true posterior was estimated through Monte Carlo sampling of the evidence lower bound, as expressed in Equation (4.4), where D represents the dataset and M denotes the number of Monte Carlo samples.

The loss function utilised for training the Bayesian Neural Network model was a combination of the ELBO and cross-entropy. The weight assigned to the ELBO component of the loss function was modulated based on the current mini-batch, as depicted in Equation 5.15, where N represents the number of mini-batches per epoch and n denotes the current mini-batch. This weighting approach was employed such that the early mini-batches were more influenced by Bayesian complexity and later mini-batches focused more on learning from the training data. In the experiments, the Bayesian neural network model trained with the Bayes by Backprop method was trained using a weighted combination of ELBO and cross-entropy, with the ELBO estimated using three Monte Carlo samples during training. All other training settings were equivalent to those utilised for the training of the convolutional neural network model.

$$\pi_n = \frac{2^{N-n}}{2^N - 1} \tag{5.15}$$

After training, a CNN model can be transformed into a Bayesian neural network using the Laplace approximation method (MacKay, 1992). The Laplace approximation is a technique for approximating the posterior of a model as a Gaussian distribution centred on the learned weights. The curvature of the approximation is estimated through the use of approximations to the Hessian matrix (Botev et al., 2017) at the MAP. In the experiments, Laplace approximation was applied only to the last layer of the neural network due to hardware constraints, however, it has been demonstrated by Kristiadi et al. (2020) that this approach can lead to improved calibration and estimation of predictive uncertainty. After performing the Laplace approximation, a predictive probability can be computed by averaging Monte Carlo samples.

5.6.4 Results

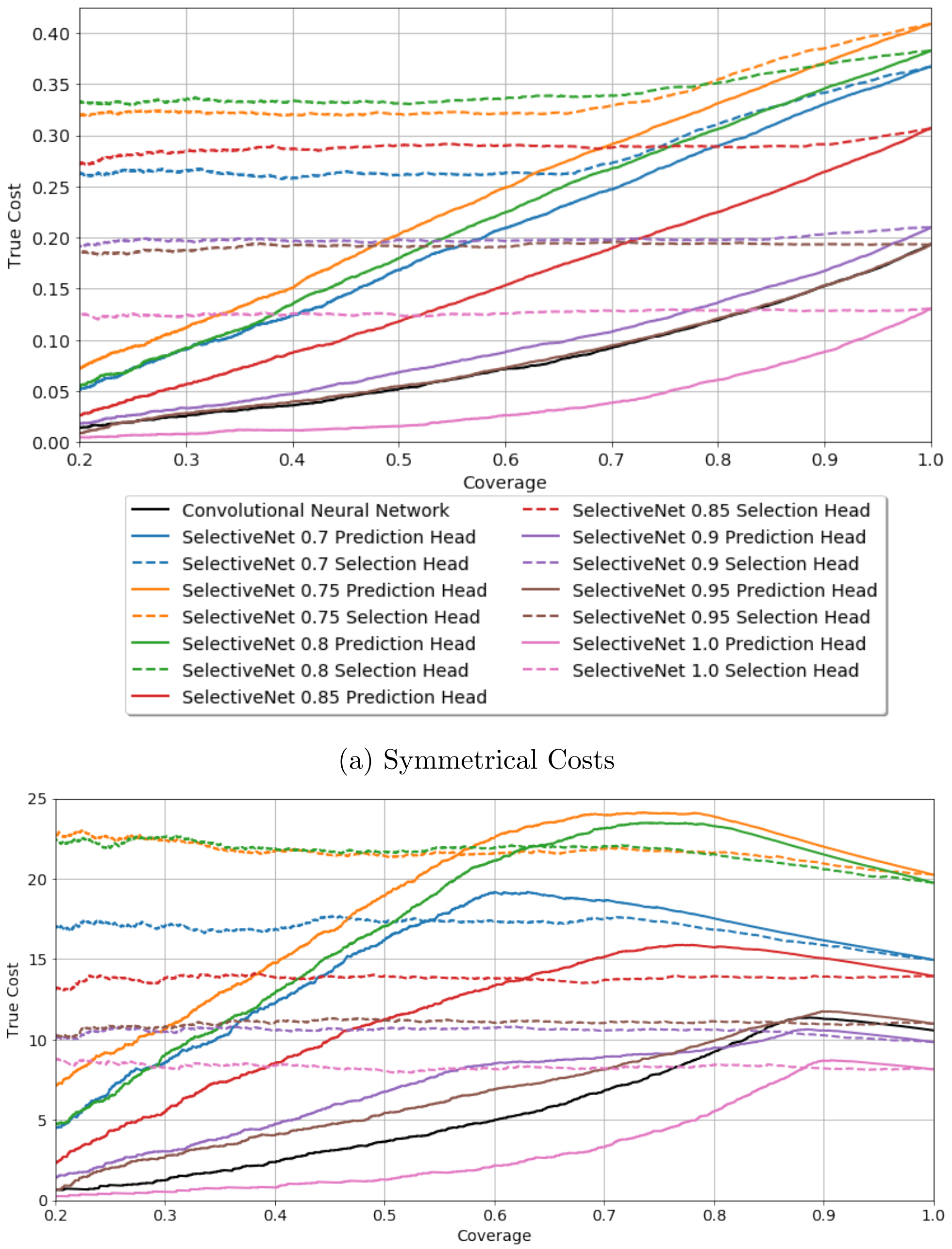
The findings presented in this section are not exhaustive, and complete cost coverage curves for all models and methods of selective classification are available in Appendix B.

SelectiveNet, EC-SelectiveNet and Target Coverage

Figure 5.8 illustrates the outcomes of the investigation carried out on SelectiveNet, which was trained with different target coverages ranging from 0.7 to 1.0. The utilisation of the selective head of the SelectiveNet model resulted in suboptimal performance for selective classification in both settings. In contrast, using the predictive head in symmetrical settings led to improved performance compared to the selective head, resulting in a similar curve shape as that of the CNN. However, it is worth noting that the base performance, and consequently, the area under the cost coverage curve, was worse in all cases except for SelectiveNet trained with a target coverage of 1.0. This observation is consistent with the results of the binary classification experiments and can be ascribed to the SelectiveNet model's higher number of weights, which enables it to learn a superior model when trained with a target coverage of 1.0, as its learning is not constrained. In an asymmetrical cost scenario, the use of predictive heads to select based on estimated cost (EC-SelectiveNet) yielded similar results to those of the binary experiments, where the cost coverage curves trended upwards until they approached their target coverage and then demonstrated better performance. The selective head's poor performance in multi-class asymmetrical circumstances compared to binary settings could be attributed to a more difficult environment in which the model struggled to learn how to reject these images. However, the predictive head did not suffer as much in the transition from binary to multi-class settings, probably due to the presence of more decision boundaries, which provided the predictive head with more rejection information.

Effect of Temperature Scaling

The outcomes of the experiments involving the convolutional neural network and EC-SelectiveNet (trained with a target coverage of 1.0) in both symmetrical and asymmetrical cost scenarios are illustrated in Figure 5.9. The choice of the EC-SelectiveNet model with a target coverage of 1.0 was predicated on its superior performance relative to other SelectiveNet models, as evidenced in Figure 5.8. The findings demonstrate that the utilisation of temperature scaling to improve prediction calibration had a negligible impact on the cost coverage curves in symmetrical cost settings. However, in multi-class situations characterised by asymmetrical costs, temperature scaling had no effect on the convolutional neural network but exhibited a detrimental effect when used in conjunction with the EC-SelectiveNet model. This phenomenon could conceivably be attributed to the process of temperature scaling, wherein predictions are systematically calibrated within a symmetrical setting. When presented to an asymmetrical setting this might inadvertently entail suboptimal performance, with the performance gap increasing as the setting becomes more asymmetrical.



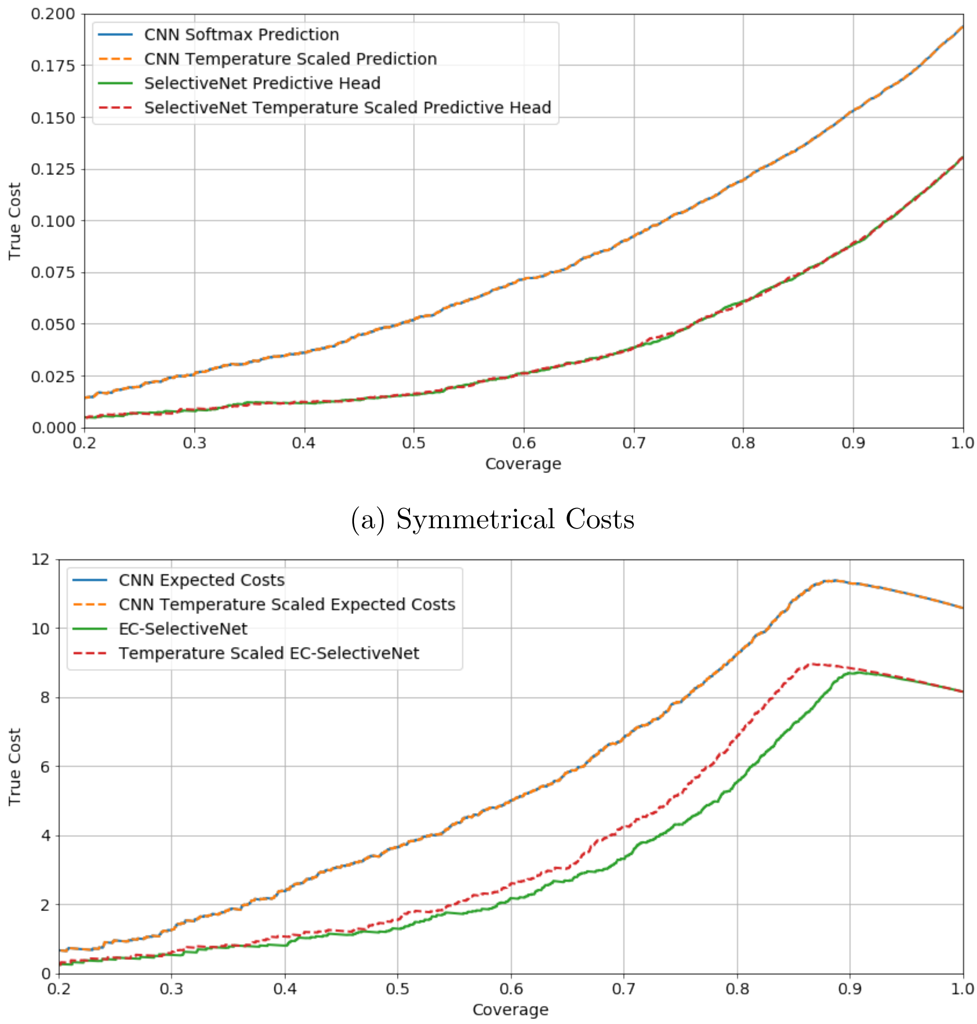
(a) Symmetrical Costs

(b) Asymmetrical Costs

Figure 5.8: Results with experiments with SelectiveNet.

Methods Bayesian Neural Networks

In the multi-class experiments, three distinct implementations of Bayesian neural networks were examined. For each Bayesian neural network, 100 samples were taken, and the selection was based on the average of the Monte Carlo samples. The variance of the samples is presented in Figure 5.10. In a symmetrical cost scenario, both the sample average and variance of Monte Carlo dropout and Laplace approximation performed

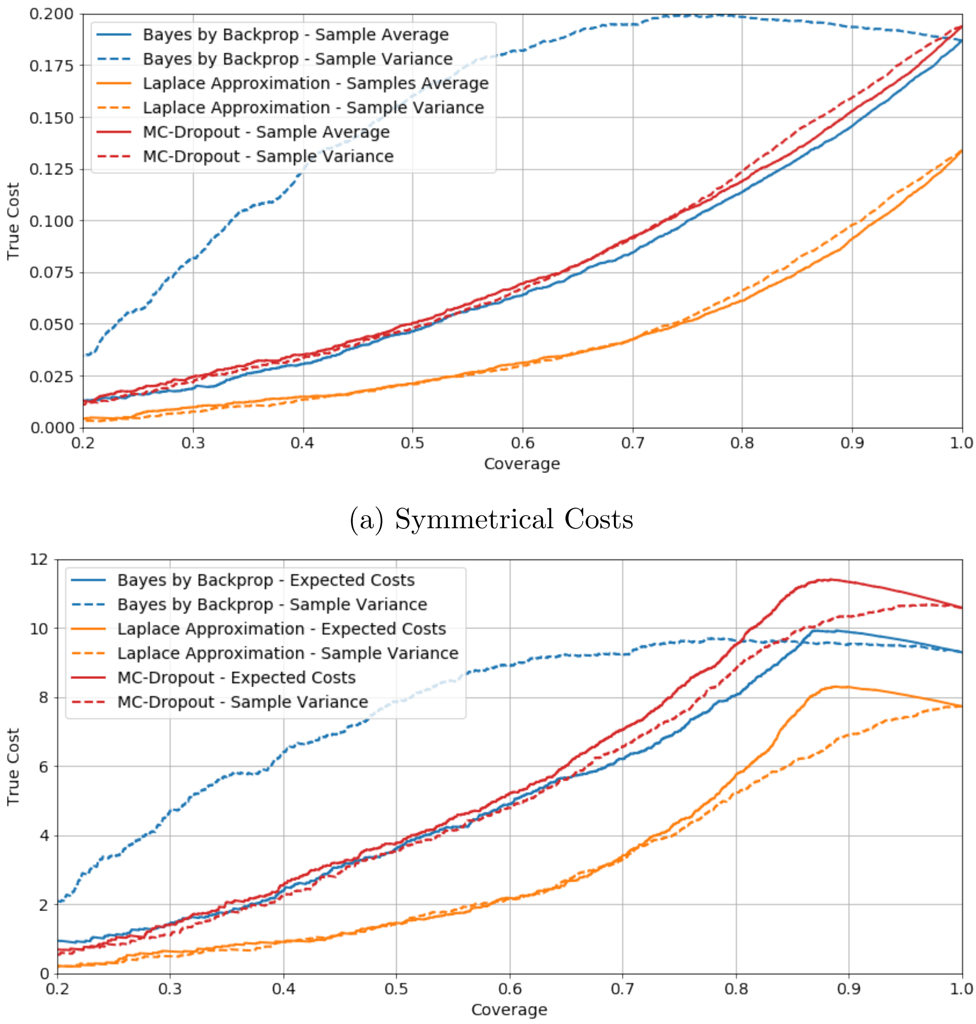


(a) Symmetrical Costs

(b) Asymmetrical Costs

Figure 5.9: Results with experiments with temperature scaling.

similarly, with the average slightly outperforming the variance. However, the model trained using Bayes by Backprop exhibited very poor performance when using the sample variance, in comparison to the average. In the asymmetrical cost scenario, the variance was found to be a significantly better method for selective classification than the expected costs of the sample average. The disparity in performance for the Monte Carlo dropout and Laplace approximation methods was evident in the initial 100% to 80% coverage before the curves converge. This was not the case with Bayes by Backprop, as the performance with sample variance remained considerably worse than expected costs using sample average.



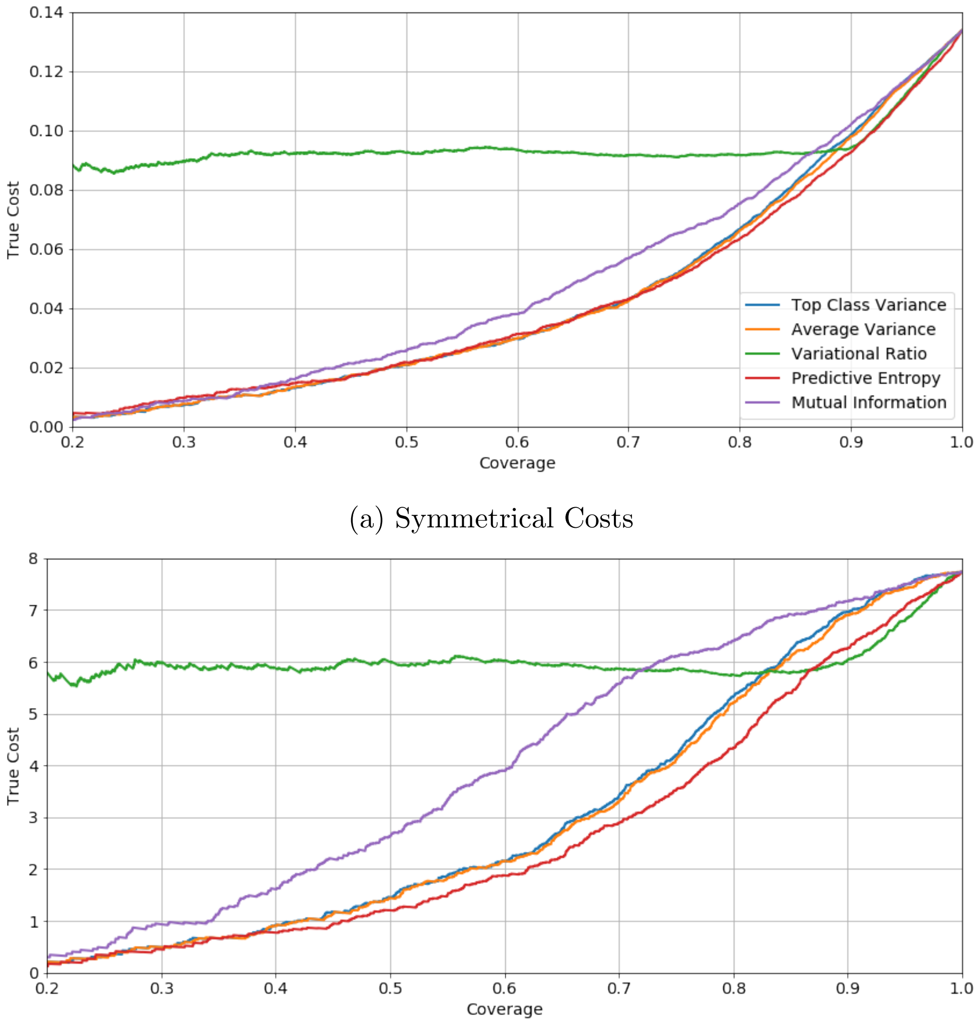
(a) Symmetrical Costs

(b) Asymmetrical Costs

Figure 5.10: Results with experiments with Bayesian Neural Networks.

Measures of Uncertainty

In the multi-class experiments, various measures of uncertainty were employed with Bayesian neural networks' Monte Carlo samples. The results presented in Figure 5.11 depict the outcomes of selective classification with Laplace Approximation models, as they demonstrated the best performance, as shown in Figure 5.10. The results indicate that the two measures of variance (average over all classes and top class only) perform comparably, with average variance having a slight edge. Predictive entropy emerged as the best-performing measure of uncertainty for selective classification in both symmetrical and asymmetrical cost settings. While mutual information exhibited inferior performance when compared to variance and predictive entropy, it performed significantly worse in an asymmetrical setting. The variational ratio was the poorest measure, it displayed similar performance to other metrics until 90% coverage where the performance flattened meaning that the selections to reject are random.



(a) Symmetrical Costs

(b) Asymmetrical Costs

Figure 5.11: Results with experiments with measures of uncertainty.

5.7 Conclusion

The aim of this chapter was to enhance comprehension of the performance of various selective classification methods for skin lesion images, utilising asymmetrical costs in both binary (triage setting) and multi-class (disease) classification contexts. Additionally, the effectiveness of selective classification techniques when dealing with lesion types not present in the training data was investigated. The experimental results indicate that SelectiveNet, in general, was less effective compared to other selective classification methods, and only exhibited improvement in performance when trained with a target coverage of 1.0. Furthermore, in cases where SelectiveNet was trained with a target coverage of 1.0, its prediction head performed better than the CNN performance. In the asymmetrical setting, EC-SelectiveNet, trained with a target coverage of 1.0, consistently outperformed all other methods in both binary and multi-class settings. This could be due to the selective part of the SelectiveNet loss function being ignored during training when a target of 1.0 is used. Consequently, using only the predictive and auxiliary heads jointly to generate a loss, allows for a better representative encoder.

The utilisation of Bayesian neural networks had a negligible impact when averaging the predictions in any setting. However, leveraging the variance of the Monte Carlo samples led to superior results in the context of asymmetric settings and yielded promising outcomes on the $S_{unknown}$ dataset. Of the three methods for Bayesian neural networks, Laplace approximation exhibited the best performance, while Bayes by Backprop surprisingly performed well when using the variance. It can be speculated that the samples taken from the Bayes by Backprop were too similar, making the variance unsuitable for selective classification purposes. Notably, the use of different uncertainty measures led to varying results, with the predictive entropy measure of uncertainty surpassing all others, particularly in an asymmetric setting, while the variational ratios performed the worst in both symmetric and asymmetric cost settings.

The experiments reveal that the utilisation of temperature scaling for the calibration of predictions, which aims to enhance selective classification, resulted in elevated costs at higher coverage levels in asymmetrical cost settings. This problem could be ascribed to temperature parameter optimisation, which has been specifically optimised for symmetrical contexts, making scaled predictions unsuitable for addressing asymmetrical circumstances. Addressing this issue could be a promising area for future study, involving an evaluation of calibration procedures customised to various asymmetrical settings. However, in other scenarios, the use of temperature scaling did not have any significant impact compared to the uncalibrated outcomes. As highlighted by Jaeger et al. (2022), this phenomenon could be explained by the intrinsic limitation of temperature scaling in as the scaled predictions are limited in their capacity to capture the inherent uncertainty of the model's predictions. Addressing this issue necessitates further investigation and subsequent efforts to correct this constraint. The investigation encompasses diverse selective classification settings and underlines the necessity for further efforts to advance selective classification methods and comprehend their performance in asymmetrical cost settings. Such efforts will prove instrumental in the application of classification in clinical settings, where asymmetrical costs are prevalent and not all images can be classified, necessitating the use of rejection.

The chapter's findings highlight the importance of enhancing selective classification in order to simplify its adoption into clinical practise. This requirement is reinforced by Jaeger et al. (2022) observations, which reveal the limitations of neural networks in addressing failure detection, emphasising the constraints of selective categorization. A promising path forward is to tailor calibration approaches to specific decision-making scenarios. While still in its infancy, this topic has potential (Zhao et al., 2021) which introduces novel calibration methodologies for decision-specific scenarios. This new research represents the shifting landscape of selective classification and foreshadows transformational developments that could alter its path.

Chapter 6

Evaluating Dataset Fine-Tuning

6.1 Introduction

Recent years have seen significant advancements in the use of deep learning for the dermatological classification of skin lesion images (Du-Harpur et al., 2020, Wu et al., 2022). Deep learning classifiers have been shown to be particularly effective with large datasets, such as the ISIC archive, which have enabled substantial progress in this field (Tschandl et al., 2018, Wen et al., 2021). Studies utilising high-quality datasets have reported performance that is comparable to or surpasses that of dermatologists in the classification of skin lesions (Esteva et al., 2017, Haenssle et al., 2018, Han et al., 2018, Tschandl et al., 2019). Additionally, research has investigated the ability of deep learning classifiers to accurately classify macroscopic clinical images, as opposed to solely dermoscopic images (Fujisawa et al., 2019). However, it should be noted that images acquired from primary care settings are often of variable quality, with wider and less consistent fields of view or focus, and may include visual distractions.

Current deep learning classifiers for medical images have been observed to exhibit poor generalisation across different healthcare systems, acquisition protocols, and patient populations. This phenomenon has been documented in several studies, with evaluations typically being conducted in an internal manner, where the model training and testing datasets are drawn from the same source (e.g., Han et al. (2018)). However, there remains a significant degree of uncertainty regarding the ability of these models to generalise across diverse domains, datasets, and imaging modalities.

Given the complexity of the medical imaging domain, it may be unrealistic to expect the development of a universally applicable skin lesion classifier. A more practical approach may involve the utilisation of local datasets to adapt models to specific target populations and healthcare systems. This approach would involve leveraging the knowledge acquired from these datasets to fine-tune existing models, thus increasing their applicability and performance within specific domains. This approach has been proposed in several studies and is considered a more realistic and practical solution to the challenges of generalisation in deep learning-based medical image analysis (Glocker et al., 2022).

The phenomenon of domain shift, wherein the distribution of features in the new population diverges from the distribution of features present in the training data, has been identified as a major obstacle in the deployment of artificial intelligence systems in clinical environments. This issue has been identified as a crucial challenge in implementing artificial intelligence systems in clinical settings, as highlighted in Kelly et al. (2019). Additionally, this weakness arising from domain shift can result in unintended consequences such as gender or racial discrimination bias (Glocker et al., 2022). The utilisation of medical datasets for the training of deep learning models is often hindered by the lack of diversity, which is often derived from a single source with a specific population distribution. Domain shift can also occur due to differences in the image capture methods, such as differences in the intensity distribution of MRI scanners at different sites (Prados et al., 2017) or staining of histological slides (Stacke et al., 2020). These complexities pose significant challenges in delivering artificial intelligence systems in clinical settings.

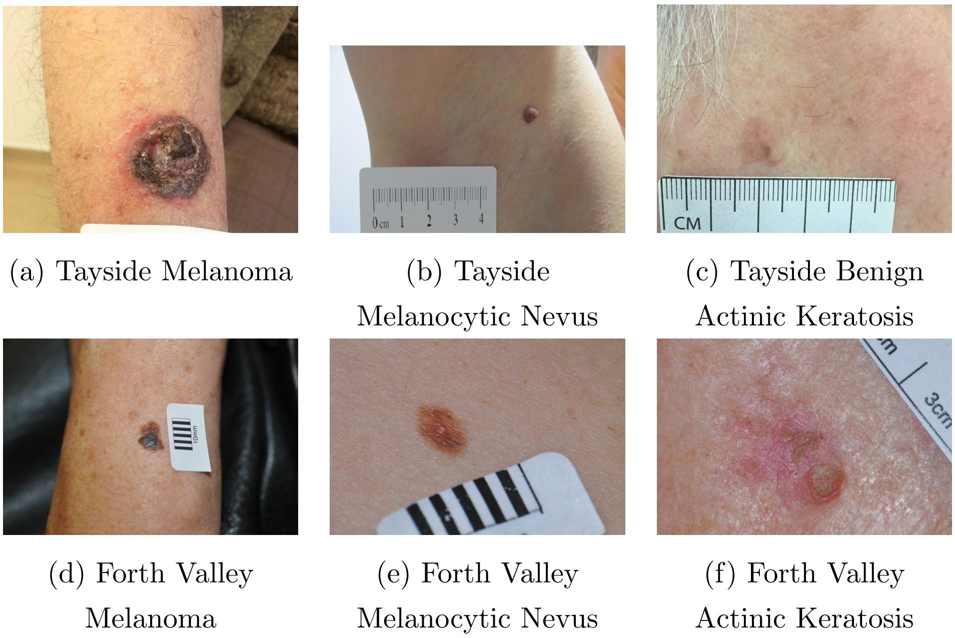
The creation and curation of labelled datasets can be a costly and time-consuming task, as highlighted in recent studies such as Chin et al. (2022). To address this issue, transfer learning (Weiss et al., 2016) has emerged as a valuable approach for utilising knowledge acquired from large datasets in other domains and applying it to a target domain with limited data. This method has been widely adopted in the field of medical image analysis, as it allows for the fine-tuning of features learned from large datasets for use in smaller datasets within the medical domain. The effectiveness of transfer learning, however, is dependent on the similarity between the source and target domains, as well as the deep learning models utilised, as noted in studies such as Matsoukas et al. (2022). For instance, the utilisation of knowledge acquired from the large ImageNet dataset (Deng et al., 2009) has been extensively applied in medical image analysis, despite the visual dissimilarity between the images in ImageNet and medical images. This approach has been demonstrated to be effective in the deep classification of the ISIC 2019 dermoscopy dataset, with evidence of feature re-use (Matsoukas et al., 2022).

