BIRMINGHAM CITY UNIVERSITY

DOCTORAL THESIS

Medical Image Classification using Deep Learning Techniques and Uncertainty Quantification

Author:
Zakaria SENOUSY

Supervisor:
Dr. Mohammed ABDELSAMEA
Prof. Mohamed GABER

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

School of Computing and Digital Technology
Birmingham City University

February 23, 2023

Declaration of Authorship

I, Zakaria SENOUSY, declare that this thesis titled, "Medical Image Classification using Deep Learning Techniques and Uncertainty Quantification" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Abstract

The emergence of medical image analysis using deep learning techniques has introduced multiple challenges in terms of developing robust and trustworthy systems for automated grading and diagnosis. Several works have been presented to improve classification performance. However, these methods lack the diversity of capturing different levels of contextual information among image regions, strategies to present diversity in learning by using ensemble-based techniques, or uncertainty measures for predictions generated from automated systems. Consequently, the presented methods provide sub-optimal results which is not enough for clinical practice. To enhance classification performance and introduce trustworthiness, deep learning techniques and uncertainty quantification methods are required to provide diversity in contextual learning and the initial stage of explainability, respectively.

This thesis aims to explore and develop novel deep learning techniques escorted by uncertainty quantification for developing actionable automated grading and diagnosis systems. More specifically, the thesis provides the following three main contributions. First, it introduces a novel entropy-based elastic ensemble of Deep Convolutional Neural Networks (DCNNs) architecture termed as *3E-Net* for classifying grades of invasive breast carcinoma microscopic images. *3E-Net* is based on a patch-wise network for feature extraction and image-wise networks for final image classification and uses an elastic ensemble based on Shannon Entropy as an uncertainty quantification method for measuring the level of randomness in image predictions. As the second contribution, the thesis presents a novel multi-level context and uncertainty-aware deep learning architecture named *MCUa* for the classification of breast cancer microscopic images. *MCUa* consists of multiple feature extractors and multi-level context-aware models in a dynamic ensemble fashion to learn the spatial dependencies among image patches and enhance the learning diversity. Also, the architecture uses Monte Carlo (MC) dropout for measuring the uncertainty of image predictions and deciding whether an input image is accurate based on the generated uncertainty score. The third contribution of the thesis introduces a novel model agnostic method (*AUQantO*) that establishes an actionable strategy for optimising uncertainty quantification for deep learning architectures. *AUQantO* method works on optimising a hyperparameter threshold, which is compared against uncertainty scores from Shannon entropy and MC-dropout. The optimal threshold is achieved based on single- and multi-objective functions which are optimised using multiple optimisation methods.

A comprehensive set of experiments have been conducted using multiple medical imaging datasets and multiple novel evaluation metrics to prove the effectiveness of our three contributions to clinical practice. First, *3E-Net* versions achieved an accuracy of 96.15% and 99.50% on invasive breast carcinoma dataset. The second contribution, *MCUa*, achieved an accuracy of 98.11% on Breast cancer histology images dataset. Lastly, *AUQantO* showed significant improvements in performance of the state-of-the-art deep learning models with an average accuracy improvement of 1.76% and 2.02% on Breast cancer histology images dataset and an average accuracy improvement of 5.67% and 4.24% on Skin cancer dataset using two uncertainty quantification techniques. *AUQantO* demonstrated the ability to generate the optimal number of excluded images in a particular dataset.

Acknowledgements

The research presented in this thesis was funded by Birmingham City University as part of the scholars program and supported by many individuals.

First and foremost, I would like to praise and thank God Almighty, who has given me infinite grace, knowledge, and opportunity to complete my PhD study and this thesis.

I would like to convey my heartfelt thanks to my supervisory team, including my first supervisor, Dr. Mohammed Abdelsamea, and my second supervisor, Prof. Mohamed Medhat Gaber, for their unwavering support during my PhD study.