In this chapter, the effectiveness of transfer learning between dermatology datasets is investigated with a focus on the utilisation of deep learning models to assist in the diagnosis of skin lesions based on community images acquired with limited control. The effectiveness of transfer learning is expected to be dependent on the size of the source datasets used for pre-training, as well as the similarity between these source datasets and the target datasets.

To explore this topic, two types of deep learning models with different inductive biases are focused on. The performance of these models is evaluated in relation to two novel datasets, specifically designed to simulate a real-world clinical setting. These datasets were gathered, trained, and tested in the context of image referrals sent to secondary-care hospital-based dermatologists from primary care. This use case was selected as it represents a common scenario in dermatology in the UK and aims to reliably identify common benign conditions in this setting.

The results from the experiments with fine-tuning between the different dataset shows that's performance is based on the proximity of the training distribution to the testing distribution. Therefore, it is critical to obtain data from the specific real-world setting in which the deep learning will be deployed. By training, and testing on two datasets, one from primary care and the other from referred images sent for medical photography, aims to demonstrate that the approach is effective in a real-world clinical setting and can be used for triage experiments.

The work discussed in this chapter is yet to be published but is intended for submission to the British Journal of Dermatology, under the authorship of Jacob Carse, Gillian Chin, Charlotte Proby, Emanuele Trucco, Colin Fleming, and Stephen McKenna. The personal contribution to this collaborative effort is reflected in the experiments carried out on the fine-tuning of the dataset.



(a) Tayside Melanoma

(b) Tayside Melanocytic Nevus

(c) Tayside Benign Actinic Keratosis

(d) Forth Valley Melanoma

(e) Forth Valley Melanocytic Nevus

(f) Forth Valley Actinic Keratosis

Figure 6.1: Example images from the Tayside and Forth Valley datasets.

6.2 Datasets

In this chapter, two datasets of community-acquired macroscopic (non-dermoscopic) images were curated. These datasets were extracted from previously stored images referred from primary to secondary care in Tayside and Forth Valley, United Kingdom. The inclusion of these datasets allows for the examination of a diverse range of macroscopic images acquired in a community setting. Both sets of images were annotated using the same procedure, explained in Section 6.2.3.

6.2.1 Tayside Dataset

The Tayside dataset (collected by Professor Colin Fleming), sourced from NHS Tayside [1], is centred on community-acquired skin lesion data, that represent real-world data capture from primary care. The images were obtained by primary care practitioners utilising a diverse array of cameras, from various cutaneous anatomical sites, utilising non-standardised lighting, framing, focusing, and acquisition settings. The dataset is intended for triage experiments, with the goal of reliably identifying common benign conditions in a real-world clinical setting. This dataset not only represents real-world teledermatology images acquired by primary care practitioners, but it also serves as a proxy for international skin

[1] NHS Tayside: `nhstayside.scot.nhs.uk`

datasets, where high-quality images and dermoscopy may not be readily available.

6.2.2 Forth Valley Dataset

The Forth Valley dataset (collected by Dr. Colin Morton) was procured by medical photographers capturing image of patients skin lesions, who were referred by primary care practitioners for specialist assessment within NHS Forth Valley [2]. The medical photographers possess a higher degree of expertise in the acquisition of photos of skin lesions, despite potentially lacking specialised knowledge in the field of dermatology. The medical photographers employ standard equipment, such as cameras, lighting, and backgrounds, to ensure uniformity in image quality. Furthermore, a standardised pattern of image capture is utilised to ensure high-quality, wide-angle, and close-up macro images.

6.2.3 Annotation Procedure

The datasets underwent a comprehensive annotation process that adhered to strict ethical guidelines and protocols for ensuring the quality and clinical usefulness of the images. This annotation procedure was developed by the Dermatology department at NHS Tayside. This included the removal of duplicate images and those that were not clinically relevant, as well as the de-identification and cropping of images to highlight any abnormalities. The images were then assigned a diagnostic label using the British Association of Dermatologists diagnostic index [3], which corresponds to the International Classification of Diseases (version 11) [4]. The assignment of these labels was conducted by two consultant-level dermatologists, registered on the UK General Medical Council specialist register of dermatologists [5], and any discrepancies were resolved by a third consultant-level dermatologist. This decision was based on all available clinical information, including pathology reports from a consultant pathologist. The labels represent real sources of data, with diagnoses for malignancy being derived from pathology, and for benign lesions, based

[2] NHS Forth Valley: `nhsforthvalley.com`

[3] BAD Clinical Guidelines: `https://www.bad.org.uk/guidelines-and-standards/clinical-guidelines/`

[4] ICD-11: `https://icd.who.int/dev11/l-derma/en`

[5] UK GMC Medical Register: `https://www.gmc-uk.org/registration-and-licensing/the-medical-register/`

on the opinions of consultant dermatologists. The full annotation procedure is presented in detail for further reference.

1. Ensure the necessary Caldicott/Ethical approvals are in place.

2. Proceed to open the first patient record.

3. Proceed to view the images. For each image, determine whether to retain it as potentially diagnostically useful, retain as a negative control, or discard it.

    (a) Retain as potentially diagnostically useful if any triage-level information can be obtained from the image.

    (b) Retain as negative control if a clinical image of skin without even triage level information i.e., where photographic information is insufficient to make a skin diagnosis. This may include images which are blurred by background details e.g., scarring or tattooing.

    (c) Discard the image if: a. No skin clinical images present, e.g., a clinical letter with no images may be in your system, X-ray image b. Duplicate image, i.e., multiple images, of which you will choose the best single view.

4. If retaining the image, anonymise where necessary. This may involve the removal of distinguishing features e.g., a tattoo, a label with patient details, or full-face views. Minimise cropping to ensure the remaining image has a maximum resolution.

5. Where multiple skin lesions, attempt to crop the image to ensure one lesion per image. Minimise cropping to ensure the remaining image has a maximum resolution.

6. For multiple skin lesions of the same disease process/widespread disease, if not possible to crop sufficiently, ensure all skin lesions have the same diagnostic label and are the same disease process.

7. Label the image with the diagnosis using BAD Diagnostic Index.

6.3 Fine-Tuning Experiments

This section details the datasets, training parameters, experimental setup, and results for the experiments with dermatology cross dataset fine-tuning. The code and full results used within this section can be found on the project GitHub repository [6].

6.3.1 Datasets

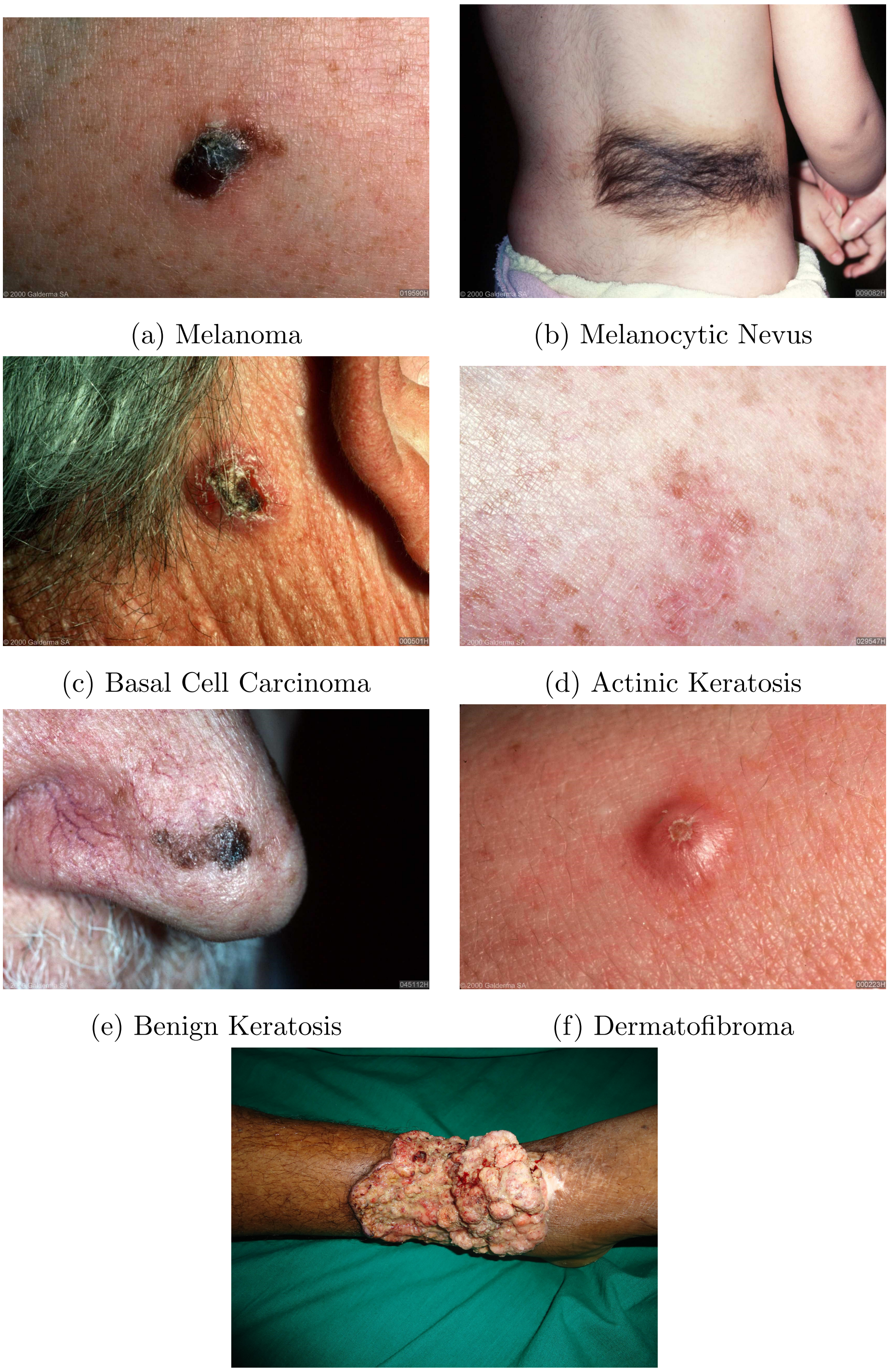
In this chapter's experiments, images from two public domain skin lesion datasets are utilised in addition to the datasets from Tayside and Forth Valley. These datasets were the ISIC 2019 dataset (Codella et al., 2018, Combalia et al., 2019, Tschandl et al., 2018) and the SD-260 dataset (Yang et al., 2019). The ISIC datasets are among the largest publicly available, with ISIC 2019 containing over 26,000 skin lesion images labelled with diagnoses. These images were acquired using dermatoscopes, which tend to produce well-centered, zoomed in, and consistent resolution images. In contrast, the SD-260 images were acquired in less controlled environments with varying imaging devices, resulting in more variation in colour, exposure, illumination, resolution, and scale (Figure 6.2). Visually, they are qualitatively similar to the Tayside and Forth Valley datasets but are acquired from a Chinese population.

To facilitate deep learning experiments, the four skin lesion datasets used in this chapter were intentionally restricted to a set of seven diagnostic categories. The specific seven categories were chosen based on their representation across the four datasets. For example, vascular lesions were excluded from the ISIC 2019 dataset, and classes such as angioma and solar lentigo were excluded from the Tayside dataset because they were not well represented in the other two datasets. The ISIC 2019, SD-260, Tayside, and Forth Valley data subsets used in the chapter contained 25,078, 13,814, 2,213, and 1,510 images, respectively (Table 6.1).

6.3.2 Training Parameters

In this chapter, two methods for image classification using deep learning were employed. The first method was a CNN, specifically an EfficientNet architecture (Tan and Le, 2019) with a compound coefficient of 7 and an additional fully connected layer of 512 neurons preceding the output

[6] GitHub Repository: `github.com/UoD-CVIP/Lesion-Classifier`



(a) Melanoma

(b) Melanocytic Nevus

(c) Basal Cell Carcinoma

(d) Actinic Keratosis

(e) Benign Keratosis

(f) Dermatofibroma

(g) Squamous Cell Carcinoma

Figure 6.2: Example images from the SD-260 dataset (Yang et al., 2019).

Table 6.1: Number of images per diagnosis in each dataset.

		ISIC 2019	SD-260	Tayside	Forth Valley	***Total***
Benign	Actinic Keratosis (B52)	867	1,434	414	143	*2,858*
	Dermatofibroma (X9002)	239	303	56	77	*675*
	Naevus, Melanocytic (X31z)	12,875	1,401	575	530	*15,381*
	Seborrhoeic Keratosis (X01)	2,624	1,133	537	289	*4,583*
Malignant	Melanoma (X41) / Melanoma in situ (X40)	4,522	7,094	78	204	*11,898*
	Squamous Cell Carcinoma (X12) / Squamous Cell Carcinoma in situ (X11)	628	17	175	89	*909*
	Basal Cell Carcinoma (X20)	3,323	2,432	378	178	*6,311*
Total		*25,078*	*13,814*	*2,213*	*1,510*	42,615

layer. The second method was a SWIN-B transformer (Liu et al., 2021), a state-of-the-art visual transformer network for image classification.

Subsequent training sessions were conducted for 40 epochs, with the model being saved each time the lowest validation loss was achieved. The weights were optimised using stochastic gradient descent with batches of 16 images and a triangular[2] cyclical scheduler (Smith, 2017), which alternated the learning rate between 10^{-5} and 10^{-2} and the momentum between 0.8 and 0.9. All images were pre-processed by cropping, resizing to 224 x 224 pixels, and normalising the pixel values between 0.0 and 1.0. To improve fine-tuning generalisation, data augmentation was used, specifically a variety of geometric and photometric changes applied at random when sampling. Horizontal and vertical flips, cropping and padding, affine transforms, Gaussian, average, and median blurring, sharpening, adding to channels, and multiplying channels by arbitrary amounts specifically.

Given an image, the deep network models predicted class probabilities for each of the seven diagnostic classes. These probabilities were constrained, by definition, to sum to one. Three of the seven classes represented malignant lesions. By summing the probabilities of these three classes, the probability that the observed lesion is malignant, assuming the class probabilities computed were well-calibrated (Chapter 4). This was used to evaluate the ability of the EfficientNet CNN and SWIN transformer models to identify malignant lesions. Specifically, ROC curves were used to quantify the sensitivity-specificity trade-offs that can be obtained on the macroscopic image datasets.

6.3.3 Experiment Setup

The ability of the models to classify lesion images from each dataset was first evaluated after training on images from that same dataset. When utilising the larger datasets, ISIC 2019 and SD-260, a disjoint split of 60% for training, 20% for validation, and 20% for testing was employed. These splits were static, and all training and testing with the datasets utilised the same splits. In contrast, given the limited size of the Tayside and Forth Valley datasets, 10-fold cross-validation was utilised to estimate performance, with each fold containing disjoint training (70%), validation (20%), and test (10%) sets. These splits were also static across each fold, with the performance being estimated by averaging the 10 test results.

Subsequently, cross-dataset performance was evaluated by measuring the class-balanced accuracy of each model when tested on data from datasets not used for training the model. Deep classifiers pre-trained with ImageNet (Deng et al., 2009) were trained on data from each of the datasets and then tested on each of the datasets. The composition of the training sets was identical to those used in the internal data classification experiment.

Finally, the effect of transfer learning between the dermatology datasets using an EfficientNet CNN and SWIM transformer models was evaluated. It is noted that all models in the experiments were pre-trained on ImageNet. Further pre-training was then performed on dermatology datasets. The effect of transfer between the large ISIC 2019 and SD-260 datasets was evaluated first. Secondly, the effect of transfer when the target domains (test data) were Tayside and Forth Valley data was investigated, utilising multi-dataset pre-training sequences, such as pre-training on SD-260 data and then training on Forth Valley training data, denoted "SD-260, Forth Valley" or pre-training on ISIC 2019 data, then on SD-260 data, and finally on Tayside training data, denoted "ISIC 2019, SD-260, Tayside".

Bootstrapping is used to evaluate model performance (100 bootstraps were sampled); mean class accuracy is obtained for each bootstrap, and the average mean class accuracy is presented together with the 95% confidence and 95% tolerance intervals (intervals within which 95% of bootstrapped test results lie). To describe the sensitivity-specificity trade-offs obtained when categorising test pictures as benign or malignant, ROC curves were constructed.

6.3.4 Results

Table 6.2 presents the balanced class accuracies obtained using an EfficientNet CNN and SWIN transformer models when trained and tested on data from the four datasets. The diagonal entries, in bold, reflect the internal accuracies obtained when disjoint test and training sets from the same source dataset were used. The results of models trained using the comparably smaller NHS datasets and then tested against the larger datasets were excluded from these experiments. This omission is due to the result not being necessary for the primary objective of this work, which is to examine the results when the models are tested against the moderately sized NHS datasets.

Table 6.2: Mean class accuracy test results for EfficientNet CNN and SWIN classifiers trained on each of the four datasets. Bootstrapped 95% confidence intervals are in parentheses.

		Testing Dataset			
Model	Training Dataset	ISIC 2019	SD-260	Tayside	Forth Valley
CNN	ISIC 2019	**0.975 (±0.0002)**	0.810 (±0.0007)	0.823 (±0.0006)	0.850 (±0.0007)
	SD-260	0.875 (±0.0004)	**0.957 (±0.0005)**	0.852 (±0.0006)	0.871 (±0.0006)
	Tayside			**0.881 (±0.0006)**	0.857 (±0.0002)
	Forth Valley			0.846 (±0.0002)	**0.891 (±0.0007)**
SWIN	ISCI 2019	**0.978 (±0.0002)**	0.788 (±0.0006)	0.812 (±0.0005)	0.853 (±0.0008)
	SD-260	0.880 (±0.0004)	**0.971 (±0.0005)**	0.859 (±0.0006)	0.870 (±0.0007)
	Tayside			**0.876 (±0.0007)**	0.864 (±0.0002)
	Forth Valley			0.850 (±0.0002)	**0.898 (±0.0007)**

Overall, in the absence of training data from the target domains, training on SD-260 data provided the best test accuracies on all test datasets. Models trained on ISIC data did not generalise well to the other datasets, while models trained on SD-260 data generalised slightly better but still poorly. Fine-tuning generalisation between the Tayside and Forth Valley datasets was better, with relatively small drops in test accuracy. Models trained on Tayside data achieved test results on Forth Valley data only 2.4% and 1.2% lower than test results on Tayside data.

Additionally, it was observed that models trained and tested on ISIC data did not benefit from pre-training on SD-260; ISIC test results with SD-260 pre-training were 98.5% and 98.5% for EfficientNet CNN and SWIN, respectively. Similarly, models trained and tested on SD-260 data did not benefit from pre-training on ISIC data. SD-260 test results with ISIC pre-training were 97.2% and 96.2% for EfficientNet CNN and SWIN, respectively.

This occurrence may be linked to the appearance of negative trans-

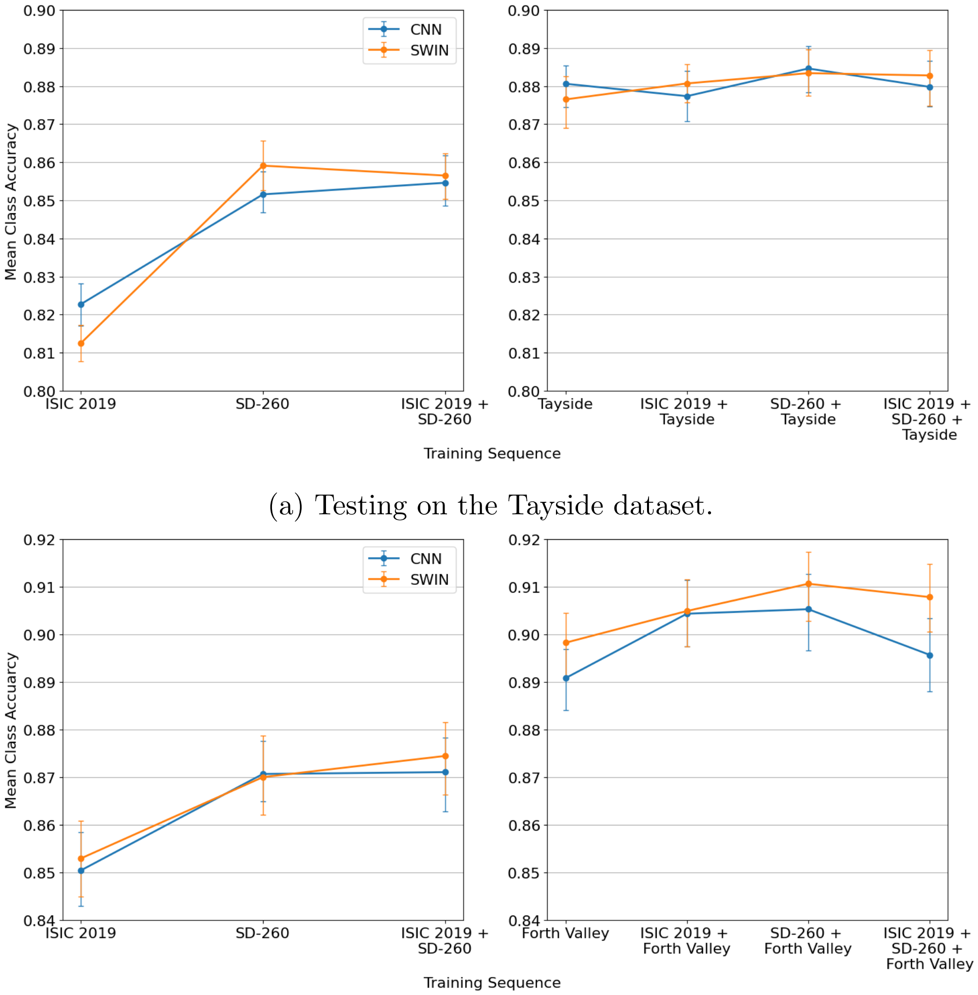
Table 6.3: Mean class accuracy for EfficientNet CNN and SWIN classifiers tested on Tayside and Forth Valley data having been trained with various multi-dataset training sequences. 95% confidence intervals are in parentheses.

		Testing Datasets	
Model	Training Datasets	Tayside	Forth Valley
CNN	SD-260, ISIC 2019	0.827 (±0.0007)	0.847 (±0.0007)
	ISIC 2019, SD-260	0.855 (±0.0006)	0.871 (±0.0008)
	ISIC 2019, Tayside	0.877 (±0.0006)	0.875 (±0.0002)
	SD-260, Tayside	**0.885 (±0.0006)**	0.877 (±0.0003)
	ISIC 2019, SD-260, Tayside	0.880 (±0.0006)	0.869 (±0.0002)
	ISIC 2019, Forth Valley	0.851 (±0.0002)	0.904 (±0.0007)
	SD-260, Forth Valley	0.861 (±0.0002)	**0.905 (±0.0007)**
	ISIC 2019, SD-260, Forth Valley	0.850 (±0.0002)	0.896 (±0.0007)
SWIN	SD-260, ISIC 2019	0.833 (±0.0005)	0.870 (±0.0007)
	ISIC 2019, SD-260	0.856 (±0.0006)	0.874 (±0.0008)
	ISIC 2019, Tayside	0.881 (±0.0006)	0.878 (±0.0002)
	SD-260, Tayside	**0.885 (±0.0006)**	0.882 (±0.0002)
	ISIC 2019, SD-260, Tayside	0.883 (±0.0007)	0.888 (±0.0007)
	ISIC 2019, Forth Valley	0.856 (±0.0002)	0.905 (±0.0007)
	SD-260, Forth Valley	0.862 (±0.0002)	**0.911 (±0.0007)**
	ISIC 2019, SD-260, Forth Valley	0.857 (±0.0002)	0.908 (±0.0008)

fer, in which traits gained during the initial ISIC pre-training phase may have deleterious consequences when applied to the subsequent SD-260 training regimen. As indicated by the existing literature (Wang et al., 2019), negative transfer can be caused by a variety of causes, including inconsistencies in feature alignment, discordant patterning, instances of overfitting, and, most significantly, domain shift. In this specific context, the perceptible disparity in the relevance of knowledge gained from the ISIC domain versus its application to the SD-260 training domain may underpin the emergence of optimisation challenges, potentially culminating in entrapment within local minima configurations.

Table 6.3 reports balanced class accuracies when EfficientNet CNN and SWIN transformers were tested on Tayside and Forth Valley data after various multi-dataset training sequences. In each case, pre-training on SD-260 data followed by fine-tuning on data from the target domain was found to be effective.

The results are also illustrated in Figure 6.3 which shows how test accuracies changed when different datasets were used for training and transfer learning. It was observed that training only on the large ISIC



(a) Testing on the Tayside dataset.

(b) Testing on the Forth Valley dataset.

Figure 6.3: Mean class accuracy results for EfficientNet CNN (blue) and SWIN (orange) benign-malignant classifiers tested on (a) Tayside and (b) Forth Valley data. Bars indicate 95% tolerance intervals computed using bootstrap. The plots on the left are from classifiers trained only on public domain datasets. The plots on the right are cross-validation results from classifiers pre-trained using public domain datasets and then fine-tuned using data from the target domain, i.e., Tayside and Forth Valley, respectively.

and SD-260 datasets gave relatively poor results, while training on a small dataset from the target domain, after pre-training only on ImageNet, performed better. The most accurate models were obtained by pre-training on the large dermatology datasets followed by further training on data from the target domain.

Figure 6.4 illustrates the extent to which models trained for the Tayside domain were able to generalise to the Forth Valley domain, and vice-versa. The solid curves plot the accuracies obtained when training on data from the test domain, with and without pre-training on the large public domain datasets. Dashed curves plot accuracies when training on data from the other test domain. The drops in accuracy when generalising between Tayside and Forth Valley dataset were 2-3% using the SWIN transformer model.

The 95% confidence intervals shown in Tables 6.2 and 6.3 are always small, always encompassing a range less than $\pm, 0.001$. This result indicates a higher level of statistical dependability in the estimated values of the parameters under consideration. It specifically suggests that the models under consideration have a remarkable amount of robustness, and the predictions they provide have a respectable level of resilience. In contrast, the visual representations shown in Figures 6.3, 6.4, and 6.5 use 95% tolerance intervals rather than confidence intervals since the latter were regarded too small. These tolerance intervals represent the upper and lower boundaries of the generated bootstrap samples' central 95% distribution.

Finally, binary classification performance in the form of ROC curves is reported in Figure 6.5. These curves were generated using models pre-trained on SD-260 prior to training on either Tayside or Forth Valley data. It was observed that the Forth Valley curves dominated the Tayside curves, and curves for models trained on the test domain dominated curves trained on the other domain. This occurrence is arguably due to the noticeably higher quality of the Forth Valley photos, as seen in Figure 6.1, primarily in terms of properties such as lighting, spatial alignment, and proportional scale. As a result, the classification model can have a concentrated focus on the knowledge of the key lesion characteristics avoiding involvement with these variables.

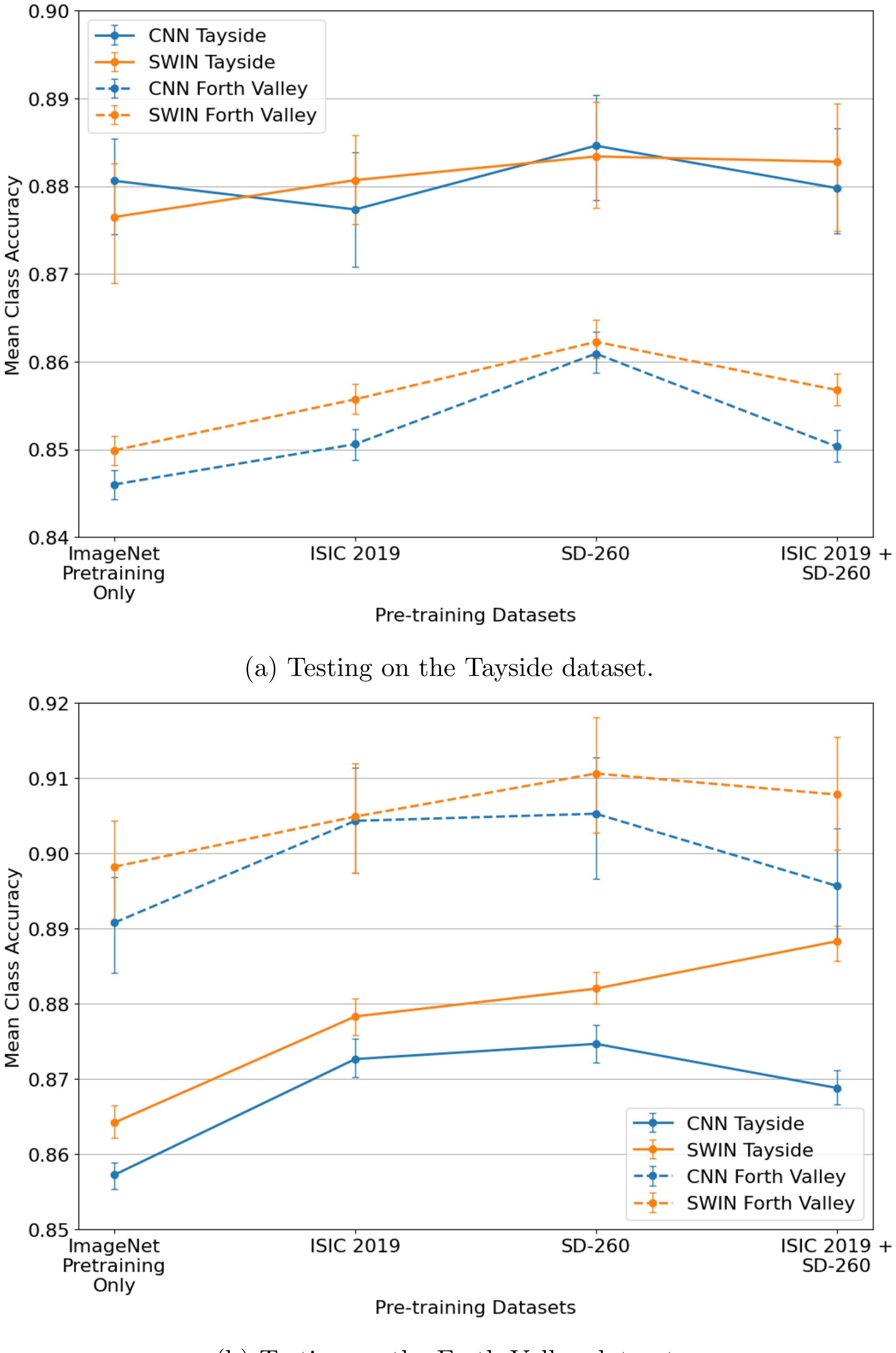
6.4 Conclusion

It is well-established that deep learning classifiers benefit from large training sets. However, it is also known that test performance is negatively affected by variations in image acquisition, image quality, and the population being imaged. This presents challenges in the development of dermatology diagnostic systems that can be deployed in multiple sites, as such conditions can vary geographically and over time. Curation of large datasets can be costly at a local level, yet learning systems need to be attuned to local conditions.

The results suggest that the SWIN transformer should be preferred over the EfficientNet convolutional neural network. The transformer obtained accuracies of 98% and 97% on the ISIC 2019 and SD-260 test datasets, provided it was trained on data from the same domain. However, cross-domain fine-tuning (training on ISIC 2019 and testing on SD-260) led to a weaker 79% accuracy. Training on SD-260 and testing on ISIC 2019 yielded a better 88% accuracy, possibly due to the more varied nature of the SD-260 dataset (Table 6.2).

The ISIC 2019 and SD-260 datasets are relatively large, though still not comparable to datasets used for deep learning in computer vision. The focus of this chapter was how to obtain good performance on challenging local data with limited availability of diagnostic labels. The two NHS datasets used to explore this question were sourced from Tayside and Forth Valley, the latter having images of more consistent and higher quality than the former. Transformers yielded accuracies of 88% and 90%, respectively, when trained and tested on data from these domains. This was better than results obtained by training on the larger SD-260 and ISIC 2019 datasets (Figure 6.3), highlighting the benefit of training with data from the local target domain. It was found that by pre-training on SD-260 and then on data from the local target domain, further increases to 89% and 91% were obtained (Figure 6.4, solid blue lines).

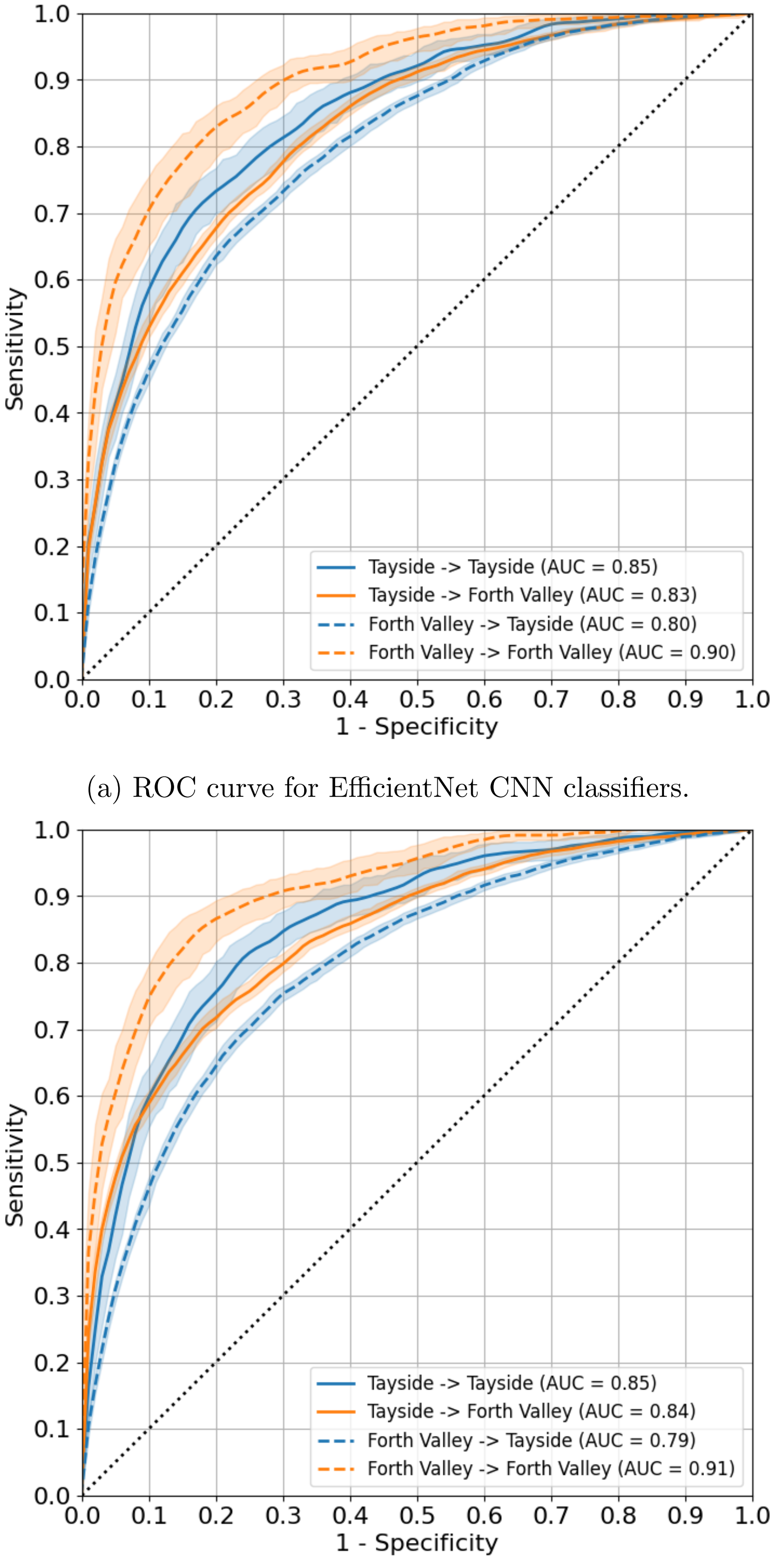
Given their geographical proximity, Tayside and Forth Valley data are from similar populations, so one might expect good cross-domain fine-tuning generalisation between them. However, training on data from one (with appropriate pre-training for transfer) and testing on the other gave accuracies 2-3% lower than testing on the same domain. This is likely due to differences in image acquisition. ROC curves were used to indicate sensitivity-specificity trade-offs (Figure 6.5). In a triage setting, relatively low specificity can be acceptable in return for very high sensitivity, especially for melanoma. A combination of pre-training on public macroscopic data, followed by tuning to local data, gave promising results. However, further improvements are needed for deployment in a real clinical pathway.



(a) Testing on the Tayside dataset.

(b) Testing on the Forth Valley dataset.

Figure 6.4: Mean class accuracy results for EfficientNet CNN (blue) and SWIN (orange) benign-malignant classifiers tested on (a) Tayside and (b) Forth Valley data. Bars indicate 95% tolerance intervals computed using bootstrap. The x-axis indicates the datasets used for pretraining and the image legend indicates the fine-tuning dataset. All classifiers were first pre-trained with ImageNet. Solid lines: fine-tuned on Tayside; Dashed lines: fine-tuned on Forth Valley.



(a) ROC curve for EfficientNet CNN classifiers.

(b) ROC curve for SWIN transformer classifiers.

Figure 6.5: ROC curves for (a) CNN and (b) SWIN classifiers tested on Tayside (blue) and Forth Valley (orange) datasets. All curves were generated using classifiers pre-trained on SD-260 prior to fine-tuning on either Tayside data (solid lines) or Forth Valley data (dashed lines). Shaded areas indicate 95% tolerance intervals computed using bootstrap.

Chapter 7

Conclusions and Recommendations

7.1 Summary of Contributions

This thesis represents a contribution to the field of medical image analysis and machine learning, pursuing two primary aims. Firstly, it aims to address the challenge of limited annotated data, particularly for histopathological whole slide images, by utilising active learning and unsupervised learning techniques. Secondly, it seeks to enhance the accuracy of asymmetrical selective classification for skin lesion images. Additionally, the thesis presents secondary contributions in the areas of predictive probability calibration and dataset fine-tuning generalisation.

7.1.1 Scarcity of Annotations

How can a deep learning model be effectively trained to achieve optimal performance when faced with a scarcity of annotations, and a large corpus of unannotated data?

To achieve this aim, the first choice of method investigated was active learning, as it is a type of machine learning that seeks to select the most beneficial unannotated data for the model to annotate (Settles, 2009). However, this led to the identification of limitations of active learning and its application to histopathology patches, where nuclei patches may be difficult to annotate. Chapter 2 introduced an active-learning framework designed to select tiles composed of multiple patches for annotation with a view to making the annotation task easier and thus increasing annotation throughput (Carse and McKenna, 2019). The efficacy of this framework was evaluated using various query strategies on nuclei classification tasks by employing CNNs trained on small patches containing single nuclei.

The results suggest that traditional active-learning approaches are less effective when applied to deep-learning models, while specialised active-learning techniques for deep-learning fail to outperform random sampling baselines. This phenomenon has been previously noted in literature on active deep-learning (Ren et al., 2021) and underscores the need for more robust active-learning methods in this domain. Although this chapter demonstrates that active learning holds promise as a means to address these challenges, further research is required to achieve significant improvements on tasks such as those presented herein. This motivated an investigation into unsupervised learning techniques as a complementary approach.

As discussed in Chapter 2, active learning alone may not provide representative enough features for a deep learning model to learn from the annotated data. This leads to the investigation of unsupervised representative learning techniques, which can enable a model to learn generalisable representative features from the unannotated data, which can then be fine-tuned. Chapter 3 proposed a modification to the CPC framework (van den Oord et al., 2018) for digital pathology patch classification. The modification involved using an alternative infilling-style mask to construct the latent context and a multi-directional PixelCNN autoregressor (van den Oord et al., 2016).

The experiments conducted to evaluate the proposed modification to the CPC framework revealed that the original implementation of CPC is not well-suited for patch-based digital pathology tasks. However, the proposed multi-directional modifications to the CPC led to better results and improved classification accuracies on transfer learning tasks, where access to annotated data is limited (Carse, Carey and McKenna, 2021). Thus, the combination of active learning and unsupervised representation learning holds promise in digital pathology tasks.

While not inherently tied to immediate clinical use, the concurrent application of these approaches has the potential to aid engineers and researchers in their efforts to train machine learning models specialised for the processing of medical images. This technique could result in cost savings and the elimination of time-consuming annotation duties for doctors, whose engagement in data annotation would be reduced.

7.1.2 Asymmetrical Selective Classification

To what extent can selective classification techniques be applied in order to mitigate the costs associated with asymmetrical misdiagnosis of skin lesion images?

Chapter 5 presents an investigation into the efficacy of selective classification as a potential solution to asymmetrical misdiagnosis in skin lesion images. The research aims to address this issue through the exploration of cost-sensitive classification techniques in both binary triage and multi-class disease classification scenarios using a dermatology dataset. The chapter draws on the expertise of clinical dermatologists to provide asymmetrical misclassification costs based on healthcare economic estimations. Methods for uncertainty estimation with neural networks and probability calibration were evaluated, and a novel modification to SelectiveNet (Geifman and El-Yaniv, 2019), known as EC-SelectiveNet (Carse, Süveges, Hogg, Trucco, Proby, Fleming and McKenna, 2021), was proposed.

The results suggest that SelectiveNet exhibited inferior performance compared to other selective classification methods, except for when it was trained with a target coverage of 1.0. In contrast, EC-SelectiveNet, trained with a target coverage of 1.0, consistently outperformed all other methods in both binary and multi-class settings in the presence of asymmetrical costs. The utilisation of Bayesian neural networks had minimal effect on the predictions when averaged in any setting. Interestingly, the use of various uncertainty measures resulted in different outcomes, with the predictive entropy measure surpassing all others, particularly in an asymmetric setting, while the variational ratios performed poorly in both symmetric and asymmetric cost settings. The data also revealed that using temperature scaling to calibrate predictions to optimise discerning classification resulted in increased costs, particularly at higher coverage thresholds in the context of imbalanced cost scenarios. While this strategy may be useful in situations requiring limited coverage, its benefits may not outweigh the negative consequences identified in situations requiring heightened coverage. The chapter evaluated diverse selective classification settings and underscores the need for further research to advance selective classification methods and comprehend their performance in asymmetrical cost settings.

These efforts are critical in the context of adopting classification approaches in clinical settings, which are characterised by pervasive asymmetrical costs and an inherent limitation on the feasibility of classifying all images. This situation emphasises the importance of incorporating rejection mechanisms. The combination of these techniques with established clinical workflows has the potential to yield exclusive predictions only for data instances that can be accurately classified by the model. Even if this threshold is only 20%, the resulting reduction in clinician burden corresponds to a significant one-fifth reduction.

7.1.3 Secondary Contributions

Predictive Probability Calibration

Calibration denotes the systematic procedure of conforming the anticipated probabilities of a model with the authentic probabilities of the target variable (Guo et al., 2017). Poor calibration performance in modern deep neural networks can hinder the calculation of a model's uncertainty, thereby affecting uncertainty-dependent techniques like active learning and cost-sensitive decision-making (Carse, Süveges, Hogg, Trucco, Proby, Fleming and McKenna, 2021). To address this, Chapter 4 presented an empirical investigation of calibration techniques on two medical image classification tasks: multi-class dermatology classification and binary histopathology image classification. The chapter implemented temperature scaling, optimising the temperature parameter using various calibration measures rather than the standard negative log-likelihood. This method was applied to networks trained with one-hot encoding and cross-entropy loss, as well as networks trained with focal loss and label smoothing. Two Bayesian neural network approaches were also utilised for comparison. The results demonstrated that while alternative calibration metrics may not provide significant advantages for tuning temperature, temperature scaling of networks trained with focal loss and appropriate hyperparameters exhibited robust performance in terms of both calibration and accuracy across both datasets (Carse et al., 2022). The calibration techniques discussed in this thesis are directly applicable to clinical applications, as they provide a means to improve clinicians' understanding of prediction uncertainties and the level of trust that they can appropriately ascribe to recommendations derived from neural network analyses.

Dataset Fine-Tuning

In chapters 2, 3, 4 and 5, open-source datasets were utilised to train and evaluate deep learning models. However, it is important to note that although large open-source datasets are useful for experimentation, the models produced may not be suitable for clinical use (Wu et al., 2022). To address this concern, Chapter 6 conducts an investigation to determine the generalisability of models trained with open-source data when fine-tuned to locally collected macroscopic datasets from primary care referrals. Two types of neural networks, a CNN and a transformer, were employed to evaluate the model's generalisation performance on two open-source datasets and two smaller locally collected datasets from the NHS. The findings emphasise the significance of assessing the fine-tuning of deep learning algorithms for macroscopic skin lesion images in real-world settings. Moreover, the chapter highlights the potential benefits of utilising large public macroscopic datasets for pre-training and fine-tuning the algorithms with smaller local datasets. This could be clinically useful as it means that models could be fine-tuned for individual populations instead of focusing on large general models expected to perform on a global distribution.

7.2 Limitations and Future Work

This section describes the limitations of this thesis as well as outlining potential avenues for further research and areas for improvement.

7.2.1 Annotator Efficient Active Learning

Chapter 2 presented an investigation into improving annotation throughput on deep active learning methods for histopathology patches, with the goal of expanding the volume of annotations gathered while minimising annotation costs. The proposed approach entails simplifying the annotation task to optimise annotator time allocation. Notably, there has been limited research on enhancing annotator efficiency for medical image analysis (Ren et al., 2021), and further investigation is warranted across diverse modalities. One possible avenue for exploration involves the application of established methods that address asymmetrical annotation costs, such as CEREALS (Mackowiak et al., 2018), to tasks like whole-slide segmentation, where the costs of obtaining annotations are typically uniform (Budd et al., 2021).

The experiments relied on simulated active learning scenarios in which annotations had already been collected and provided automatically upon query. This approach allowed for rapid development of active learning query strategies but did not enable investigation into how a human in the loop would interact with the active learning scenario. Another limitation of the experiments was the dataset used, which was an open-source dataset that had been pre-filtered by selecting the best examples and removing any challenging examples or anomalies that could arise in real-world scenarios where an active learning query system would need to be employed. Therefore, these experiments may not fully capture the complexities and challenges of active learning in real-world settings. Both of these can be mitigated in future by using real unannotated datasets conducting a case study by having clinicians annotate sets of queried data from different query strategies.

The evidence in Chapter 2 shows that current active learning query strategies don't perform much better than random querying. This suggests that there is room for improvement in developing new strategies. One promising research direction would be to explore the integration of batch-aware active learning and semi-supervised learning. This could be achieved by designing an active learning query strategy that places emphasis on improving the semi-supervised learning performance. One such attempt has been made for a scoring query strategy like CEAL (Wang et al., 2016), which merges softmax response active learning with the semi-supervised method of pseudo labelling. Despite its rudimentary nature, this approach demonstrated encouraging results and is open to further improvement.

7.2.2 Unsupervised Representation Learning

Chapter 3 delved into the topic of unsupervised representation learning, an area that has witnessed significant progress and continues to be a subject of active development. Although much of the research in this field has been centred around image datasets such as ImageNet (Deng et al., 2009), it is important to note that the developed methods may not be optimally suited for medical image datasets. An alternative approach involves disentanglement methods that aim to learn a model capable of identifying and disentangling the underlying factors in the observable data. A recent survey paper by Liu et al. (2022) provides an overview of such methods. In the context of histopathological patches, disentanglement can facilitate the production of feature representations that capture variations in slide staining, among other factors. These features, when used in conjunction with transfer learning, can yield improved accuracy and generalisation performance on new data.

7.2.3 Predictive Probability Calibration

The experimental findings presented in Chapter 4 indicate that temperature scaling (Guo et al., 2017) optimised with any measure of calibration can effectively enhance calibration for multi-class classification tasks. As a post-hoc calibration method, temperature scaling is readily applicable to pre-trained models and does not interfere with the model training process. These results suggest that future research should prioritise the further development and refinement of post-hoc calibration methods. While temperature scaling involves the learning of a single scaling value for model calibration, recent work investigates alternative post-hoc calibration methods that warrant further exploration (Song et al., 2021). In order to assess the performance of these methods, appropriate evaluation metrics must be developed and investigated.

The evaluation of calibration is a crucial aspect in the development of classification systems. Despite its significance, the determination of accurate calibration measures remains a subject of active research, although efforts have been made to incorporate them into various medical image analysis studies (Maier-Hein et al., 2022). In Chapter 4, the KDE-ECE (Zhang et al., 2020) approach was employed to quantify the calibration of the trained models. This measure evaluates the overall calibration performance across the confidence and correctness distribution. It is noteworthy that KDE-ECE is just one of several calibration measures and should be used in conjunction with other approaches, such as the maximum calibration error, which indicates the maximum calibration error of a bin rather than the weighted average of errors. Furthermore, there is scope for exploring alternative calibration measures, such as those proposed by Nixon et al. (2019).

7.2.4 Asymmetrical Selective Classification

In Chapter 5, experiments are presented on selective classification of skin lesion images with both symmetrical and asymmetrical misdiagnosis costs. The asymmetrical costs utilised in this chapter were derived from rough estimates provided by a consultant-level dermatologist. However, these costs are subject to variation based on local healthcare economics and are likely to change over time. Consequently, the development of additional asymmetrical costs for realistic settings is necessary for the evaluation of future algorithms in a practical environment.

The findings presented in Chapter 5 suggest that both Bayesian and SelectiveNet (Geifman and El-Yaniv, 2019) models encountered challenges in symmetrical and asymmetrical environments during selective classification experiments. The most effective model was found to be the standard CNN model utilising expected costs, which exhibited good calibration and performance. Temperature scaling (Guo et al., 2017), the only calibration method evaluated, did not meet expectations. However, optimising decision calibration (Zhao et al., 2021) of the models could potentially enhance selective classification performance in asymmetrical misclassification cost environments. This could involve extending asymmetrical decision calibration to facilitate selective classification of skin lesions by incorporating the option to reject an image due to excessive expected loss.

7.2.5 Dataset Fine-Tuning

In Chapter 6, the generalisability of models trained with skin lesion datasets by fine-tuning to other datasets is investigated. Based on the findings of this investigation, it is concluded that future research efforts should prioritise the exploration of techniques that can effectively adapt cross-domain models. This should be done by considering the varying costs associated with misdiagnoses of different types, with a particular focus on making cost-sensitive classification decisions (Guan et al., 2021, Carse, Süveges, Hogg, Trucco, Proby, Fleming and McKenna, 2021). To improve the accuracy of classification decisions, incorporating well-calibrated classifiers is recommended. This approach would enable the implementation of selective classification decisions. Furthermore, to enhance the performance of deep learning algorithms, it is essential to acquire and annotate data in a prospective manner during dermatology consultant triaging and clinical work. This would ultimately lead to the creation of larger local datasets that would be beneficial in improving the performance of the algorithms. However, it is important to acknowledge that the findings of this chapter, similar to other studies, are limited by the restricted number of critical diagnostic categories that are examined. Although including more diagnostic categories comes with its own issues known as the long tail problem where the more diseases are included, they are less represented in the training data (Roy et al., 2022).

References

Akrami, H., Joshi, A. A., Li, J., Aydore, S. and Leahy, R. M. (2020), Brain lesion detection using a robust variational autoencoder and transfer learning, *in* 'International Symposium on Biomedical Imaging (ISBI)', IEEE, pp. 786–790.

Arjovsky, M., Chintala, S. and Bottou, L. (2017), Wasserstein generative adversarial networks, *in* 'International Conference on Machine Learning (ICML)', PMLR, pp. 214–223.

Bachman, P., Hjelm, R. D. and Buchwalter, W. (2019), 'Learning representations by maximizing mutual information across views', *International Conference on Neural Information Processing Systems (NIPS)* pp. 15535–15545.

Bayramoglu, N. and Heikkilä, J. (2016), Transfer learning for cell nuclei classification in histopathology images, *in* 'Transferring and Adapting Source Knowledge in Computer Vision Workshop (Task-CV)', Springer.

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

Bengio, Y., Courville, A. and Vincent, P. (2013), 'Representation learning: A review and new perspectives', *Transactions on Pattern Analysis and Machine Intelligence* **35**(8), 1798–1828.

Bengio, Y., Yao, L., Alain, G. and Vincent, P. (2013), 'Generalized denoising auto-encoders as generative models', *International Conference on Neural Information Processing Systems (NIPS)* pp. 899–907.

Blundell, C., Cornebise, J., Kavukcuoglu, K. and Wierstra, D. (2015), Weight uncertainty in neural network, *in* ‘International Conference on Machine Learning (ICML)’, PMLR, pp. 1613–1622.

Botev, A., Ritter, H. and Barber, D. (2017), Practical Gauss-Newton optimisation for deep learning, *in* ‘International Conference on Machine Learning (ICML)’, PMLR, pp. 557–565.

Budd, S., Robinson, E. C. and Kainz, B. (2021), ‘A survey on active learning and human-in-the-loop deep learning for medical image analysis’, *Medical Image Analysis* **71**, 102062.

Caron, M., Bojanowski, P., Joulin, A. and Douze, M. (2018), Deep clustering for unsupervised learning of visual features, *in* ‘European conference on computer vision (ECCV)’, Springer, pp. 132–149.

Carse, J., Alvarez Olmo, A. and McKenna, S. (2022), Calibration of deep medical image classifiers: An empirical comparison using dermatology and histopathology datasets, *in* ‘Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE)’, Springer, pp. 89–99.

Carse, J., Carey, F. and McKenna, S. (2021), Unsupervised representation learning from pathology images with multi-directional contrastive predictive coding, *in* ‘International Symposium on Biomedical Imaging (ISBI)’, IEEE, pp. 1254–1258.

Carse, J. and McKenna, S. (2019), Active learning for patch-based digital pathology using convolutional neural networks to reduce annotation costs, *in* ‘European Congress on Digital Pathology (ECDP)’, Springer, pp. 20–27.

Carse, J., Süveges, T., Hogg, S., Trucco, E., Proby, C., Fleming, C. and McKenna, S. (2021), Robust selective classification of skin lesions with asymmetric costs, *in* ‘Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE)’, Springer, pp. 112–121.

Chang, H., Han, J., Zhong, C., Snijders, A. M. and Mao, J.-H. (2017), ‘Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications’, *Transactions on Pattern Analysis and Machine Intelligence* **40**(5), 1182–1194.

Chen, B.-C., Chen, C.-S. and Hsu, W. H. (2014), Cross-age reference coding for age-invariant face recognition and retrieval, *in* 'European Conference on Computer Vision (ECCV)', Springer, pp. 768–783.