I would like to gratefully thank Dr. Mohammed Abdelsamea for his ideas, guidance, and continuous motivation. Without his guidance and support, I would not be here at this step of completing my PhD study. I could not have imagined having a better supervisor and director for my PhD study.

I am tremendously thankful to Prof. Mohamed Medhat Gaber for his insight, vast knowledge, ideas, and patience during my studies. I am part of this amazing journey because of his belief in me and his great assistance and support in bringing me to Birmingham City University as a PhD student.

I would like to share my special thanks to my wife, Maram, for her support and help during difficult times and throughout my research journey. She has a great contribution to keep things smooth and flexible during my PhD studies. I couldn't be more thankful for her kindness, assistance, and motivating me during this journey.

I am eternally thankful to my Mom and Dad, for being my strength in everything I do and the reason of me being at this stage in my life. Also, I would like to thank everyone else in my family and more especially my elder brother, Yehya, for his support, motivation, and advices which helped me in being a better person and encouraged me to accomplish my PhD studies. No words can truly express how grateful I am for sharing emotional and unforgettable moments with all of them.

Last but not least, I would like to thank my best friends for their massive support during my research journey and for making this journey enjoyable and unforgettable.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

CV	**C**omputer **V**ision
ANN	**A**rtificial **N**eural **N**etwork
DL	**D**eep **L**earning
DNN	**D**eep **N**eural **N**etwork
CNN	**C**onvolutional **N**eural **N**etwork
DCNN	**D**eep **C**onvolutional **N**eural **N**etwork
CAD	**C**omputer **A**ided **D**iagnosis
SVM	**S**upport **V**ector **M**achine
MLP	**M**ulti **L**ayer **P**erceptron
FF-DNN	**F**eed **F**orward **D**eep **N**eural **N**etwork
FCN	**F**ully **C**onvolutional **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
LSTM	**L**ong **S**hort **T**erm **M**emory
ReLU	**Re**ctified **L**inear **U**nit
WSI	**W**hole **S**lide **I**mage
CNB	**C**ore **N**eedle **B**iopsy-tissue
H&E	**H**ematoxylin (&) **E**osin
MRI	**M**agnetic **R**esonance **I**maging
3E-Net	**E**ntropy-based **E**lastic **E**nsemble **Net**work
MC	**M**onte **C**arlo
MCUa	**M**ulti-level **C**ontext (&) **U**ncertainty **a**ware
AUQantO	**A**ctionable **U**ncertainty **Q**u**ant**ification **O**ptimisation
GP	**G**aussian **P**rocesses
COBYLA	**C**onstrained **O**ptimisation **BY** **L**inear **A**pproximation
NSGA	**N**on-dominated **S**orting **G**enetic **A**lgorithm

Dedicated to the soul of my beloved mother whose unconditional love, affection, support, and prayers were my strength in everything I do.

Chapter 1

Introduction

1.1 Overview

Computer Vision (CV) is a field that aims to develop strategies to assist computers in seeing and understanding the content of visual data, such as photographs and videos. CV seeks to comprehend the content of digital images. Generally, this entails the creation of methods that seek to replicate human vision systems. Extracting a description from a digital image, which could be some objects in a scene, is one way to understand the content of the image(s) [14]. CV is a multidisciplinary field that includes elements from image processing, pattern recognition, and artificial intelligence. It commonly uses a combination of specialised techniques and general learning algorithms to analyse and comprehend visual data.

Machine learning is the use of computer systems that can learn and adapt without explicit instructions or guidelines. This is done by using algorithms and statistical models to analyse and predict data in an input [13]. Generally speaking, machine learning methods are categorised as supervised learning and unsupervised learning. For supervised learning, the machine learning model uses an algorithm that learns to map between an input variable x and an output variable y. This is based on a training dataset in which the algorithm learns the correct labels of the training samples (where labels of input samples are available). Unsupervised learning is where a machine learning model learns and distinguishes patterns from unlabelled samples (where labels of input samples are not available). The other class of machine learning techniques uses semi-supervised learning, which can deal with a limited number of labelled samples [87].