Chin, G. X., Suveges, T., Carse, J., Butt, S., Muthiah, S., Morton, C., Trucco, E., Proby, C., McKenna, S. and Fleming, C. (2022), 'Prepare to succeed: real-world image datasets for artificial intelligence in skin cancer triage', *British Journal of Dermatology* **187**(Suppl. 1)), 125.

Chow, C.-K. (1957), 'An optimum character recognition system using decision functions', *Transactions on Electronic Computers* **4**, 247–254.

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H. et al. (2018), Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC), *in* 'International Symposium on Biomedical Imaging (ISBI)', IEEE, pp. 168–172.

Cohen, G., Afshar, S., Tapson, J. and Van Schaik, A. (2017), Emnist: Extending mnist to handwritten letters, *in* 'International Joint Conference on Neural Networks (IJCNN)', IEEE, pp. 2921–2926.

Cohn, D. A., Ghahramani, Z. and Jordan, M. I. (1996), 'Active learning with statistical models', *Journal of Artificial Intelligence Research* **4**, 129–145.

Colling, P., Roese-Koerner, L., Gottschalk, H. and Rottmann, M. (2020), 'Metabox+: A new region based active learning method for semantic segmentation using priority maps', *arXiv preprint arXiv:2010.01884* .

Combalia, M., Codella, N. C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A. C., Puig, S. et al. (2019), 'Bcn20000: Dermoscopic lesions in the wild', *arXiv preprint arXiv:1908.02288* .

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B. (2016), The cityscapes dataset for semantic urban scene understanding, *in* 'Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE/CVF, pp. 3213–3223.

Cortes, C., DeSalvo, G. and Mohri, M. (2016), Learning with rejection, *in* ‘Algorithmic Learning Theory (ALT)’, Springer, pp. 67–82.

Dai, Z., Low, B. K. H. and Jaillet, P. (2020), ‘Federated bayesian optimization via thompson sampling’, *International Conference on Neural Information Processing Systems (NIPS)* pp. 9687–9699.

Darlow, L. N., Crowley, E. J., Antoniou, A. and Storkey, A. J. (2018), ‘Cinic-10 is not imagenet or cifar-10’, *arXiv preprint arXiv:1810.03505* .

Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M. and Hennig, P. (2021), ‘Laplace redux-effortless Bayesian deep learning’, *International Conference on Neural Information Processing Systems (NIPS)* pp. 20089–20103.

Demir, B. and Bruzzone, L. (2014), ‘A novel active learning method in relevance feedback for content-based remote sensing image retrieval’, *Transactions on Geoscience and Remote Sensing* **53**(5), 2323–2334.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009), Imagenet: A large-scale hierarchical image database, *in* ‘Conference on Computer Vision and Pattern Recognition (CVPR)’, IEEE/CVF, pp. 248–255.

Dimitriou, N., Arandjelović, O. and Caie, P. D. (2019), ‘Deep learning for whole slide image analysis: an overview’, *Frontiers in Medicine* **6**, 264.

Donahue, J., Krähenbühl, P. and Darrell, T. (2016), ‘Adversarial feature learning’, *arXiv preprint arXiv:1605.09782* .

Donahue, J. and Simonyan, K. (2019), ‘Large scale adversarial representation learning’, *International Conference on Neural Information Processing Systems (NIPS)* pp. 10542–10552.

Du-Harpur, X., Watt, F., Luscombe, N. and Lynch, M. (2020), ‘What is ai? applications of artificial intelligence to dermatology’, *British Journal of Dermatology* **183**(3), 423–430.

Ducoffe, M. and Precioso, F. (2018), ‘Adversarial active learning for deep networks: a margin based approach’, *arXiv preprint arXiv:1802.09841* .

El-Yaniv, R. et al. (2010), 'On the foundations of noise-free selective classification.', *Journal of Machine Learning Research* **11**(5).

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

Farahani, R. Z. and Hekmatfar, M. (2009), *Facility location: concepts, models, algorithms and case studies*, Springer Science & Business Media.

Ferrer, L. (2022), 'Analysis and comparison of classification metrics', *arXiv preprint arXiv:2209.05355* .

Folmsbee, J., Brandwein-Weber, M. and Doyle, S. (2021), Whole slide semantic segmentation: large scale active learning for digital pathology, *in* 'Medical Imaging: Digital Pathology', Vol. 11603, SPIE, pp. 83–93.

Freeman, L. C. (1965), *Elementary applied statistics: for students in behavioral science*, New York: Wiley.

Frenkel, L. and Goldberger, J. (2021), Network calibration by class-based temperature scaling, *in* 'European Signal Processing Conference (ESPC)', IEEE, pp. 1486–1490.

Fujisawa, Y., Otomo, Y., Ogata, Y., Nakamura, Y., Fujita, R., Ishitsuka, Y., Watanabe, R., Okiyama, N., Ohara, K. and Fujimoto, M. (2019), 'Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis', *British Journal of Dermatology* **180**(2), 373–381.

Fumera, G. and Roli, F. (2002), Support vector machines with embedded reject option, *in* 'Pattern Recognition with Support Vector Machines', Springer, pp. 68–82.

Gal, Y. and Ghahramani, Z. (2016), Dropout as a bayesian approximation: Representing model uncertainty in deep learning, *in* 'International Conference on Machine Learning (ICML)', PMLR, pp. 1050–1059.

Gal, Y., Islam, R. and Ghahramani, Z. (2017), Deep bayesian active learning with image data, *in* 'International Conference on Machine Learning (ICML)', PMLR, pp. 1183–1192.

Gal, Y. et al. (2016), *Uncertainty in deep learning*, University of Cambridge.

Gamper, J., Alemi Koohbanani, N., Benet, K., Khuram, A. and Rajpoot, N. (2019), Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification, *in* 'European Congress of Digital Pathology (ECDP)', Springer, pp. 11–19.

Gamper, J., Koohbanani, N. A., Benes, K., Graham, S., Jahanifar, M., Khurram, S. A., Azam, A., Hewitt, K. and Rajpoot, N. (2020), 'Pannuke dataset extension, insights and baselines', *arXiv preprint arXiv:2003.10778* .

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R. et al. (2021), 'A survey of uncertainty in deep neural networks', *arXiv preprint arXiv:2107.03342* .

Geifman, Y. and El-Yaniv, R. (2017), 'Selective classification for deep neural networks', *International Conference on Neural Information Processing Systems (NIPS)* pp. 4885–4894.

Geifman, Y. and El-Yaniv, R. (2019), Selectivenet: A deep neural network with an integrated reject option, *in* 'International Conference on Machine Learning (ICML)', PMLR, pp. 2151–2159.

Gidaris, S., Singh, P. and Komodakis, N. (2018), Unsupervised representation learning by predicting image rotations, *in* 'International Conference on Learning Representations (ICLR)'.

Glocker, B., Jones, C., Bernhardt, M. and Winzeck, S. (2022), 'Risk of bias in chest x-ray foundation models', *arXiv preprint arXiv:2209.02965* .

Glorot, X. and Bengio, Y. (2010), Understanding the difficulty of training deep feedforward neural networks, *in* 'International Conference on Artificial Intelligence and Statistics (AIStat)', JMLR, pp. 249–256.

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

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014), Generative adversarial networks, *in* 'International Conference on Neural Information Processing Systems (NIPS)', Curran Associates Inc., pp. 2672–2680.

Gorriz, M., Nieto, X. G. I., Carlier, A. and Faure, E. (2017), Cost-effective active learning for melanoma segmentation, *in* 'Workshop on Machine Learning for Health (ML4H)', pp. 1–5.

Guan, D., Huang, J., Xiao, A. and Lu, S. (2021), Domain adaptive video segmentation via temporal consistency regularization, *in* 'International Conference on Computer Vision (ICCV)', IEEE/CVF, pp. 8053–8064.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A. C. (2017), Improved training of wasserstein gans, *in* 'International Conference on Neural Information Processing Systems (NIPS)', Curran Associates Inc., pp. 5769–5779.

Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q. (2017), On calibration of modern neural networks, *in* 'International Conference on Machine Learning (ICML)', PMLR, pp. 1321–1330.

Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N. and Halpern, A. (2016), 'Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)', *arXiv preprint arXiv:1605.01397* .

Gutmann, M. and Hyvärinen, A. (2010), Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, *in* 'International Conference on Artificial Intelligence and Statistics (AIStat)', JMLR, pp. 297–304.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A. et al. (2018), 'Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists', *Annals of Oncology* **29**(8), 1836–1842.

Han, S. S., Kim, M. S., Lim, W., Park, G. H., Park, I. and Chang, S. E. (2018), 'Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm', *Journal of Investigative Dermatology* **138**(7), 1529–1538.

He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R. (2020), Momentum contrast for unsupervised visual representation learning, *in* 'Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE/CVF, pp. 9729–9738.

He, K., Zhang, X., Ren, S. and Sun, J. (2016), Deep residual learning for image recognition, *in* 'Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE/CVF, pp. 770–778.

Hellman, M. E. (1970), 'The nearest neighbor classification rule with a reject option', *Transactions on Systems Science and Cybernetics* **6**(3), 179–185.

Henaff, O. (2020), Data-efficient image recognition with contrastive predictive coding, *in* 'International Conference on Machine Learning (ICML)', PMLR, pp. 4182–4192.

Hendrycks, D., Mazeika, M. and Dietterich, T. (2018), 'Deep anomaly detection with outlier exposure', *arXiv preprint arXiv:1812.04606* .

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J. and Lakshminarayanan, B. (2019), 'Augmix: A simple data processing method to improve robustness and uncertainty', *arXiv preprint arXiv:1912.02781* .

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A. and Bengio, Y. (2018), 'Learning deep representations by mutual information estimation and maximization', *arXiv preprint arXiv:1808.06670* .

Hou, L., Singh, K., Samaras, D., Kurc, T. M., Gao, Y., Seidman, R. J. and Saltz, J. H. (2016), Automatic histopathology image analysis with cnns, *in* 'New York Scientific Data Summit (NYSDS)', IEEE, pp. 1–6.

Houlsby, N., Huszár, F., Ghahramani, Z. and Lengyel, M. (2011), 'Bayesian active learning for classification and preference learning', *arXiv preprint arXiv:1112.5745* .

Hu, B., Tang, Y., Eric, I., Chang, C., Fan, Y., Lai, M. and Xu, Y. (2018), 'Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks', *Journal of Biomedical and Health Informatics* **23**(3), 1316–1328.

Islam, M. and Glocker, B. (2021), Spatially varying label smoothing: Capturing uncertainty from expert annotations, *in* 'International Conference on Information Processing in Medical Imaging (IPMI)', Springer, pp. 677–688.

Jaeger, P. F., Lüth, C. T., Klein, L. and Bungert, T. J. (2022), 'A call to reflect on evaluation practices for failure detection in image classification', *arXiv preprint arXiv:2211.15259* .

Jin, X., An, H., Wang, J., Wen, K. and Wu, Z. (2021), Reducing the annotation cost of whole slide histology images using active learning, *in* 'International Conference on Image Processing and Machine Vision (IPMV)', pp. 47–52.

Jones, O., Matin, R., van der Schaar, M., Bhayankaram, K. P., Ranmuthu, C., Islam, M., Behiyat, D., Boscott, R., Calanzani, N., Emery, J. et al. (2022), 'Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review', *The Lancet Digital Health* **4**(6), e466–e476.

Jospin, L. V., Laga, H., Boussaid, F., Buntine, W. and Bennamoun, M. (2022), 'Hands-on bayesian neural networks—a tutorial for deep learning users', *IEEE Computational Intelligence Magazine* **17**(2), 29–48.

Kaji, S. and Kida, S. (2019), 'Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging', *Radiological Physics and Technology* **12**, 235–248.

Kampffmeyer, M., Salberg, A.-B. and Jenssen, R. (2016), Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks, *in* 'Conference on Computer Vision and Pattern Recognition Workshops', CVF, pp. 1–9.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. and King, D. (2019), 'Key challenges for delivering clinical impact with artificial intelligence', *BMC Medicine* **17**, 1–9.

Kingma, D. P. and Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Kingma, D. P. and Welling, M. (2013), 'Auto-encoding variational bayes', *arXiv preprint arXiv:1312.6114* .

Kirsch, A., Van Amersfoort, J. and Gal, Y. (2019), 'Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning', *International Conference on Neural Information Processing Systems (NIPS)* pp. 7026–7037.

Konyushkova, K., Sznitman, R. and Fua, P. (2019), 'Geometry in active learning for binary and multi-class image segmentation', *Computer Vision and Image Understanding* **182**, 1–16.

Kramer, M. A. (1991), 'Nonlinear principal component analysis using autoassociative neural networks', *AIChE Journal* **37**(2), 233–243.

Kristiadi, A., Hein, M. and Hennig, P. (2020), Being bayesian, even just a bit, fixes overconfidence in relu networks, *in* 'International Conference on Machine Learning (ICML)', PMLR, pp. 5436–5446.

Krizhevsky, A., Hinton, G. et al. (2009), *Learning multiple layers of features from tiny images*, University of Toronto.

Kullback, S. and Leibler, R. A. (1951), 'On information and sufficiency', *The Annals of Mathematical Statistics* **22**(1), 79–86.

Kurakin, A., Goodfellow, I. J. and Bengio, S. (2018), Adversarial examples in the physical world, *in* 'Artificial intelligence Safety and Security', Chapman and Hall/CRC, pp. 99–112.

Kwon, Y., Won, J.-H., Kim, B. J. and Paik, M. C. (2020), 'Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation', *Computational Statistics & Data Analysis* **142**, 106816.

Lake, B. M., Salakhutdinov, R. and Tenenbaum, J. B. (2015), 'Human-level concept learning through probabilistic program induction', *Science* **350**(6266), 1332–1338.

LeCun, Y., Bengio, Y. and Hinton, G. (2015), 'Deep learning', *Nature* **521**(7553), 436–444.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998), 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE* **86**(11), 2278–2324.

Lee, D.-H. et al. (2013), ‘Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks’, *Workshop on Challenges in Representation Learning* **3**(2), 896.

Lewis, D. D. (1995), ‘A sequential algorithm for training text classifiers: Corrigendum and additional data’, *Sigir Forum* **29**(2), 13–19.

Liang, G., Zhang, Y., Wang, X. and Jacobs, N. (2020), ‘Improved trainable calibration method for neural networks on medical imaging classification’, *arXiv preprint arXiv:2009.04057* .

Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017), Focal loss for dense object detection, *in* ‘International Conference on Computer Vision (ICCV)’, IEEE, pp. 2980–2988.

Liopyris, K., Gregoriou, S., Dias, J. and Stratigos, A. J. (2022), ‘Artificial intelligence in dermatology: challenges and perspectives’, *Dermatology and Therapy* **12**(12), 2637–2651.

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R. et al. (2018), ‘1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset’, *GigaScience* **7**(6), giy065.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B. and Sánchez, C. I. (2017), ‘A survey on deep learning in medical image analysis’, *Medical Image Analysis* **42**, 60–88.

Liu, D. C. and Nocedal, J. (1989), ‘On the limited memory BFGS method for large scale optimization’, *Mathematical Programming* **45**, 503–528.

Liu, L., Lei, W., Wan, X., Liu, L., Luo, Y. and Feng, C. (2020), Semi-supervised active learning for covid-19 lung ultrasound multi-symptom classification, *in* ‘International Conference on Tools with Artificial Intelligence (ICTAI)’, IEEE, pp. 1268–1273.

Liu, X., Sanchez, P., Thermos, S., O’Neil, A. Q. and Tsaftaris, S. A. (2022), ‘Learning disentangled representations in the imaging domain’, *Medical Image Analysis* **80**, 102516–102516.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021), Swin transformer: Hierarchical vision transformer using shifted windows, *in* ‘International Conference on Computer Vision (ICCV)’, IEEE/CVF, pp. 10012–10022.

MacKay, D. J. (1992), ‘Bayesian interpolation’, *Neural Computation* **4**(3), 415–447.

Mackowiak, R., Lenz, P., Ghori, O., Diego, F., Lange, O. and Rother, C. (2018), ‘Cereals-cost-effective region-based active learning for semantic segmentation’, *arXiv preprint arXiv:1810.09726* .

Madabhushi, A. and Lee, G. (2016), ‘Image analysis and machine learning in digital pathology: Challenges and opportunities’, *Medical Image Analysis* **33**, 170–175.

Maier-Hein, L., Reinke, A., Christodoulou, E., Glocker, B., Godau, P., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A. et al. (2022), ‘Metrics reloaded: Pitfalls and recommendations for image analysis validation’, *arXiv preprint arXiv:2206.01653* .

Makhzani, A. and Frey, B. (2013), ‘K-sparse autoencoders’, *arXiv preprint arXiv:1312.5663* .

Maron, R. C., Weichenthal, M., Utikal, J. S., Hekler, A., Berking, C., Hauschild, A., Enk, A. H., Haferkamp, S., Klode, J., Schadendorf, D. et al. (2019), ‘Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks’, *European Journal of Cancer* **119**, 57–65.

Matsoukas, C., Haslum, J. F., Sorkhei, M., Söderberg, M. and Smith, K. (2022), What makes transfer learning work for medical images: Feature reuse & other factors, *in* ‘Conference on Computer Vision and Pattern Recognition (CVPR)’, IEEE/CVF, pp. 9225–9234.

Misra, I. and Maaten, L. v. d. (2020), Self-supervised learning of pretext-invariant representations, *in* ‘Conference on Computer Vision and Pattern Recognition (CVPR)’, IEEE/CVF, pp. 6707–6717.

Mukherjee, S., Asnani, H., Lin, E. and Kannan, S. (2019), ‘Clustergan: Latent space clustering in generative adversarial networks’, *Conference on Artificial Intelligence (AAAI)* **33**(1), 4610–4617.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. and Dokania, P. (2020), ‘Calibrating deep neural networks using focal loss’, *Interna-*

tional Conference on Neural Information Processing Systems (NIPS) pp. 15288–15299.

Müller, R., Kornblith, S. and Hinton, G. E. (2019), 'When does label smoothing help?', *International Conference on Neural Information Processing Systems (NIPS)* p. 4694–4703.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G. and Tran, D. (2019), 'Measuring calibration in deep learning', *Uncertainty and Robustness in Deep Visual Learning Workshop (Uncertainty)* **2**(7).

Noroozi, M. and Favaro, P. (2016), Unsupervised learning of visual representations by solving jigsaw puzzles, *in* 'European Conference on Computer Vision (ECCV)', Springer, pp. 69–84.

Parikh, N., Boyd, S. et al. (2014), 'Proximal algorithms', *Foundations and Trends in Optimization* **1**(3), 127–239.

Parzen, E. (1962), 'On estimation of a probability density function and mode', *The Annals of Mathematical Statistics* **33**(3), 1065–1076.

Platt, J. (1999), 'Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods', *Advances in Large Margin Classifiers* **10**(3), 61–74.

Pop, R. and Fulop, P. (2018), 'Deep ensemble bayesian active learning: Addressing the mode collapse issue in monte carlo dropout via ensembles', *arXiv preprint arXiv:1811.03897* .

Prados, F., Ashburner, J., Blaiotta, C., Brosch, T., Carballido-Gamio, J., Cardoso, M. J., Conrad, B. N., Datta, E., Dávid, G., De Leener, B. et al. (2017), 'Spinal cord grey matter segmentation challenge', *Neuroimage* **152**, 312–329.

Prechelt, L. (2012), 'Early stopping—but when?', *Neural Networks: Tricks of the Trade* pp. 53–67.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H. and Raiko, T. (2015), 'Semi-supervised learning with ladder networks', pp. 3546–3554.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X. and Wang, X. (2021), 'A survey of deep active learning', *Computing Surveys* **54**(9), 1–40.

Roelofs, R., Cain, N., Shlens, J. and Mozer, M. C. (2022), Mitigating bias in calibration error estimation, *in* 'International Conference on Artificial Intelligence and Statistics (AIStat)', PMLR, pp. 4036–4054.

Roy, A. G., Ren, J., Azizi, S., Loh, A., Natarajan, V., Mustafa, B., Pawlowski, N., Freyberg, J., Liu, Y., Beaver, Z. et al. (2022), 'Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions', *Medical Image Analysis* **75**, 102274.

Roy, N. and McCallum, A. (2001), 'Toward optimal active learning through monte carlo estimation of error reduction', *International Conference on Machine Learning (ICML)* **2**, 441–448.

Sarker, I. H. (2021), 'Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions', *SN Computer Science* **2**(6), 420.

Sener, O. and Savarese, S. (2017), Active learning for convolutional neural networks: A core-set approach, *in* 'International Conference on Learning Representations (ICLR)'.

Settles, B. (2009), *Active learning literature survey*, University of Wisconsin-Madison Department of Computer Sciences.

Settles, B. and Craven, M. (2008), An analysis of active learning strategies for sequence labeling tasks, *in* 'Empirical Methods in Natural Language Processing (EMNLP)', ACL, pp. 1070–1079.

Settles, B., Craven, M. and Friedland, L. (2008), Active learning with real annotation costs, *in* 'Cost-Sensitive Learning', Vol. 1.

Settles, B., Craven, M. and Ray, S. (2007), 'Multiple-instance active learning', *International Conference on Neural Information Processing Systems (NIPS)* pp. 1289–1296.

Seung, H. S., Opper, M. and Sompolinsky, H. (1992), Query by committee, *in* 'Workshop on Computational Learning Theory', pp. 287–294.

Shannon, C. E. (1948), 'A mathematical theory of communication', *The Bell System Technical Journal* **27**(3), 379–423.

Shao, W., Sun, L. and Zhang, D. (2018), Deep active learning for nucleus classification in pathology images, *in* 'International Symposium on Biomedical Imaging (ISBI)', IEEE, pp. 199–202.

Shi, X., Dou, Q., Xue, C., Qin, J., Chen, H. and Heng, P.-A. (2019), An active learning approach for reducing annotation cost in skin lesion analysis, *in* 'Machine Learning in Medical Imaging (MLMI)', Springer, pp. 628–636.

Shurrab, S. and Duwairi, R. (2022), 'Self-supervised learning methods and applications in medical imaging analysis: A survey', *PeerJ Computer Science* **8**, 1045.

Sirinukunwattana, K., Raza, S. E. A., Tsang, Y.-W., Snead, D. R., Cree, I. A. and Rajpoot, N. M. (2016), 'Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images', *Transactions on Medical Imaging* **35**(5), 1196–1206.

Smith, L. N. (2017), Cyclical learning rates for training neural networks, *in* 'Winter Conference on Applications of Computer Vision (WACV)', IEEE, pp. 464–472.

Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., Flach, P. et al. (2021), 'Classifier calibration: How to assess and improve predicted class probabilities: a survey', *arXiv preprint arXiv:2112.10327* .

Srinidhi, C. L., Ciga, O. and Martel, A. L. (2021), 'Deep neural network models for computational histopathology: A survey', *Medical Image Analysis* **67**, 101813.

Srivastav, D., Bajpai, A. and Srivastava, P. (2021), Improved classification for pneumonia detection using transfer learning with gan based synthetic image augmentation, *in* 'International Conference on Cloud Computing, Data Science & Engineering (Confluence)', IEEE, pp. 433–437.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014), 'Dropout: a simple way to prevent neural networks from overfitting', *The Journal of Machine Learning Research* **15**(1), 1929–1958.

Stacke, K., Eilertsen, G., Unger, J. and Lundström, C. (2020), 'Measuring domain shift for deep learning in histopathology', *Journal of Biomedical and Health Informatics* **25**(2), 325–336.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016), Rethinking the inception architecture for computer vision, *in* 'Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE/CVF, pp. 2818–2826.

Tan, M. and Le, Q. (2019), EfficientNet: Rethinking model scaling for convolutional neural networks, *in* 'International Conference on Machine Learning (ICML)', PMLR, pp. 6105–6114.

Thiagarajan, J. J., Venkatesh, B., Rajan, D. and Sattigeri, P. (2020), Improving reliability of clinical models using prediction calibration, *in* 'Uncertainty for Safe Utilization of Machine Learning (UNSURE)', Springer, pp. 71–80.

Tizhoosh, H. R. and Pantanowitz, L. (2018), 'Artificial intelligence and digital pathology: challenges and opportunities', *Journal of Pathology Informatics* **9**(1), 38.

Tschandl, P., Rosendahl, C., Akay, B. N., Argenziano, G., Blum, A., Braun, R. P., Cabo, H., Gourhant, J.-Y., Kreusch, J., Lallas, A. et al. (2019), 'Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks', *JAMA Dermatology* **155**(1), 58–65.

Tschandl, P., Rosendahl, C. and Kittler, H. (2018), 'The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions', *Scientific Data* **5**(1), 1–9.

Ulmer, D., Meijerink, L. and Cinà, G. (2020), Trust issues: Uncertainty estimation does not enable reliable OOD detection on medical tabular data, *in* 'Machine Learning for Health', PMLR, pp. 341–354.

van den Oord, A., Kalchbrenner, N. and Kavukcuoglu, K. (2016), Pixel recurrent neural networks, *in* 'International Conference on Machine Learning (ICML)', PMLR, pp. 1747–1756.

van den Oord, A., Li, Y. and Vinyals, O. (2018), 'Representation learning with contrastive predictive coding', *arXiv preprint arXiv:1807.03748* .

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T. and Welling, M. (2018), Rotation equivariant cnns for digital pathology, *in* 'Medical Image Computing and Computer Assisted Intervention (MICCAI)', Springer, pp. 210–218.

Wang, K., Zhang, D., Li, Y., Zhang, R. and Lin, L. (2016), 'Cost-effective active learning for deep image classification', *Transactions on Circuits and Systems for Video Technology* **27**(12), 2591–2600.

Wang, Z., Dai, Z., Póczos, B. and Carbonell, J. (2019), Characterizing and avoiding negative transfer, *in* 'Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE/CVF, pp. 11293–11302.

Wei, R. and Mahmood, A. (2020), 'Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey', *Access* **9**, 4939–4956.

Weiss, K., Khoshgoftaar, T. M. and Wang, D. (2016), 'A survey of transfer learning', *Journal of Big data* **3**(1), 1–40.

Wen, D., Khan, S. M., Xu, A. J., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A. K., Liu, X. et al. (2021), 'Characteristics of publicly available skin cancer image datasets: a systematic review', *The Lancet Digital Health* .

Wiener, Y. and El-Yaniv, R. (2015), 'Agnostic pointwise-competitive selective classification', *Journal of Artificial Intelligence Research* **52**, 171–201.

Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R. and Zhao, S. (2022), 'Skin cancer classification with deep learning: A systematic review', *Frontiers in Oncology* **12**.

Wu, Z., Xiong, Y., Stella, X. Y. and Lin, D. (2018), Unsupervised feature learning via non-parametric instance discrimination, *in* 'Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE/CVF, pp. 3733–3742.

Yang, J., Wu, X., Liang, J., Sun, X., Cheng, M.-M., Rosin, P. L. and Wang, L. (2019), 'Self-paced balance learning for clinical skin disease recognition', *Transactions on Neural Networks and Learning Systems* **31**(8), 2832–2846.

Yi, X., Walia, E. and Babyn, P. (2019), 'Generative adversarial network in medical imaging: A review', *Medical Image Analysis* **58**, 101552.

Zeiler, M. D. (2012), 'Adadelta: an adaptive learning rate method', *arXiv preprint arXiv:1212.5701* .

Zhang, J., Kailkhura, B. and Han, T. Y.-J. (2020), Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning, *in* 'International Conference on Machine Learning (ICML)', PMLR, pp. 11117–11128.

Zhang, Q.-s. and Zhu, S.-C. (2018), 'Visual interpretability for deep learning: a survey', *Frontiers of Information Technology & Electronic Engineering* **19**(1), 27–39.

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V. and Dalca, A. V. (2019), Data augmentation using learned transformations for one-shot medical image segmentation, *in* 'Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE/CVF, pp. 8543–8553.

Zhao, S., Kim, M., Sahoo, R., Ma, T. and Ermon, S. (2021), 'Calibrating predictions to decisions: A novel approach to multi-class calibration', pp. 22313–22324.

Zhdanov, F. (2019), 'Diverse mini-batch active learning', *arXiv preprint arXiv:1901.05954* .

Zhou, H.-Y., Lu, C., Yang, S., Han, X. and Yu, Y. (2021), Preservational learning improves self-supervised medical image models by reconstructing diverse contexts, *in* 'International Conference on Computer Vision (ICCV)', IEEE/CVF, pp. 3499–3509.

Zhuang, C., Zhai, A. L. and Yamins, D. (2019), Local aggregation for unsupervised learning of visual embeddings, *in* 'Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE/CVF, pp. 6002–6012.

Appendices

Appendix A

List of Publications and Achievements

A.1 First Authored Publications

- **Carse, J**. & McKenna, S. (2019), Active learning for patch-based digital pathology using convolutional neural networks to reduce annotation costs, *in* 'European Congress on Digital Pathology', Springer, pp. 20-27.

- **Carse, J**., Carey, F. & McKenna, S. (2021), Unsupervised representation learning from pathology images with multi-directional contrastive predictive coding, *in* '18$^{\text{th}}$ International Symposium on Biomedical Imaging', IEEE, pp. 1254-1258.

- **Carse, J**., Süveges, T., Hogg, S., Trucco, E., Proby, C., Fleming, C. & McKenna, S. (2021), Robust selective classification of skin lesions with asymmetric costs, *in* 'Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis', Springer, pp. 112-121.

- **Carse, J**., Alvarez Olmo, A. & McKenna, S. (2022) Calibration of deep medical image classifiers: An empirical comparison using dermatology and histopathology datasets, *in* 'International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging', Springer, pp. 89-99.

A.2 Contributed Publications

- Chin, G., Süveges, T., **Carse, J.**, Butt, S., Muthiah, S., Morton, C., Trucco, E., Proby, C., McKenna, S. & Fleming, C. (2022), 'Prepare to succeed: Real-world image datasets for artificial intelligence in skin cancer triage', *British Journal of Dermatology* **187**(Supplement 1), 125.

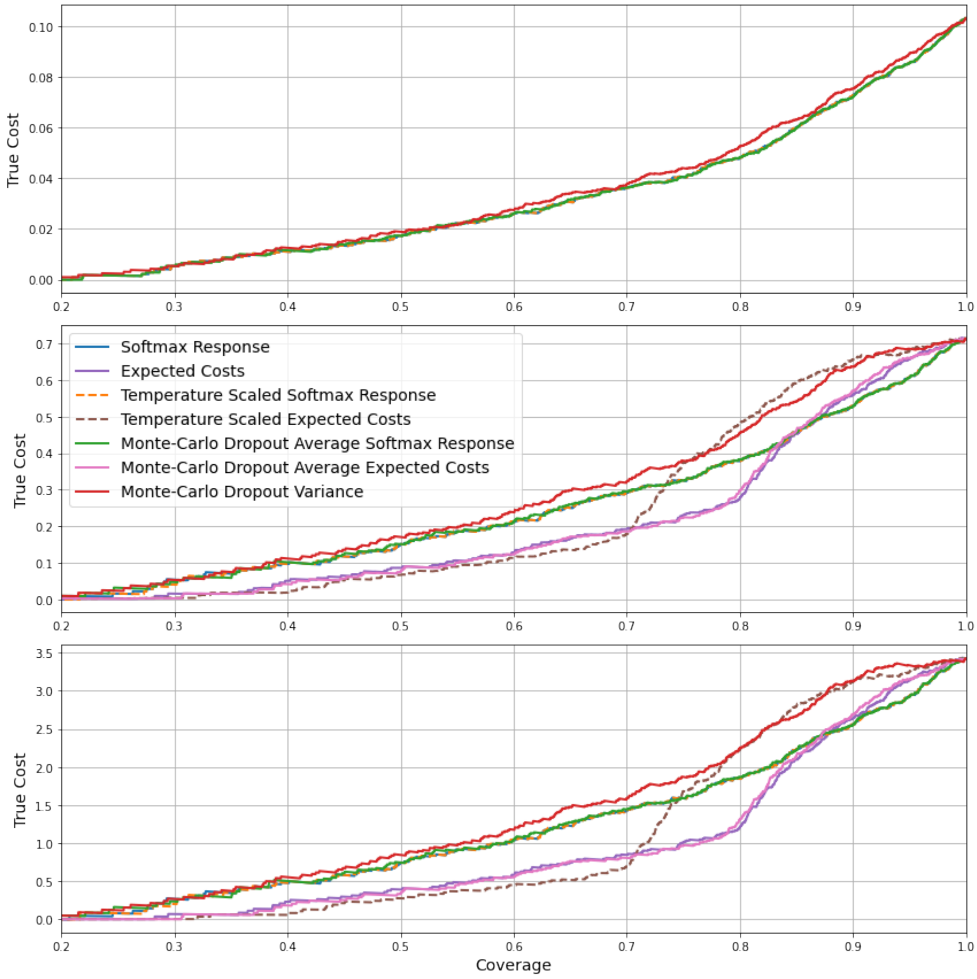
A.3 Awards

- Best Paper - University of Dundee Computing PhD Symposium 2019
- Best PhD Blitz Presentation - SINAPSE Annual General Meeting 2019
- Best Reviewer - University of Dundee Computing PhD Symposium 2022

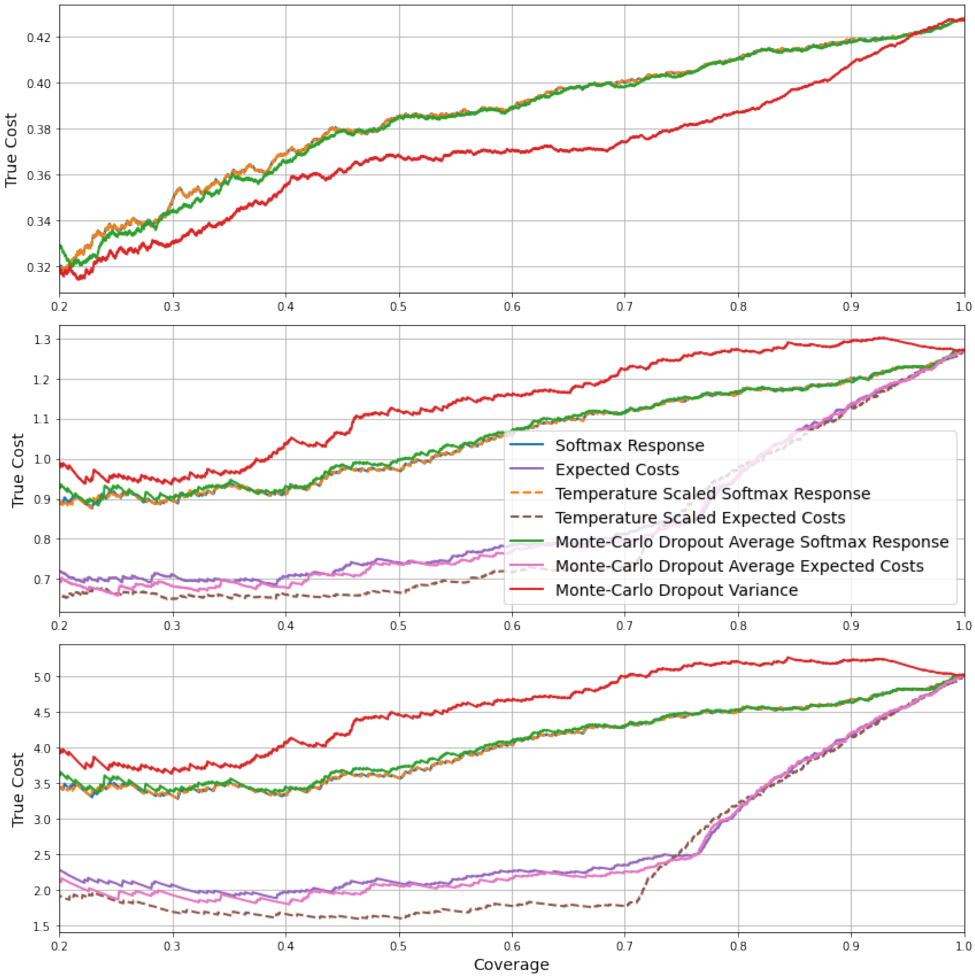
Appendix B

Selective Classification Results

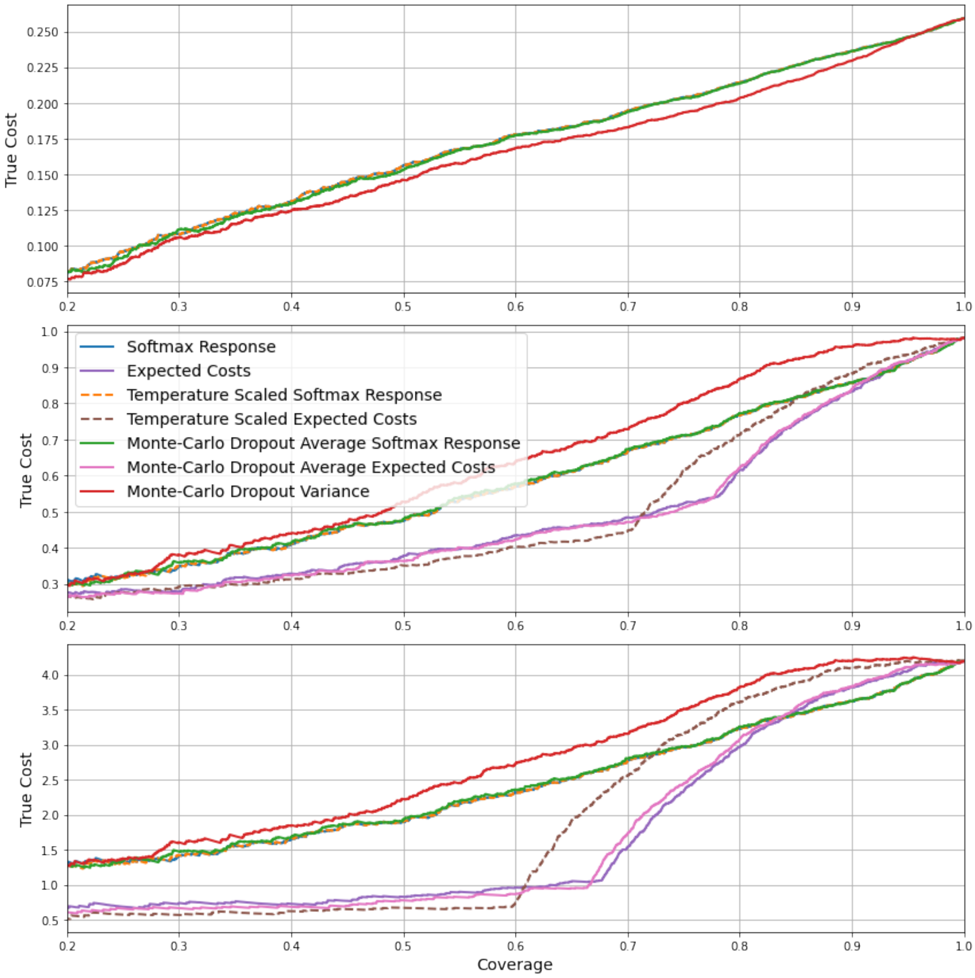
B.1 Binary Classification Experiment Results



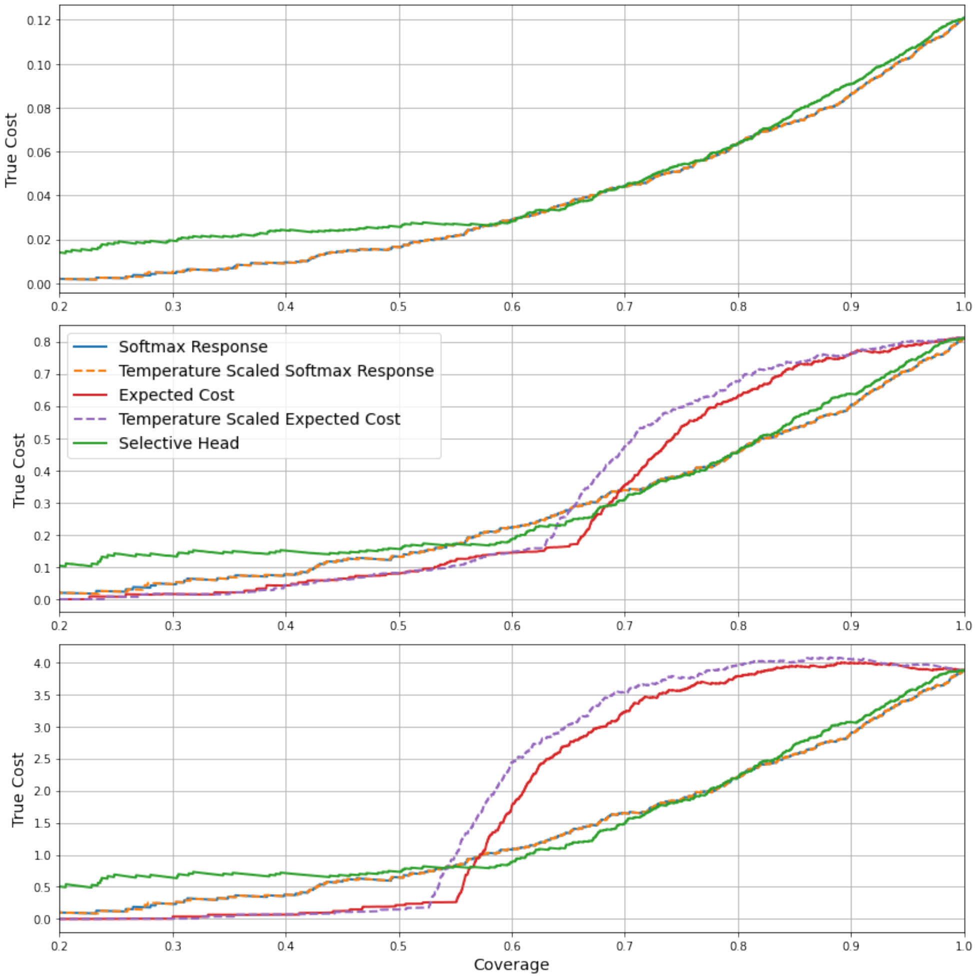
Experiments using a CNN model testing on the S_{in} distribution.



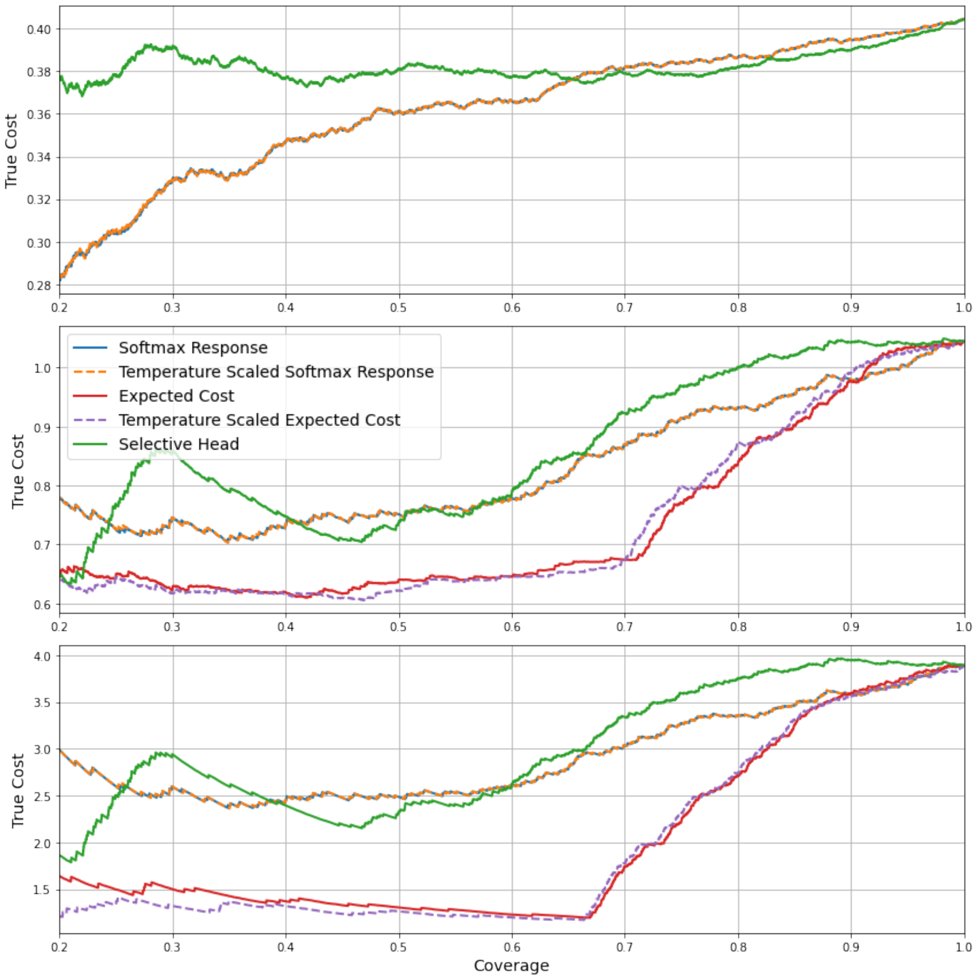
Experiments using a CNN model testing on the $S_{unknown}$ distribution.



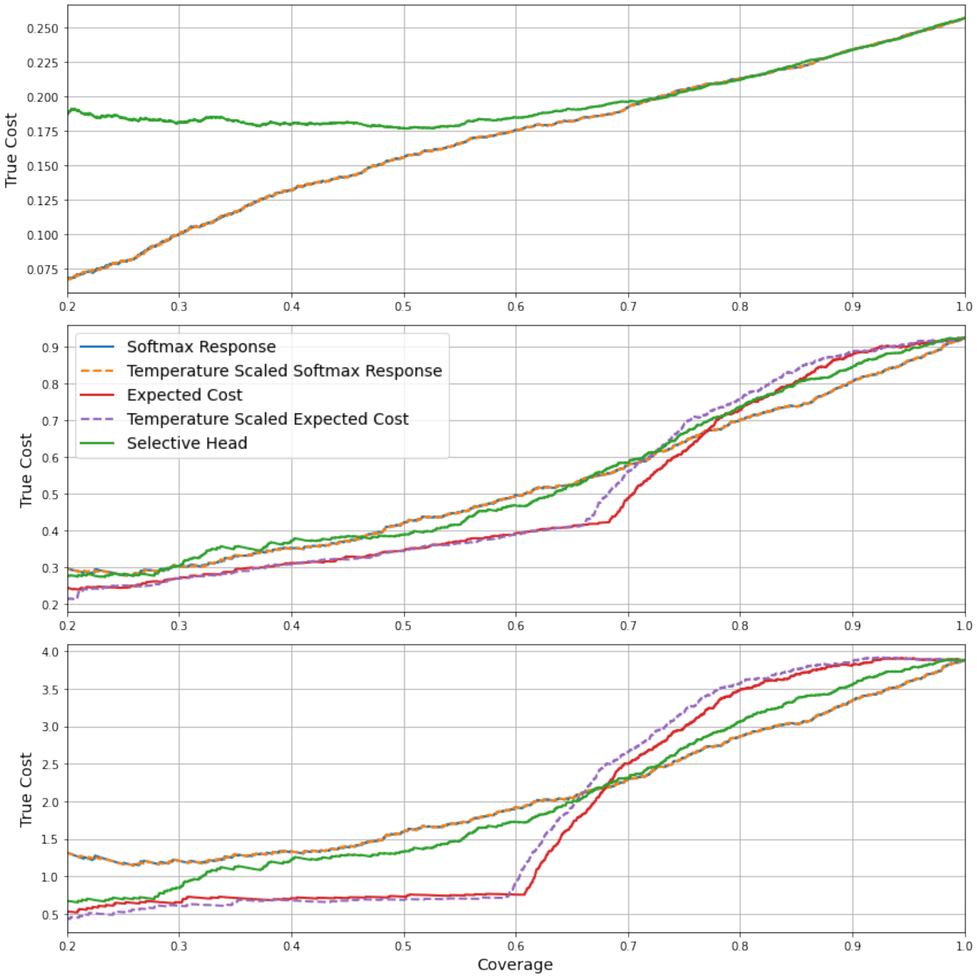
Experiments using a CNN model testing on the $S_{combined}$ distribution.



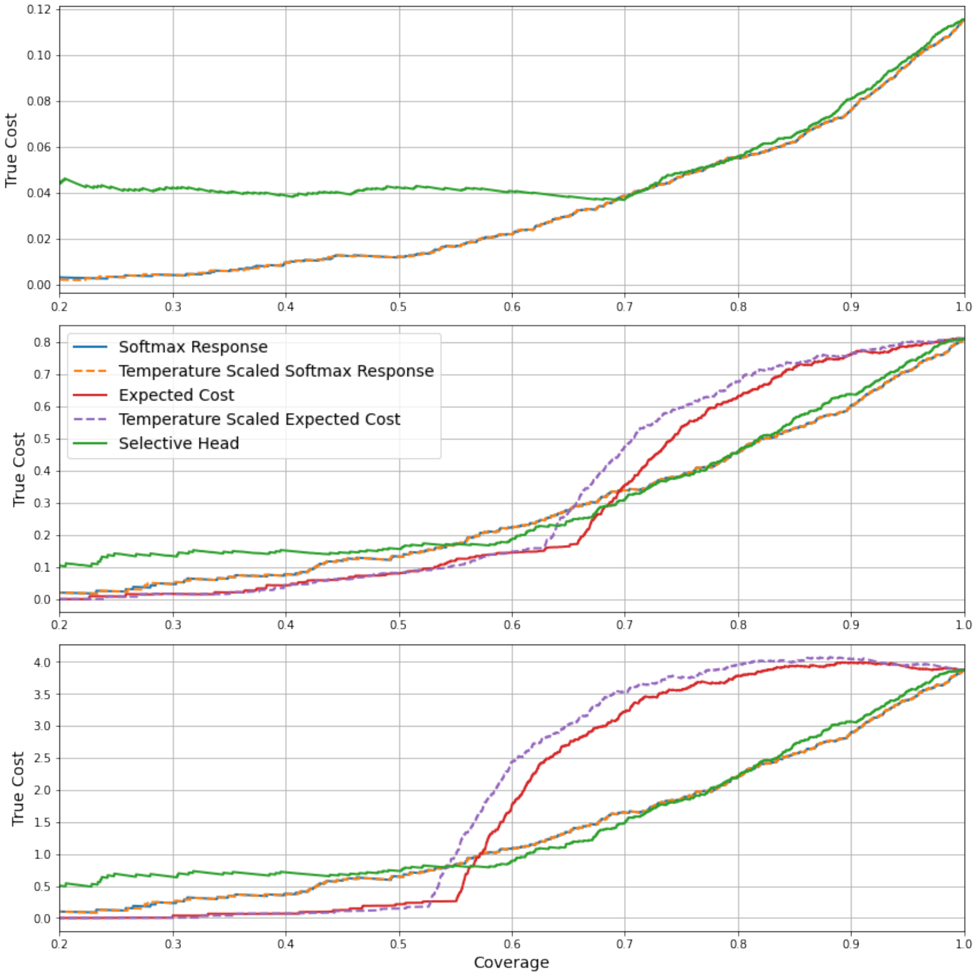
Experiments using a SelectiveNet model trained for a target coverage of 0.7, testing on the S_{in} distribution.



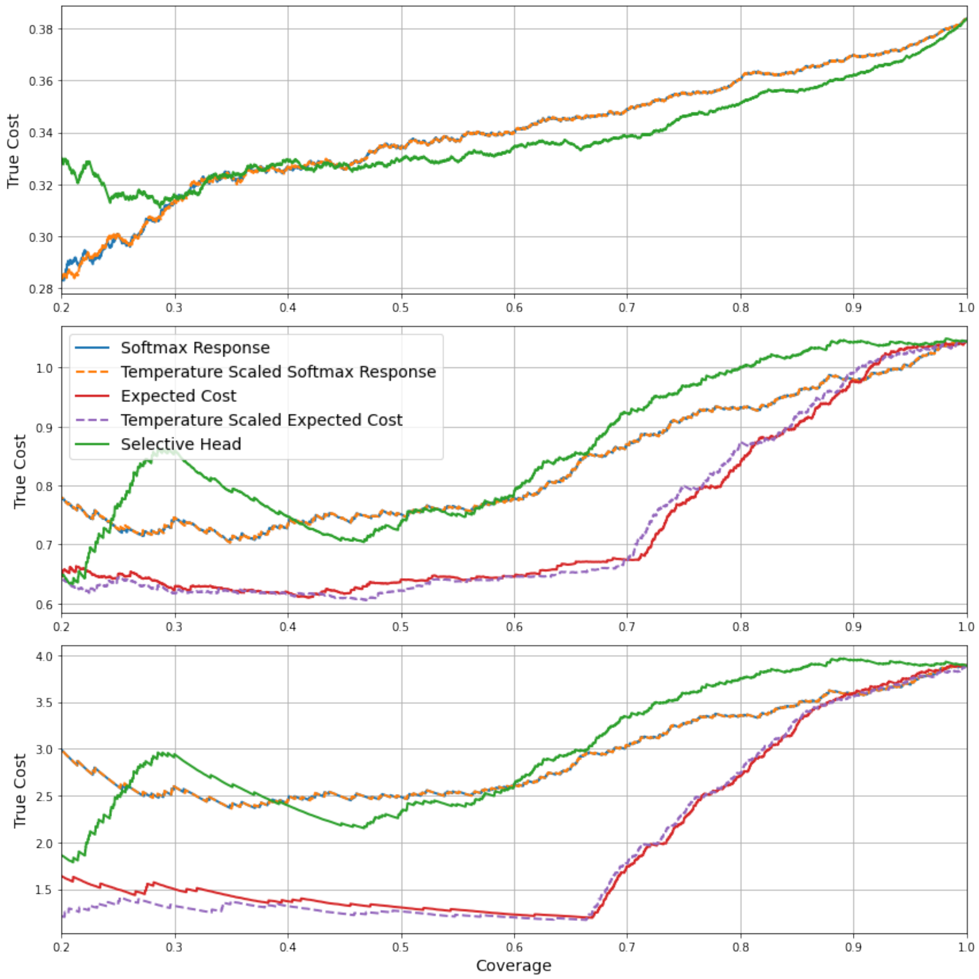
Experiments using a SelectiveNet model trained for a target coverage of 0.7, testing on the $S_{unknown}$ distribution.