Image classification is considered as one of the fundamental tasks for CV escorted by supervised learning technique where unstructured data (e.g. images) are categorised into predefined classes (i.e. labels) that are available during the training process. Traditionally, image classification approaches were based on manually engineered features, which require a high level of domain knowledge while demonstrating poor cross-domain adaptability. In recent years, deep learning has been employed to address real-world problems. Deep learning is a type of machine learning approach in which several stages of non-linear information processing in hierarchical structures are used to classify patterns and learn features [25]. Deep learning uses Artificial Neural Network (ANN), which is an attempt to mimic the human brain, to learn complex features and patterns from input data.

Deep Convolutional Neural Network (DCNN) is one of the deep learning approaches that mainly specialise in processing inputs of unstructured data (e.g. images and videos). A DCNN consists of an input layer, hidden layers, and an output layer. Hidden layers of DCNN contain: (1) convolutional layers which apply the mathematical convolution operation using image kernels (filters) applied to all regions in an image. This operation results in the generation of a feature map that

captures salient features from the input image; and (2) Pooling layers which are mainly used to decrease the variance and computational complexity by reducing dimensions of features data [32].

In the medical sector, cancer is a serious health concern that affects people around the world. According to the American Cancer Society, by the end of 2021, there was an estimate of 1.8 million new cancer cases and 608,570 cancer deaths that were projected to occur [99]. Despite significant improvements in medical research, the analysis of digital medical images remains the most widely employed sector for identifying various forms of diseases. Early diagnosis can vastly enhance the efficacy of treatment. One of the critical issues that affect the medical diagnosis process is the difficulty of manual investigation as there could be some complex samples that are challenging to analyse. Furthermore, manual investigation is considered a time-consuming process.

Advances in computer-aided diagnosis (CAD) have a significant influence on improving disease detection accuracy and lowering the amount of time medical practitioners spend on manual investigations. Deep learning is one of the approaches that has been widely employed for automated diagnosis. Deep learning algorithms have made significant progress and produced outstanding results, urging many academics to develop fair and automated solutions for a variety of medical image analysis applications. In contrast to the traditional image classification approach which relies mainly on handcrafting features, deep learning leverages the power of using DCNNs for the target of learning multi-level representations and patterns from unstructured data using numerous linear and non-linear layers. DCNNs can explore extensive statistical data structures without the need for handcrafted characteristics, making them useful in a variety of fields.

1.2 Motivation

Considerable work for medical image analysis using deep learning approaches has been proposed [65]. DCNNs showed high performance in terms of automated diagnosis. However, digitised medical images bring special challenges different from large-scale images. One of the challenges in dealing with medical images (such as histopathology images) is the high similarity of the morphological structures of multiple classes. Morphological structure in biology indicates the study of the shape, size, and structure of microorganisms and the relationship between their different components. In breast cancer histopathology images (which are digitised images generated based on the microscopic examination of biological tissues of the breast), the morphological structures in terms of nuclei distribution between benign and carcinoma images are too similar. This issue is challenging, as it introduces inter-class similarity, making it difficult for automated diagnosis systems based on DCNNs to establish and distinguish different class boundaries.

Another challenge appears for medical images is the intra-class fluctuations where the variation of image structures within particular class is very high. For instance, for a certain class, we can have samples showing different morphological structures (in case of histopathology images). This adds more difficulty for the automated diagnosis systems to clearly classify images and provide high level of confidence for the generated prediction. Based on the two challenges described above, a robust automated diagnosis system that provides different learning perspectives and a measure of classification confidence is required to handle medical imaging challenges.