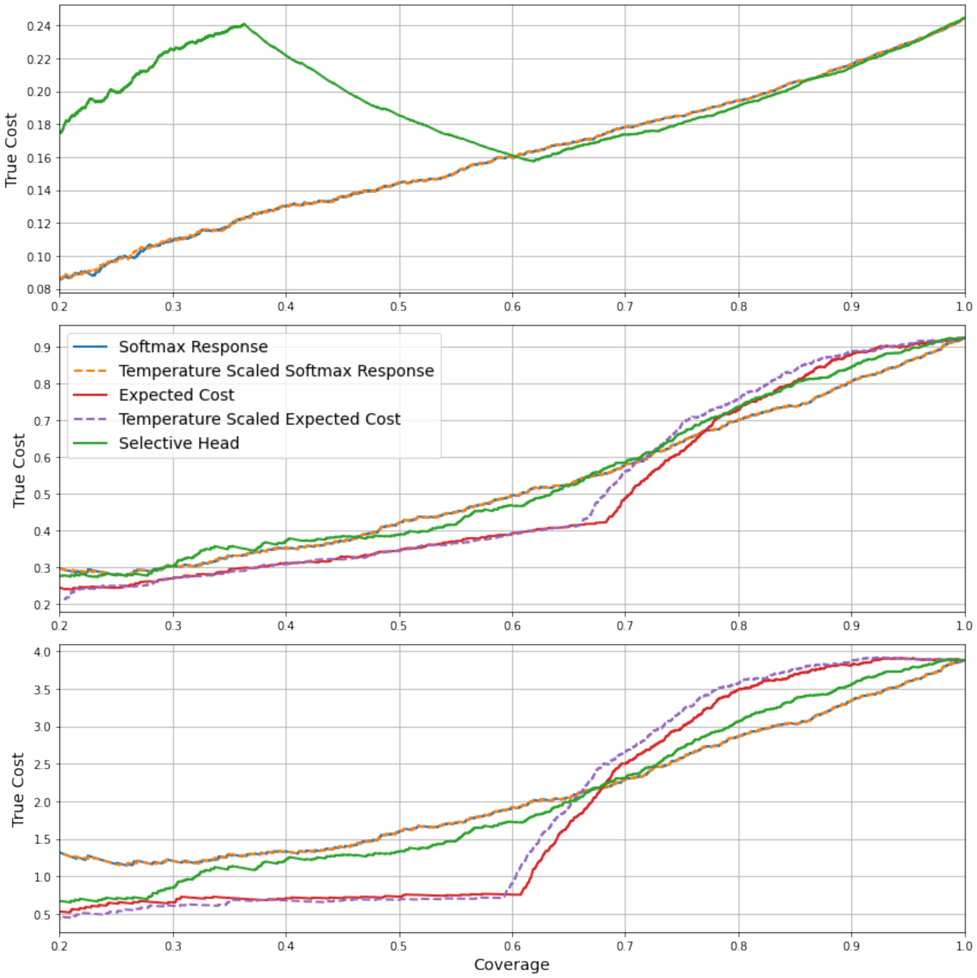
Experiments using a SelectiveNet model trained for a target coverage of 0.7, testing on the $S_{combined}$ distribution.



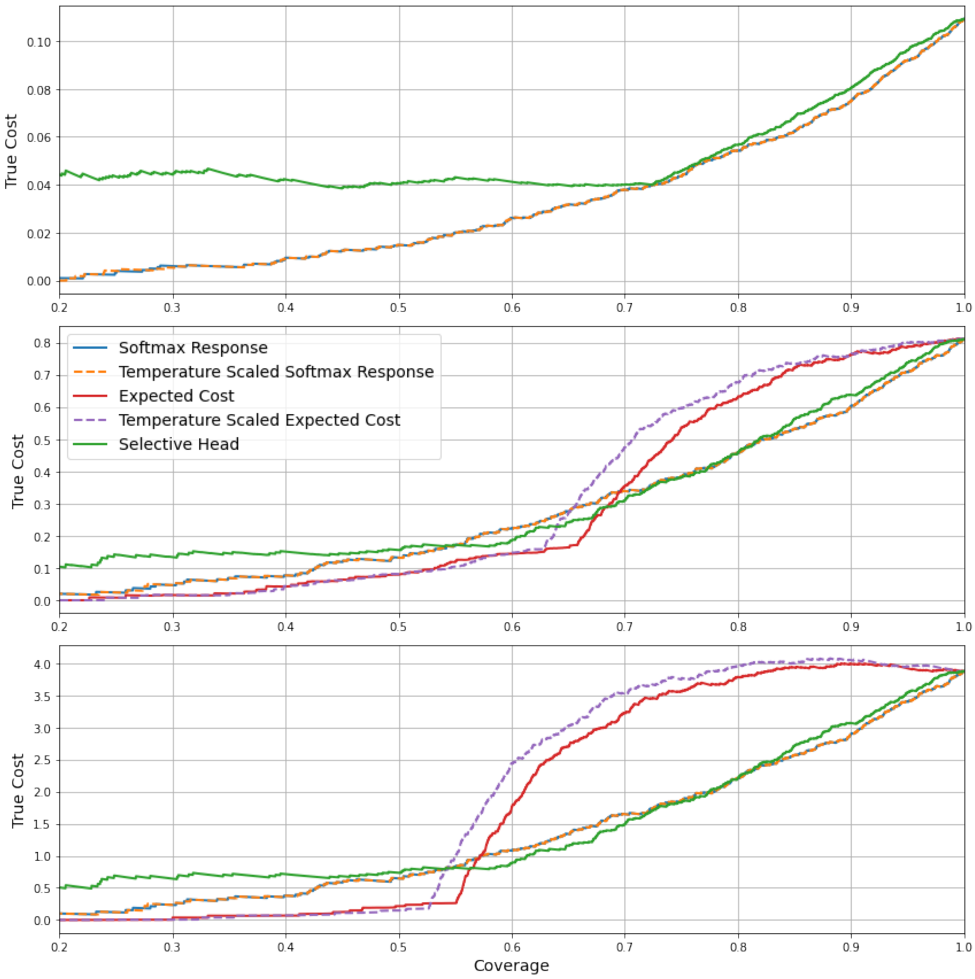
Experiments using a SelectiveNet model trained for a target coverage of 0.75, testing on the S_{in} distribution.



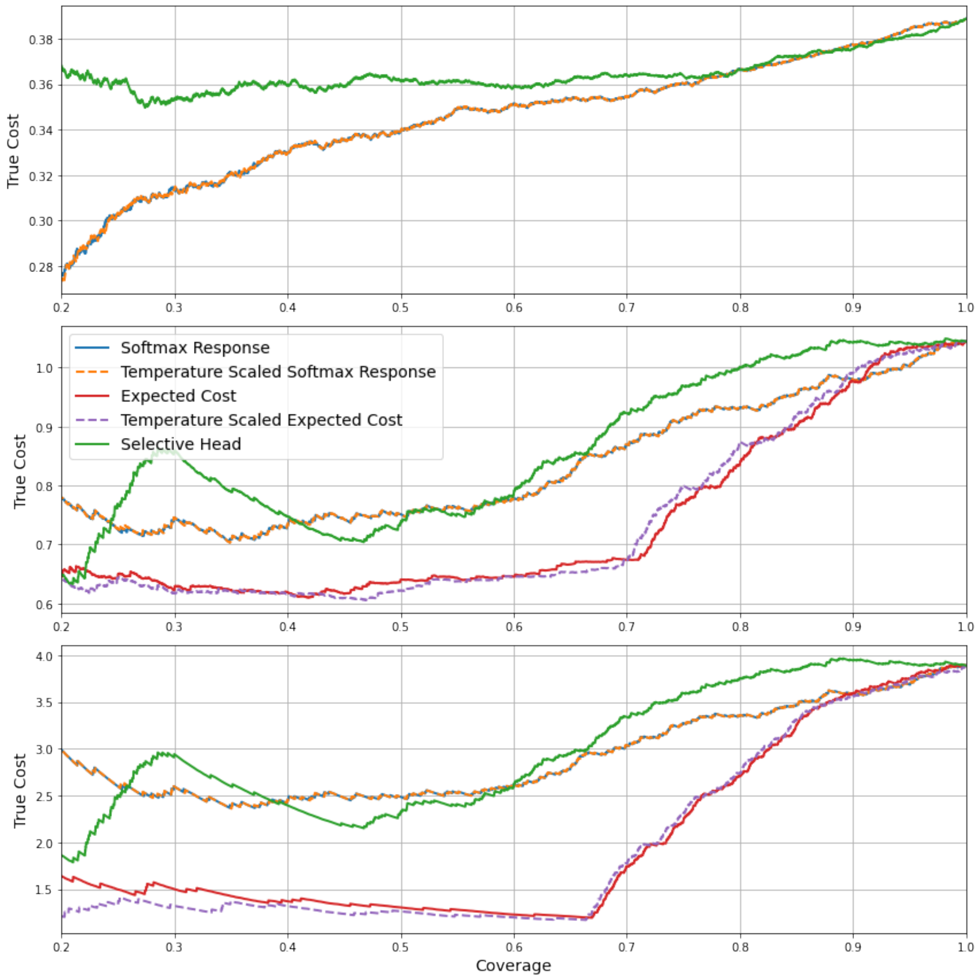
Experiments using a SelectiveNet model trained for a target coverage of 0.75, testing on the $S_{unknown}$ distribution.



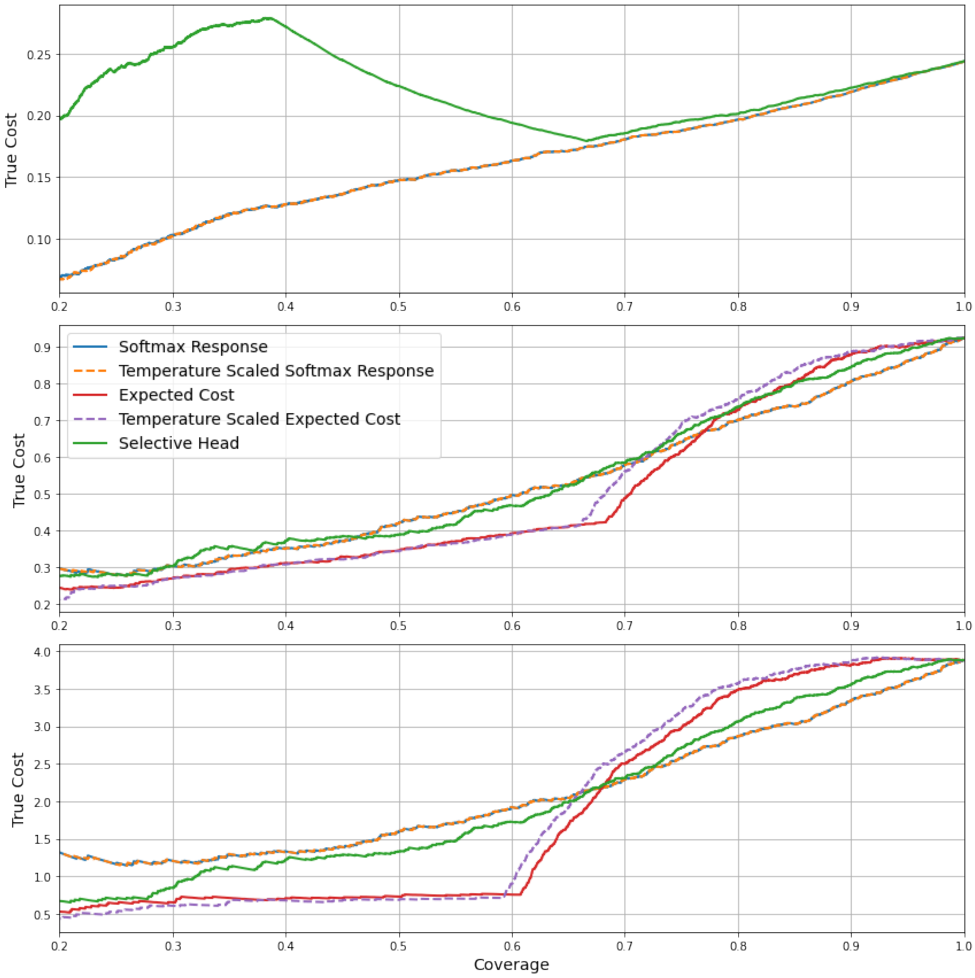
Experiments using a SelectiveNet model trained for a target coverage of 0.75, testing on the $S_{combined}$ distribution.



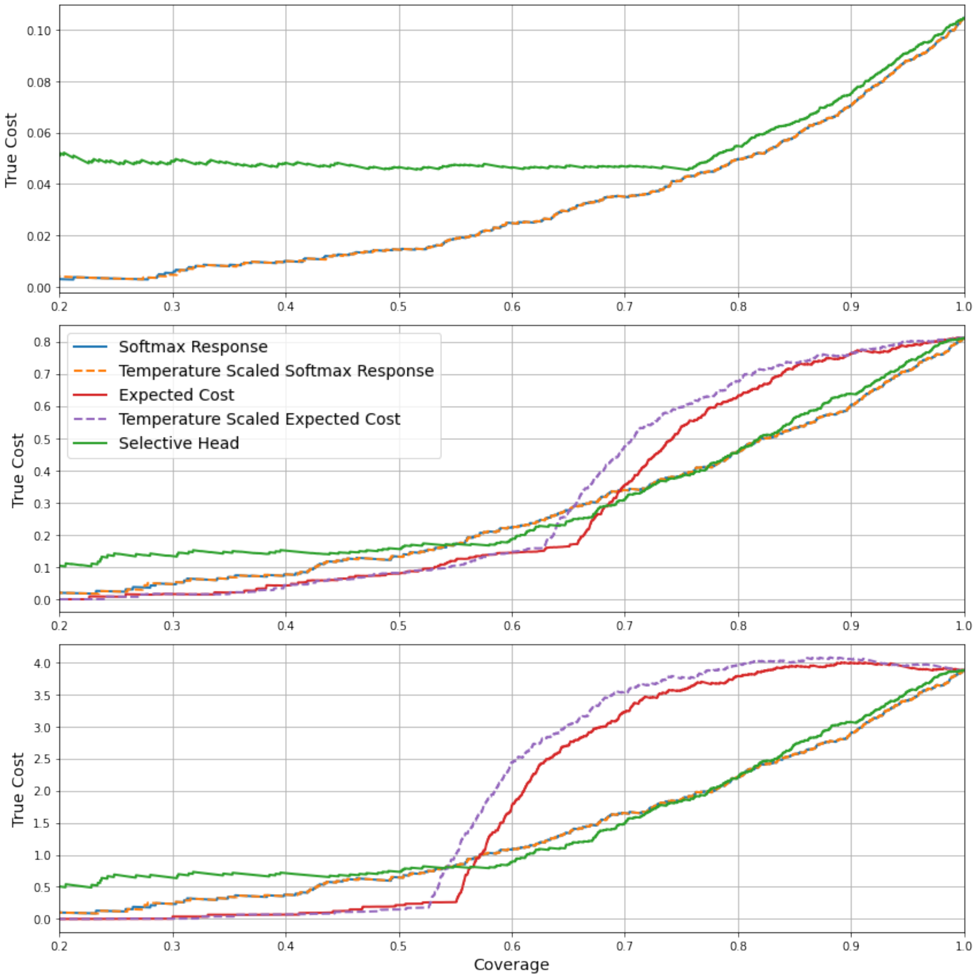
Experiments using a SelectiveNet model trained for a target coverage of 0.8, testing on the S_{in} distribution.



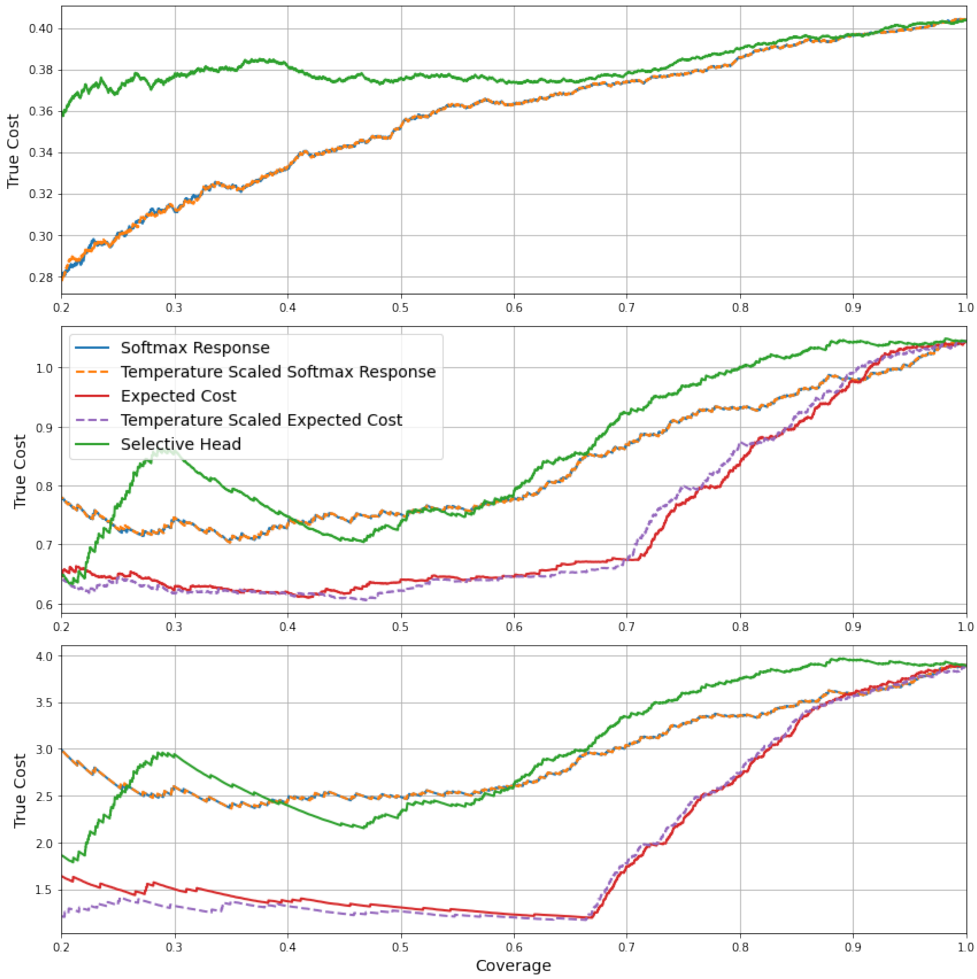
Experiments using a SelectiveNet model trained for a target coverage of 0.8, testing on the $S_{unknown}$ distribution.



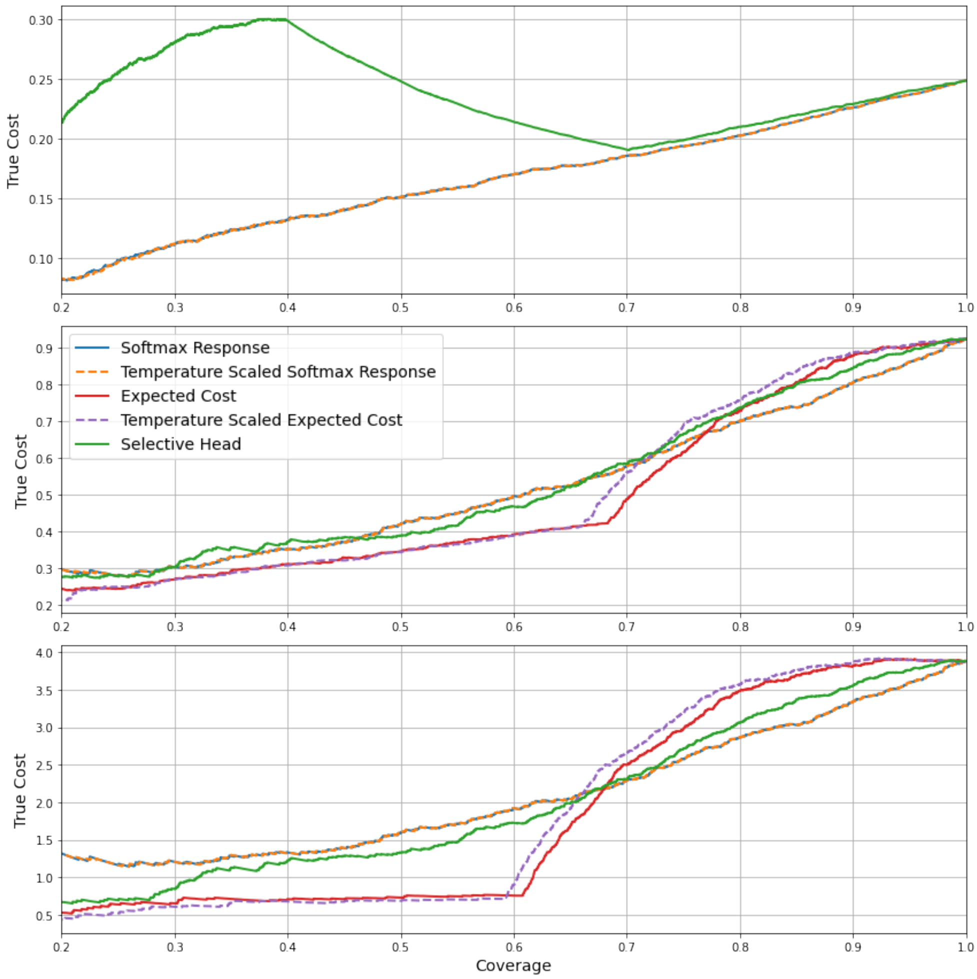
Experiments using a SelectiveNet model trained for a target coverage of 0.8, testing on the $S_{combined}$ distribution.



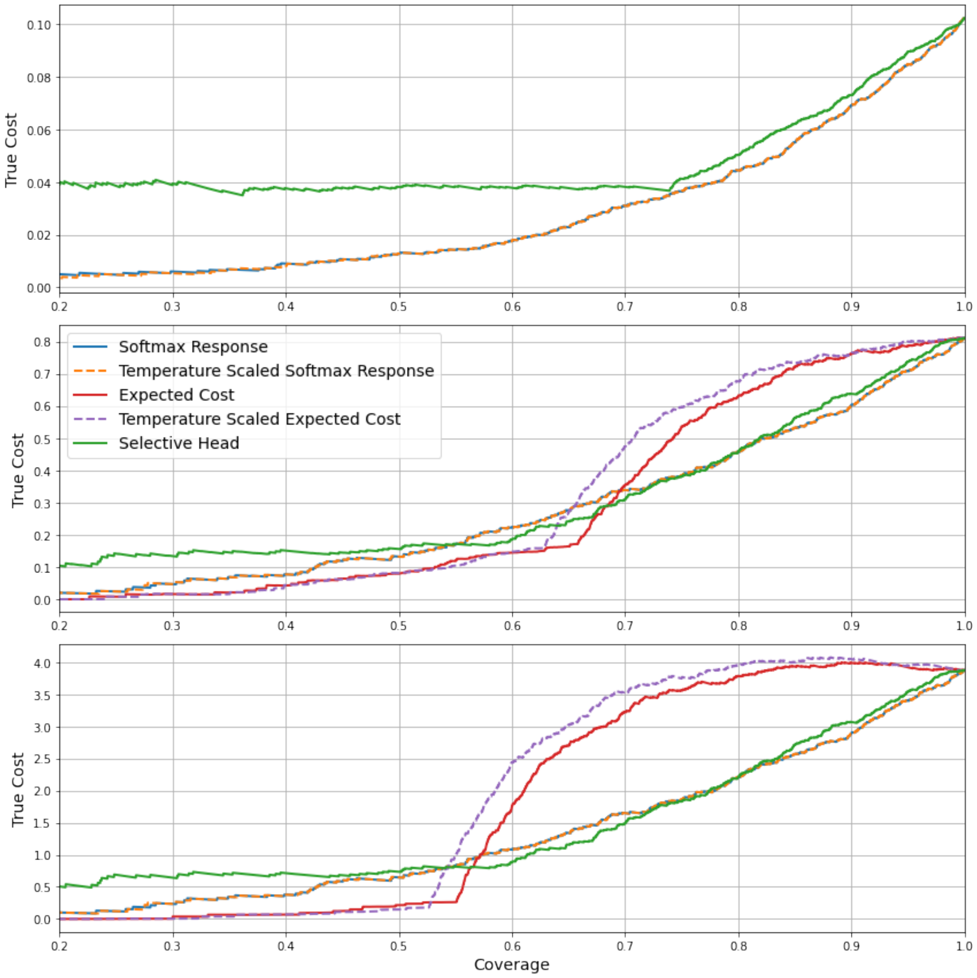
Experiments using a SelectiveNet model trained for a target coverage of 0.85, testing on the S_{in} distribution.



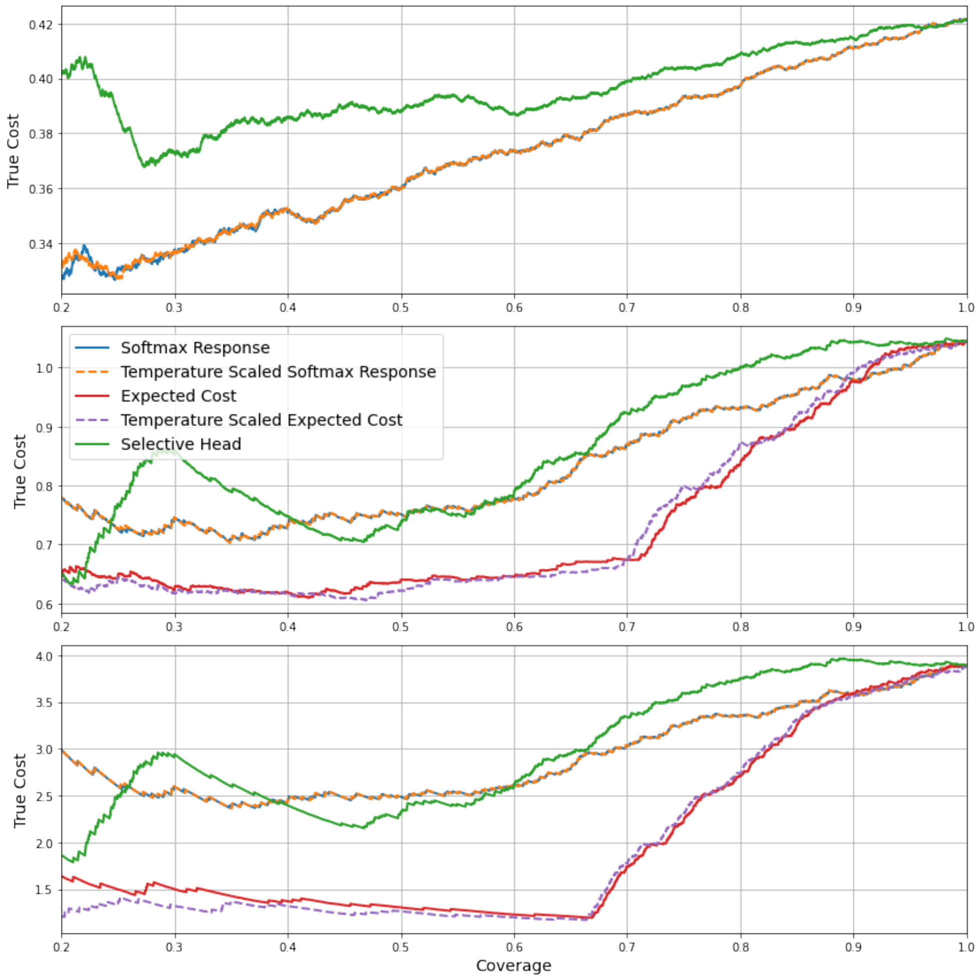
Experiments using a SelectiveNet model trained for a target coverage of 0.85, testing on the $S_{unknown}$ distribution.



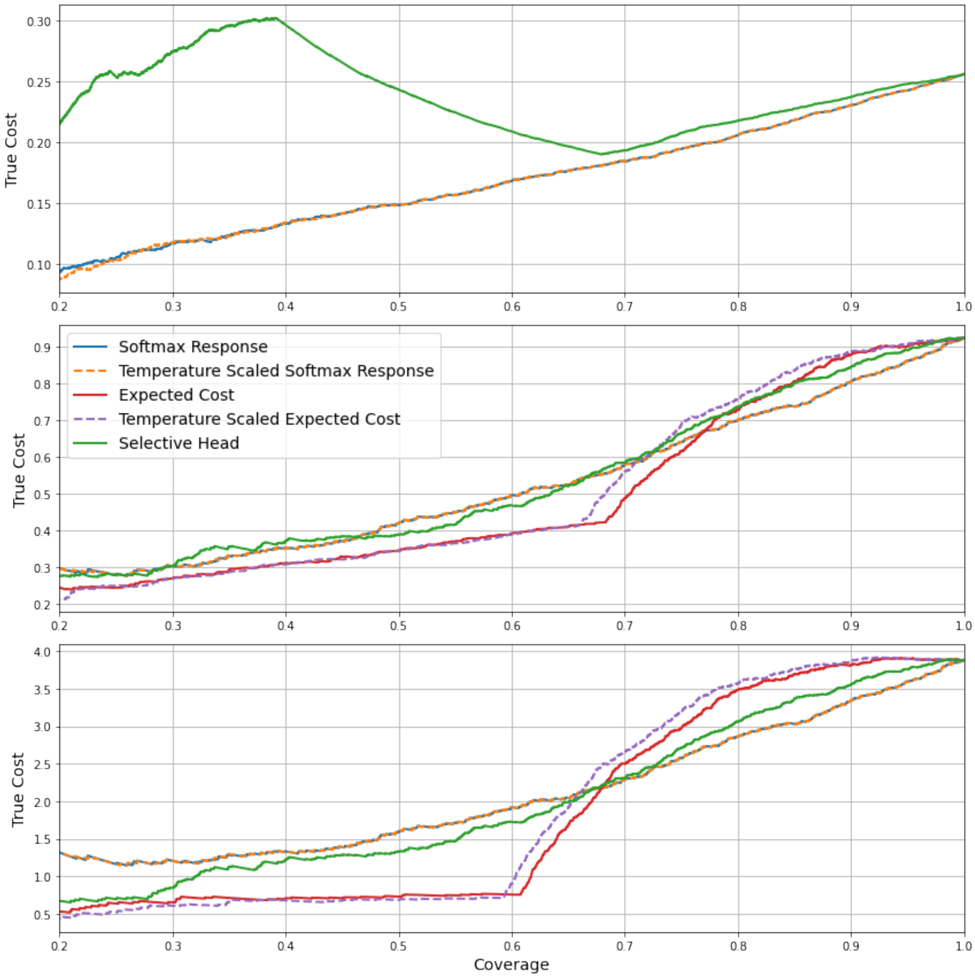
Experiments using a SelectiveNet model trained for a target coverage of 0.85, testing on the $S_{combined}$ distribution.



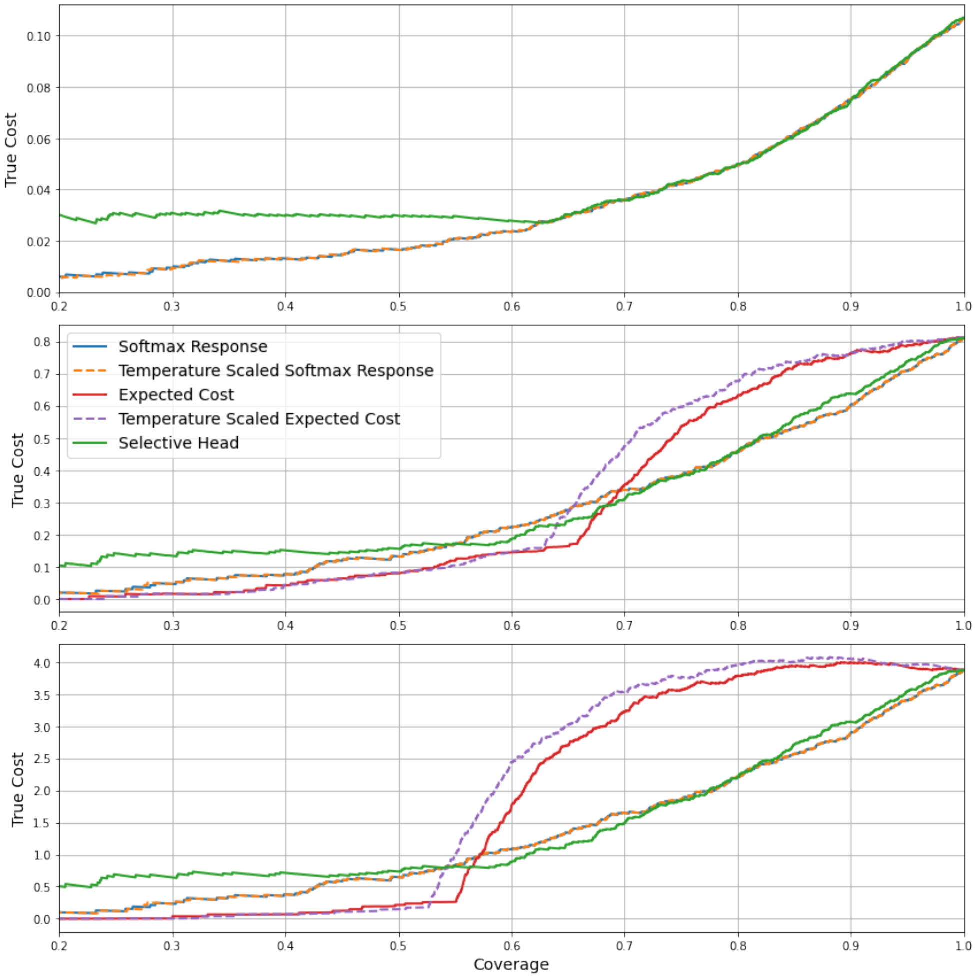
Experiments using a SelectiveNet model trained for a target coverage of 0.9, testing on the S_{in} distribution.



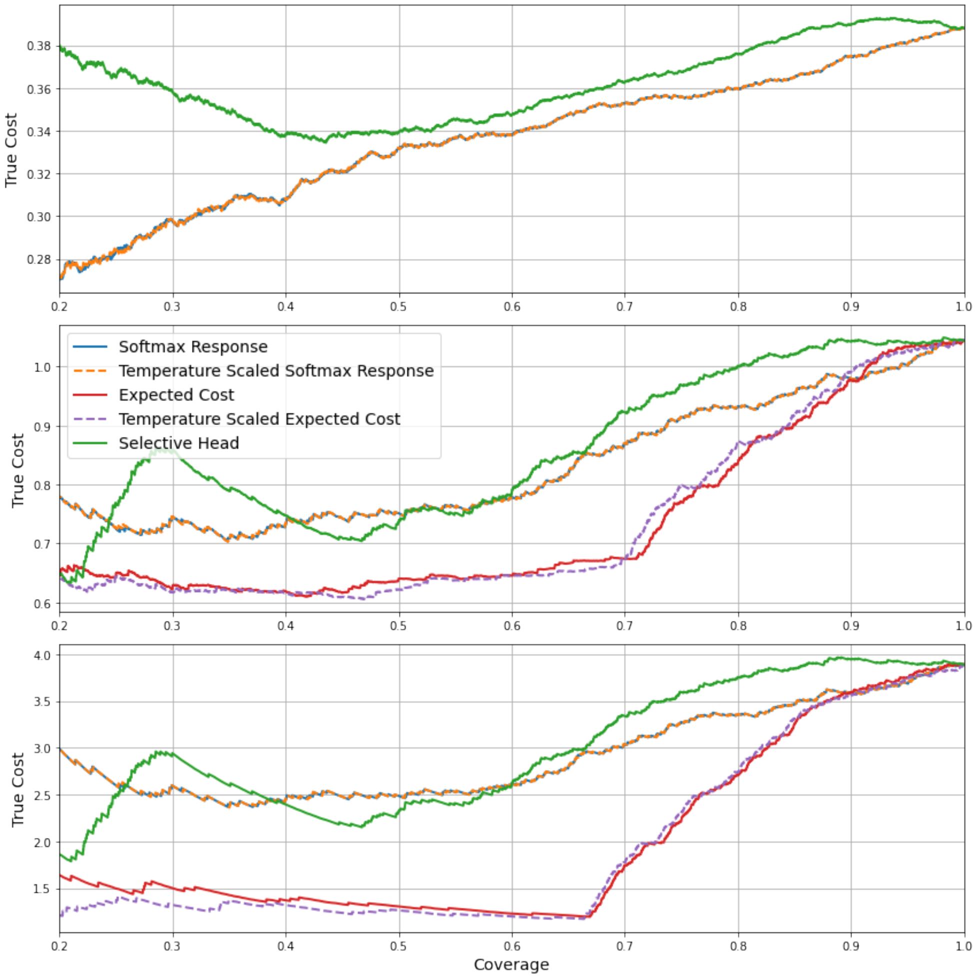
Experiments using a SelectiveNet model trained for a target coverage of 0.9, testing on the $S_{unknown}$ distribution.



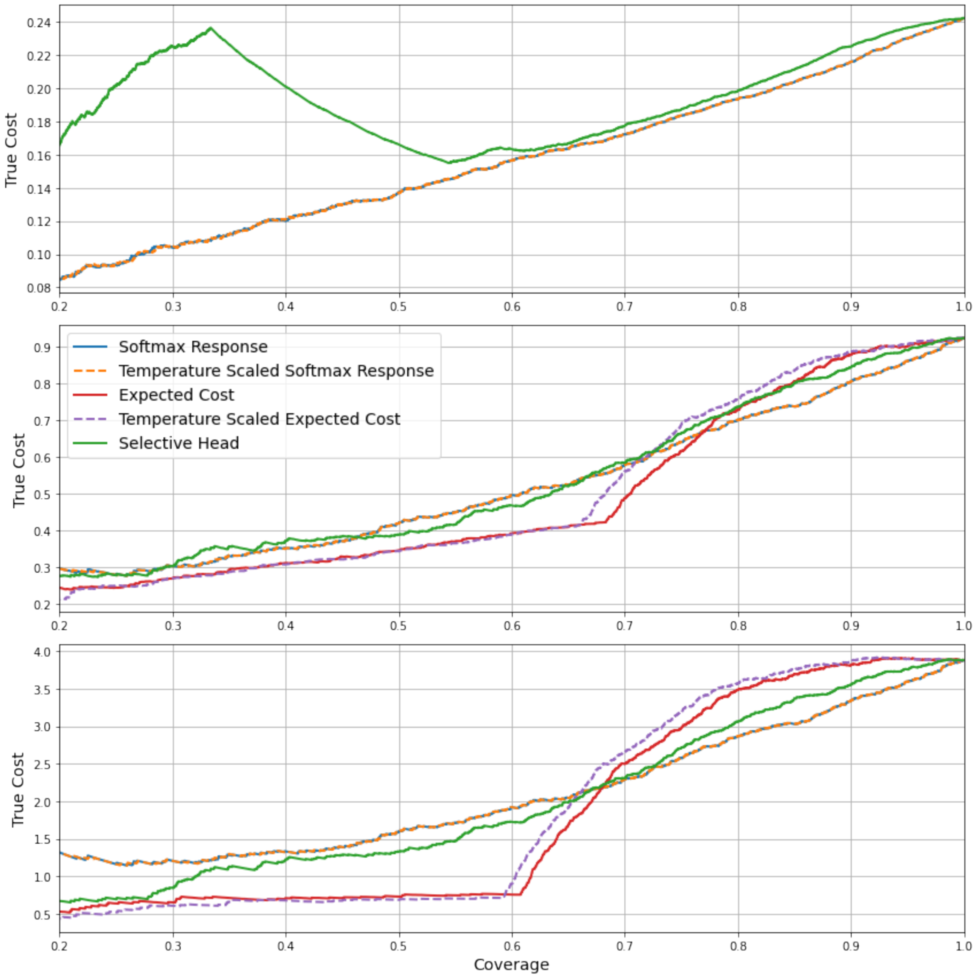
Experiments using a SelectiveNet model trained for a target coverage of 0.9, testing on the $S_{combined}$ distribution.



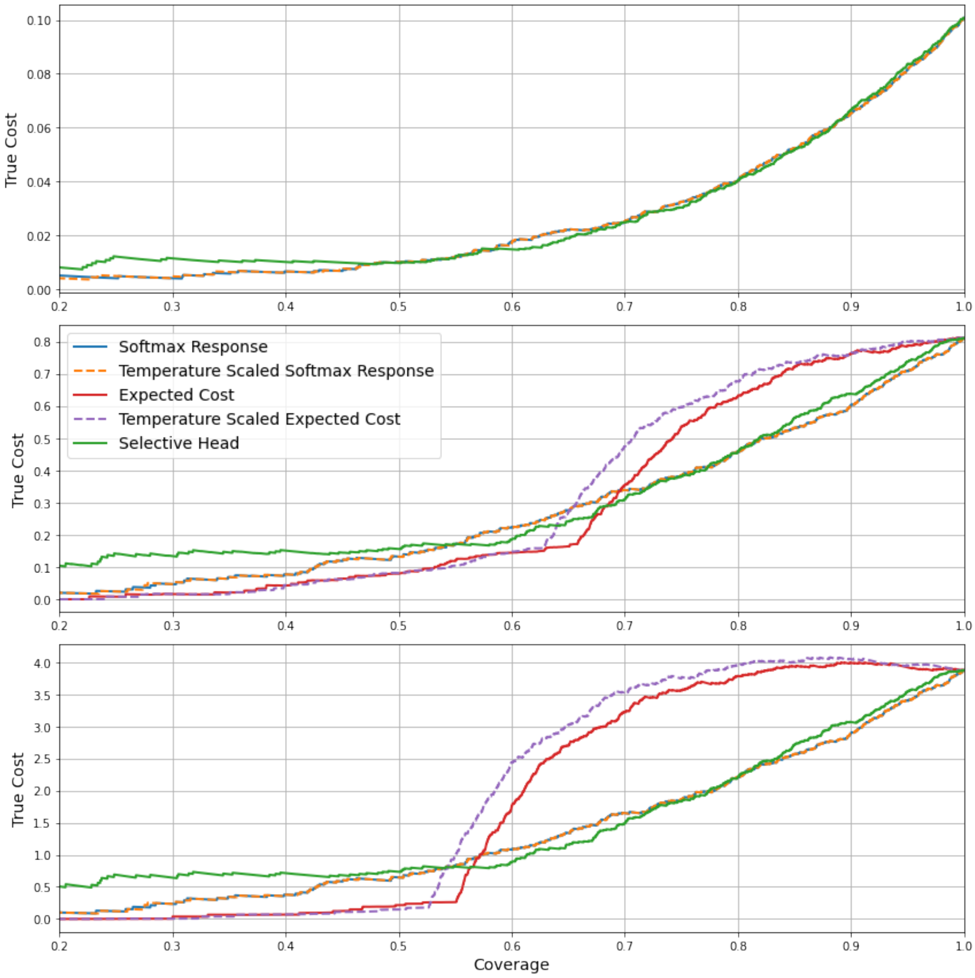
Experiments using a SelectiveNet model trained for a target coverage of 0.95, testing on the S_{in} distribution.



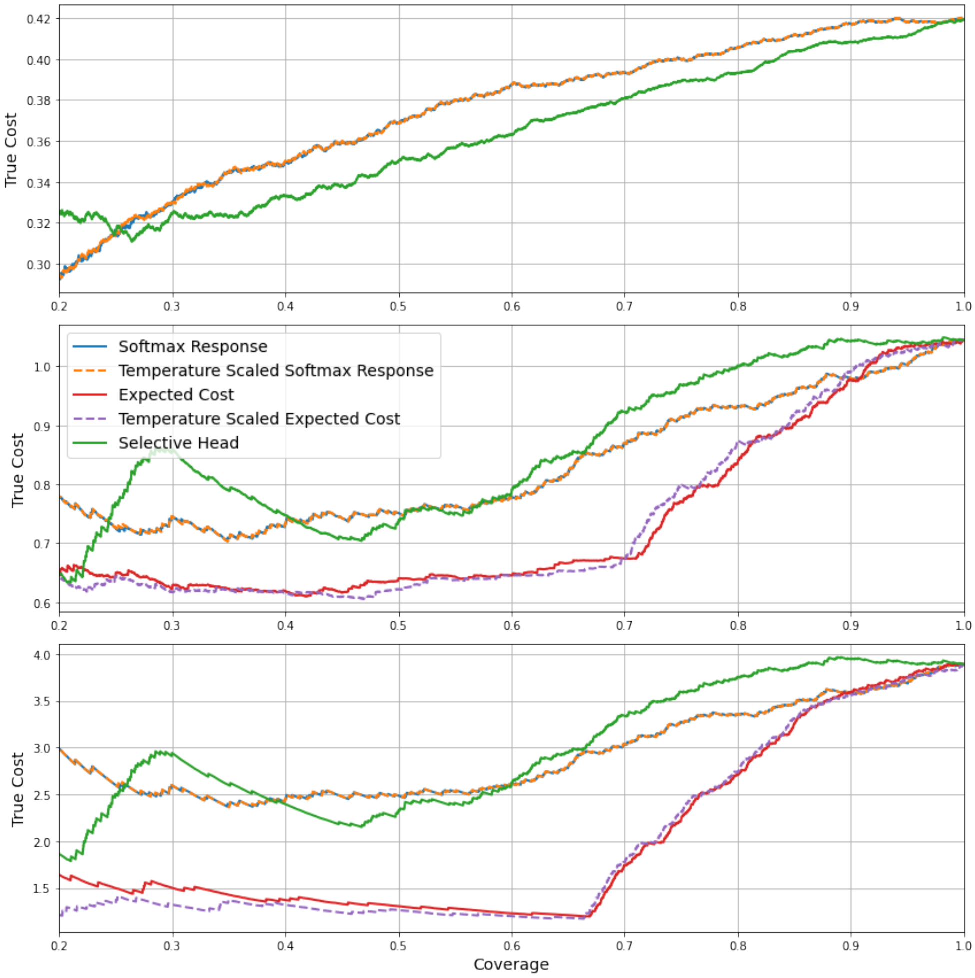
Experiments using a SelectiveNet model trained for a target coverage of 0.95, testing on the $S_{unknown}$ distribution.



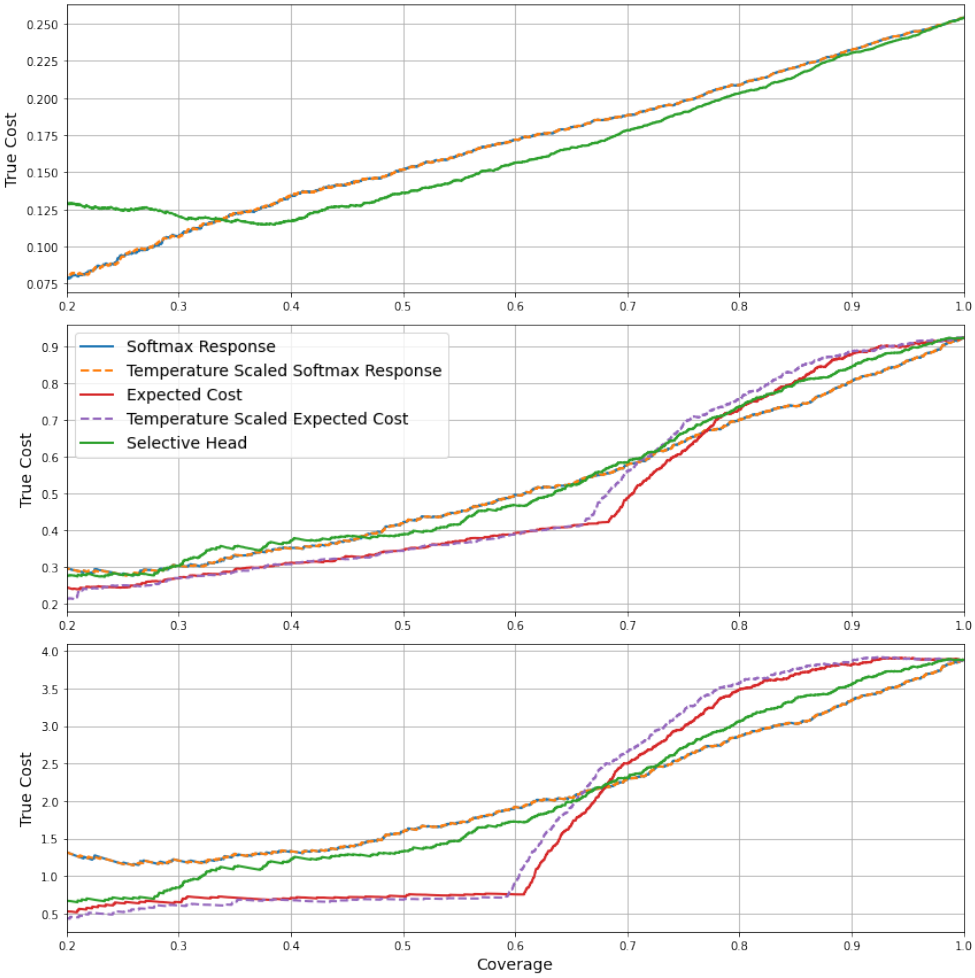
Experiments using a SelectiveNet model trained for a target coverage of 0.95, testing on the $S_{combined}$ distribution.



Experiments using a SelectiveNet model trained for a target coverage of 1.0, testing on the S_{in} distribution.

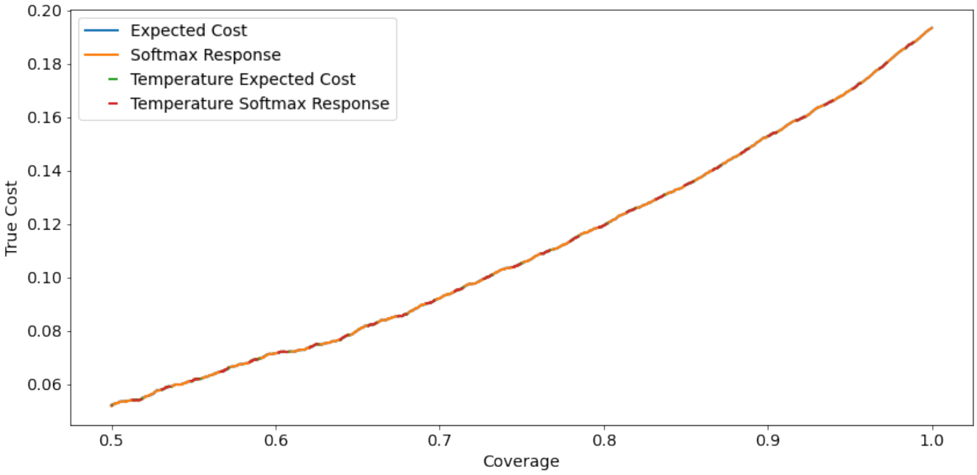


Experiments using a SelectiveNet model trained for a target coverage of 1.0, testing on the $S_{unknown}$ distribution.

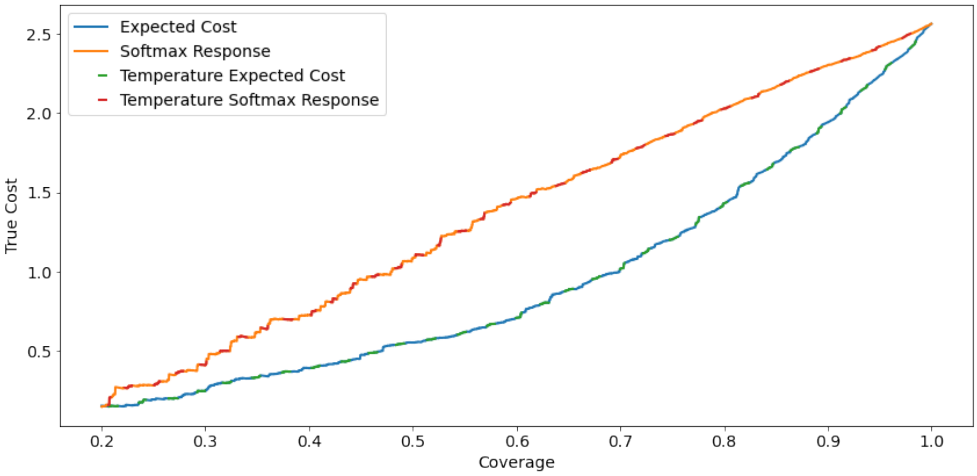


Experiments using a SelectiveNet model trained for a target coverage of 1.0, testing on the $S_{combined}$ distribution.

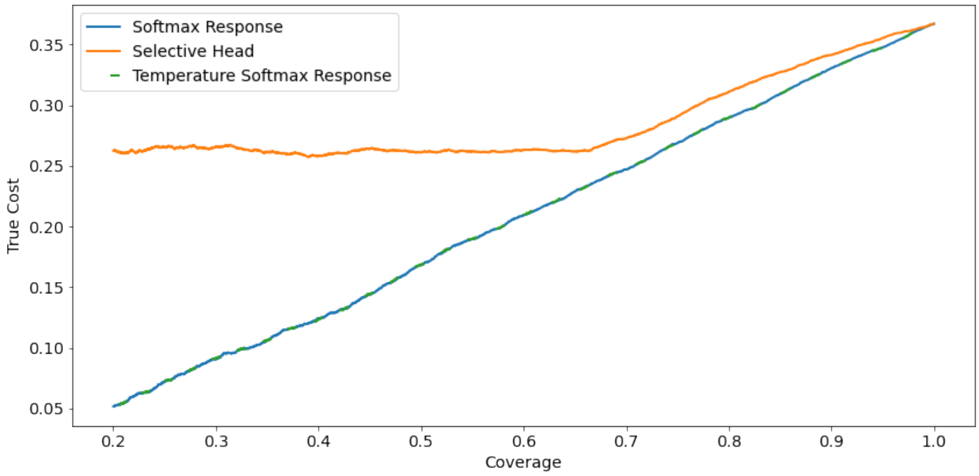
B.2 Multi-Class Classification Experiment Results



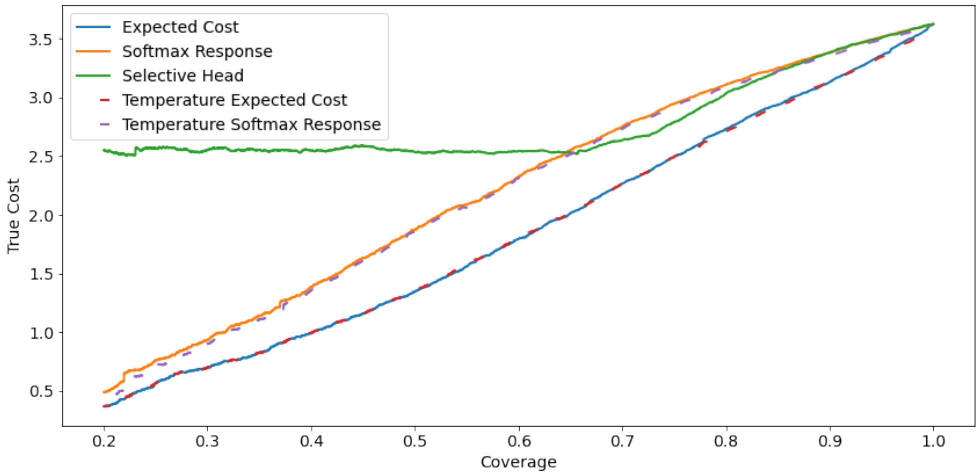
Experiments using a standard CNN with symmetrical costs.



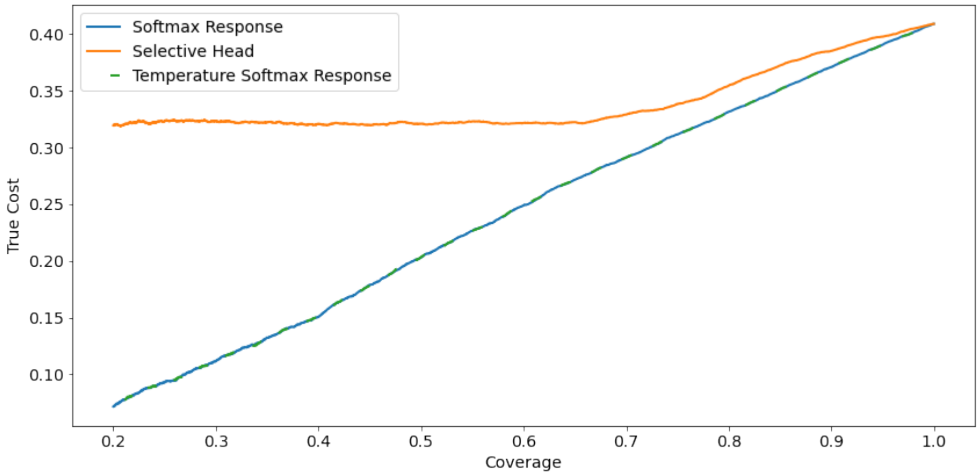
Experiments using a standard CNN with asymmetrical costs.



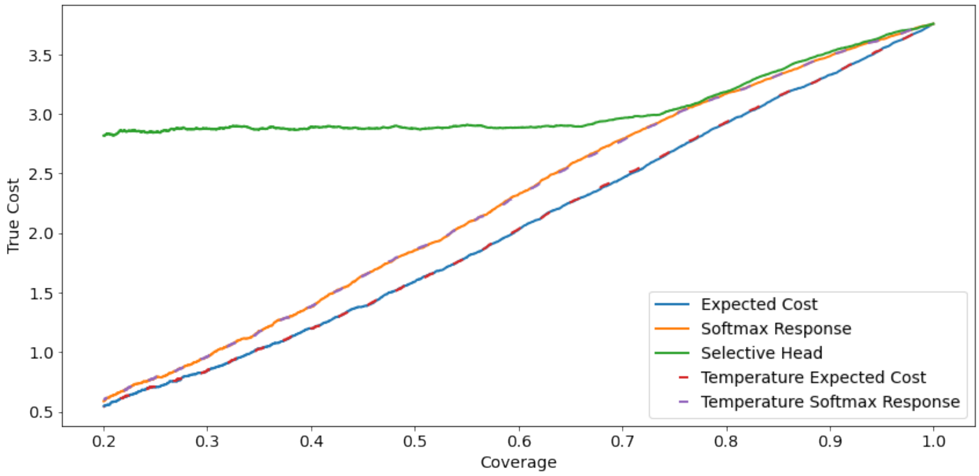
Experiments using a SelectiveNet model with a target coverage of 0.7 with symmetrical costs.



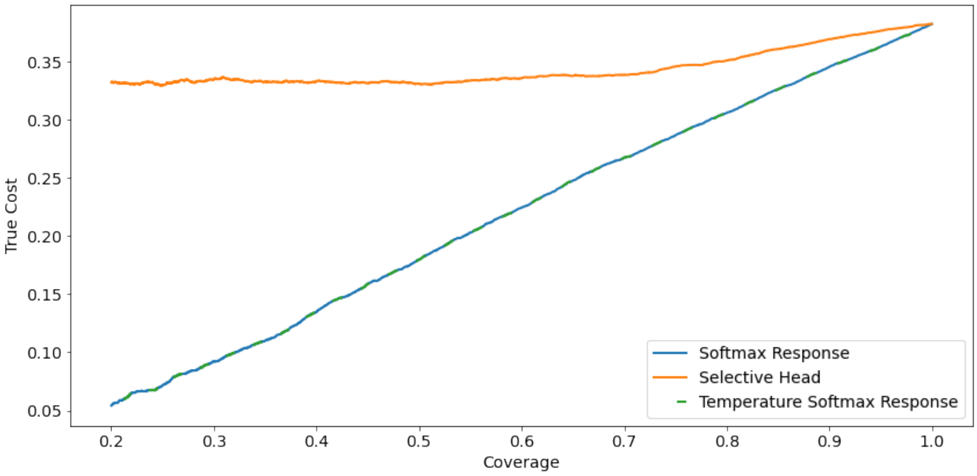
Experiments using a SelectiveNet model with a target coverage of 0.7 with asymmetrical costs.



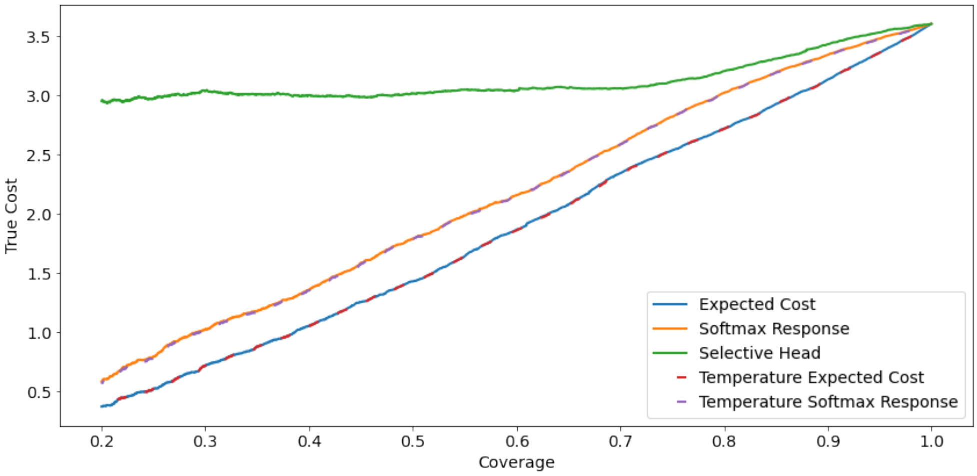
Experiments using a SelectiveNet model with a target coverage of 0.75 with symmetrical costs.



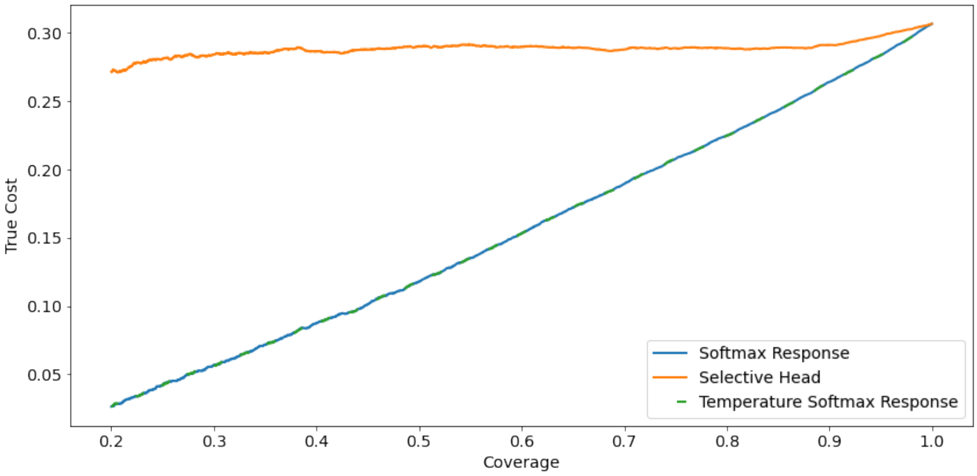
Experiments using a SelectiveNet model with a target coverage of 0.75 with asymmetrical costs.



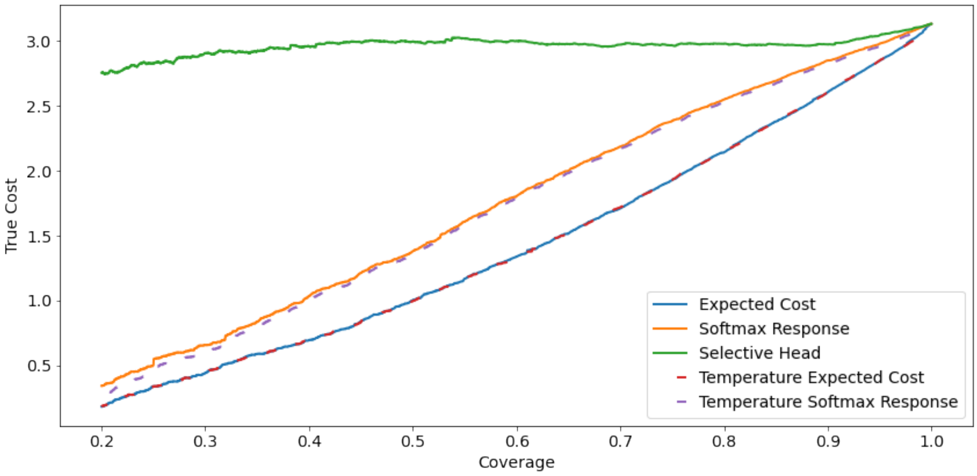
Experiments using a SelectiveNet model with a target coverage of 0.8 with symmetrical costs.



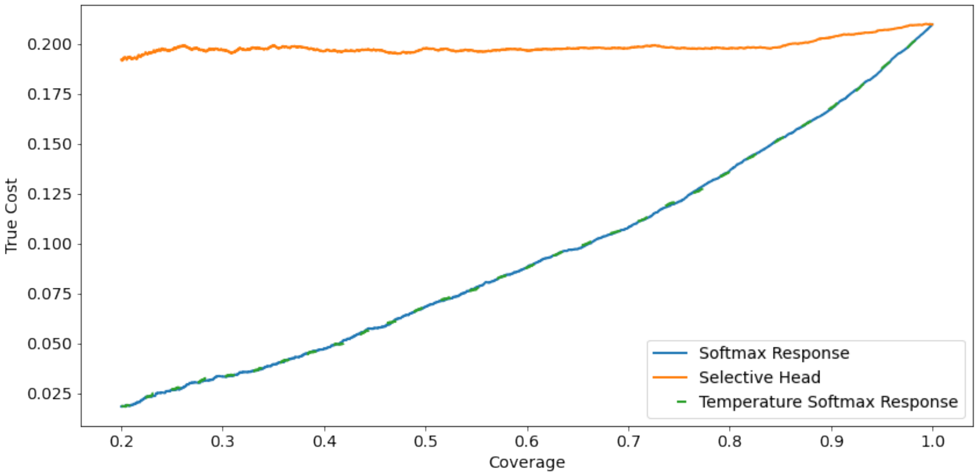
Experiments using a SelectiveNet model with a target coverage of 0.8 with asymmetrical costs.



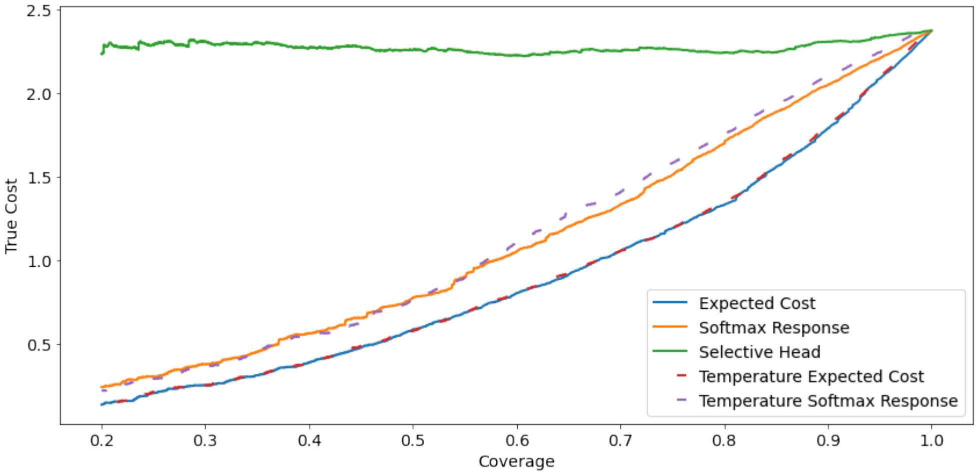
Experiments using a SelectiveNet model with a target coverage of 0.85 with symmetrical costs.



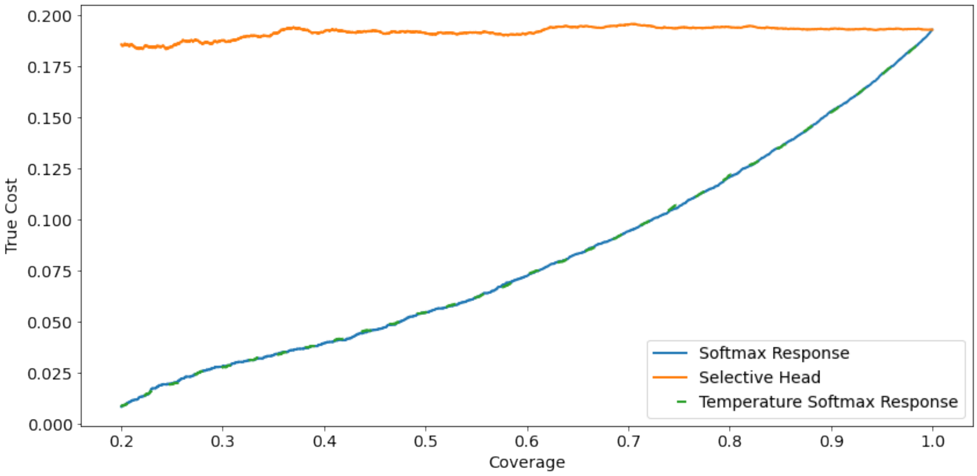
Experiments using a SelectiveNet model with a target coverage of 0.85 with asymmetrical costs.



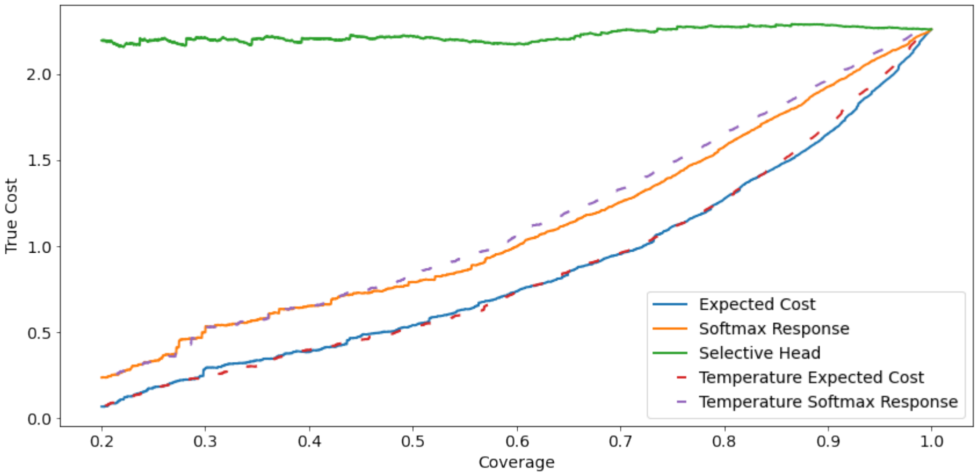
Experiments using a SelectiveNet model with a target coverage of 0.9 with symmetrical costs.



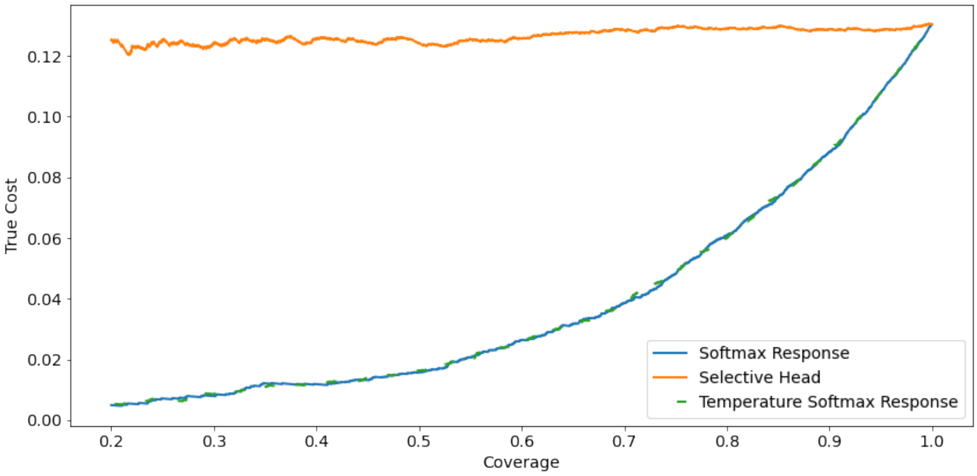
Experiments using a SelectiveNet model with a target coverage of 0.9 with asymmetrical costs.



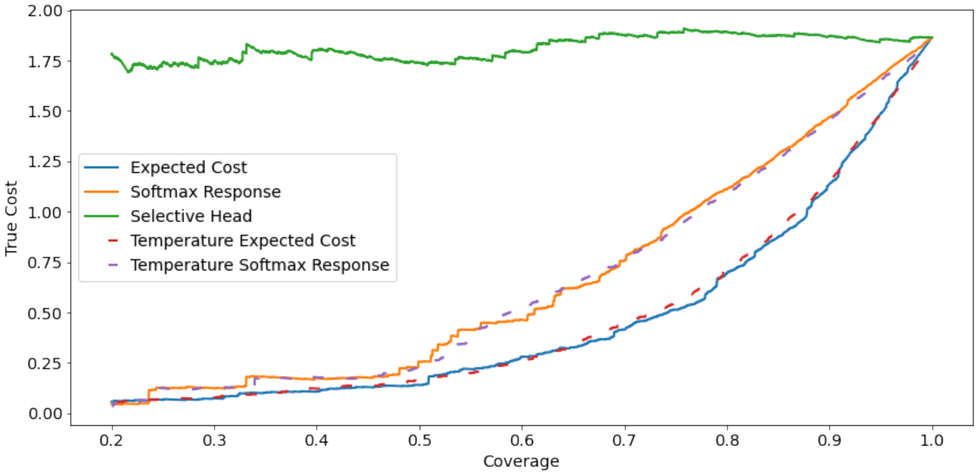
Experiments using a SelectiveNet model with a target coverage of 0.95 with symmetrical costs.



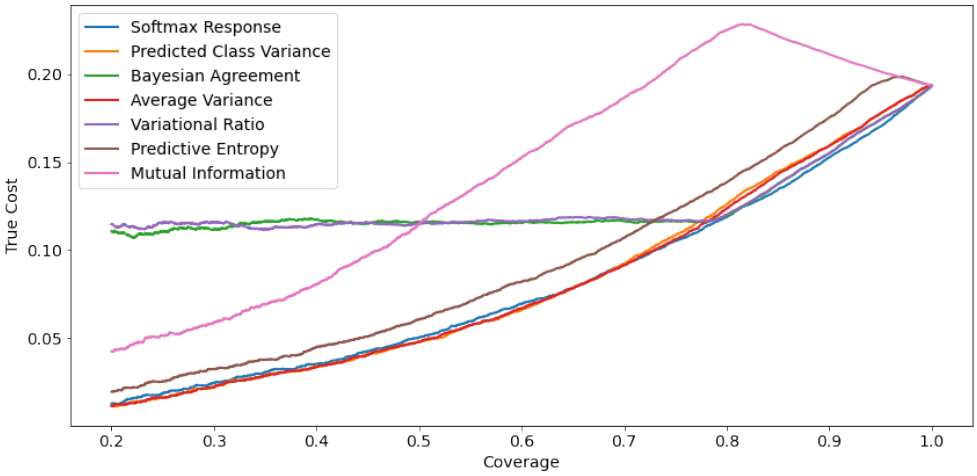
Experiments using a SelectiveNet model with a target coverage of 0.95 with asymmetrical costs.



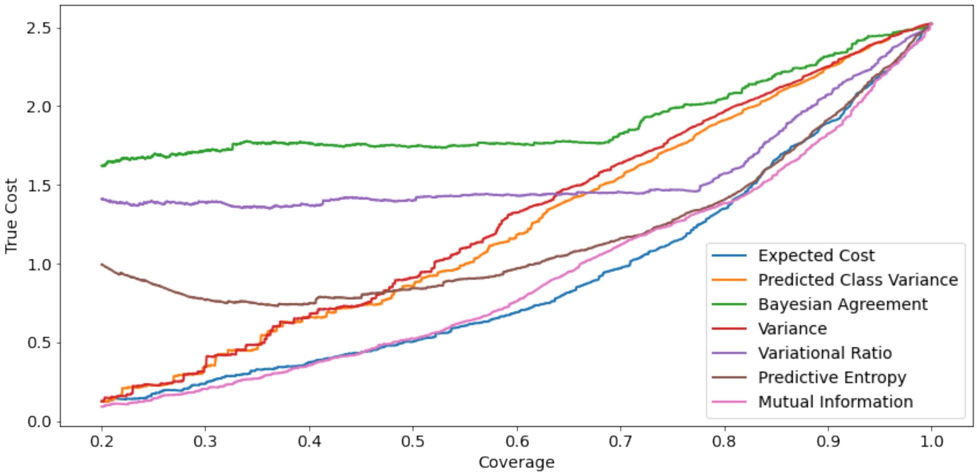
Experiments using a SelectiveNet model with a target coverage of 1.0 with symmetrical costs.



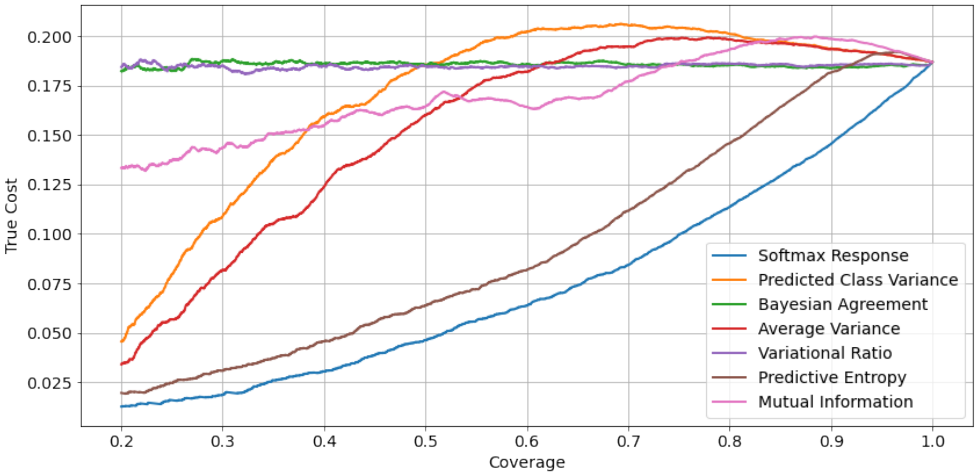
Experiments using a SelectiveNet model with a target coverage of 1.0 with asymmetrical costs.



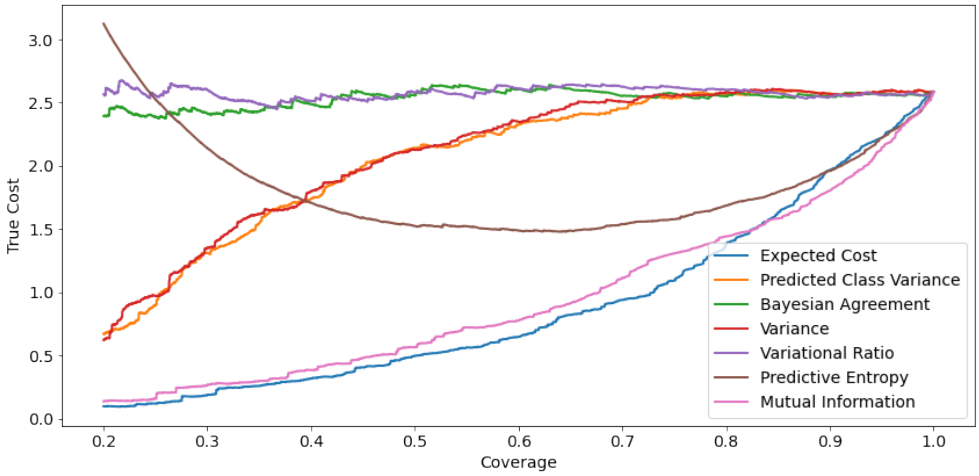
Experiments using a Bayesian model with Monte Carlo Dropout with symmetrical costs.



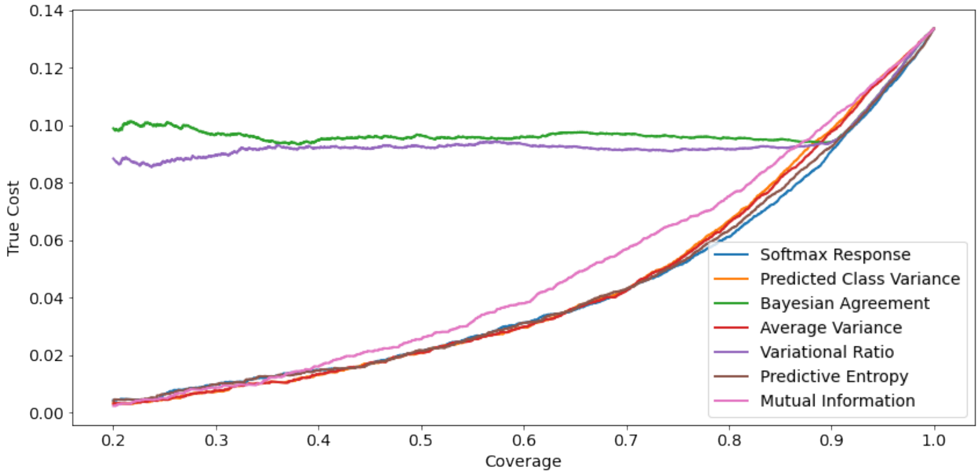
Experiments using a Bayesian model with Monte Carlo Dropout with asymmetrical costs.



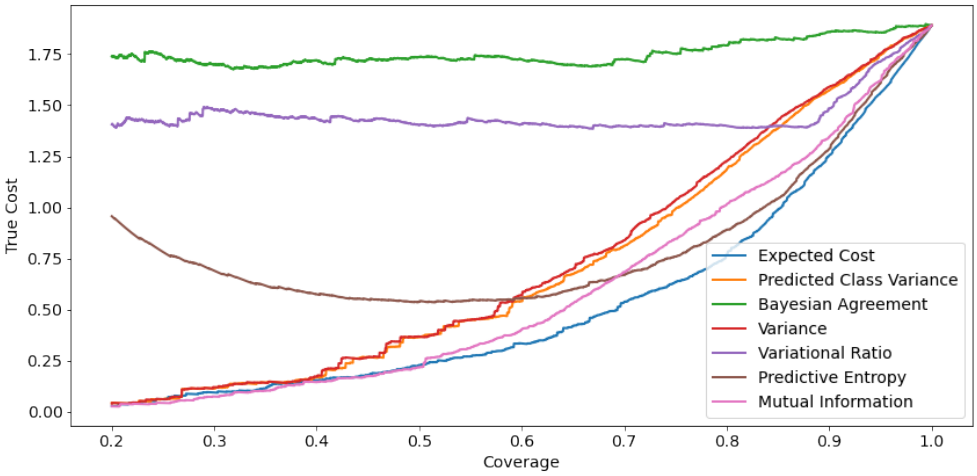
Experiments using a Bayesian model trained using Bayes-by-Backprop with asymmetrical costs.



Experiments using a Bayesian model trained using Bayes-by-Backprop with asymmetrical costs.



Experiments using a Bayesian model trained using Laplace Approximation with symmetrical costs.



Experiments using a Bayesian model trained using Laplace Approximation with asymmetrical costs.