Besides the challenges of medical images stated above, another problem occurs when we deal with a single DCNN for automated diagnosis. The ability of a single DCNN model to obtain discriminatory features is limited and usually results in sub-optimal solutions [73]. This is due to the lack of different learning perspectives for image features, which may aid in boosting the accuracy of diagnosis. As a result, an ensemble of DCNN models for automated diagnosis systems is required to preserve image description from recognised views to more exact prediction [120]. Ensemble learning is a machine learning technique that combines the predictions of separate classifiers using a combination strategy to give a more accurate final prediction of the input sample than a single classifier.

More crucially, previously proposed DCNN-based automated diagnosis systems, to the best of our knowledge, lacked a predefined measure of uncertainty, which is a critical first step toward an explainable medical image analysis. Developing an uncertainty measure component can help to identify several areas of ambiguity that could be therapeutically useful. It also allows pathologists and medical practitioners to prioritise images for annotations by rating them.

All the above-mentioned challenges motivate the need for robust and reliable automated diagnosis systems that are based on DCNNs and uncertainty measures.

1.3 Problem Statement

When building sturdy and trustworthy automated diagnosis systems, the following essential elements should be taken into account.

- **Context-awareness**: One of the important features that is crucial to build a high-performance automated diagnosis system is to introduce contextual learning for different image regions. Contextual feature information refers to the relationships, interactions, and dependencies between different elements of an image, and this information helps to understand the image content in a broader sense, including objects, their relationships, and the semantic meaning behind them. Medical images are usually high-resolution images that need to be divided into small patches (tiles) to be processed by a deep learning model. The process of sectioning images into small patches allows only learning of local representations of image patches making no place for structural and contextual information to be learned by deep learning models [35, 73, 98, 109]. Contextual information is vital for building spatial dependencies between different image patches [9, 27, 42, 119]. This learning information builds a better vision for different image regions and their context and yields higher diagnosis performance for automated systems. Figure 1.1 presents an example of how context learning is conceptualised to build spatial dependencies among different image patches.

- **Uncertainty measure**: The trustworthiness of an automated system is measured by its ability to be confident in image predictions. Uncertainty measure is another important feature that needs to be introduced for systems that mainly deal with medical images [33, 72, 79]. The sensitivity of the medical field in terms of the importance of the existence of reliable decisions for patient disease necessitates the existence of a system that introduces a measure of uncertainty for image predictions. This kind of measure aids in introducing an initial stage of explainability. Explainability is a way of stating or explaining the reason behind giving a particular diagnosis for a patient's digitised sample.

- **High accuracy and robustness**: One of the main targets of automated diagnosis systems (medical image classification models) is to reach a high level of performance and improve classification accuracy. One of the issues resulting in sub-optimal performance by deep learning models is the lack of learning diversity and variety of feature extraction from different perspectives. Ensemble learning is one of the techniques used to improve system performance [20, 46, 74, 120]. In our work, we developed a novel dynamic ensemble strategy for building an automated system based on an ensemble of DCNNs which utilise different learning perspectives. The concept of dynamic ensemble lies in the partial use of learners (classifiers) towards the final classification instead of using all classifiers to contribute to the final prediction. This depends mainly on both the input sample and its prediction confidence. Figure 1.2 presents an example of an ensemble model which applies the idea of a dynamic ensemble where only the accurate models are chosen for final image classification.

- **Actionability**: Creating an automated diagnosis system to take decisions based on the certainty of an input image is crucial, especially for medical image analysis and clinical practice [5, 70]. This step identify images that have high uncertainty to be investigated by medical professionals. This means that an automated system takes the decision whether to give a final classification for a particular image based on a high confidence level of output prediction or to avoid classifying the image (exclude image) based on a low confidence score of the output prediction. Figure 1.3 presents a workflow pipeline for actionability and how can be used to take decisions and introduce level of explainability.

Accordingly, a deep learning model for automated diagnosis can ideally be introduced based on a dynamic ensemble learning strategy, a context-awareness approach that enhances diagnosis accuracy, and an uncertainty-aware component which measures the quality of image predictions.

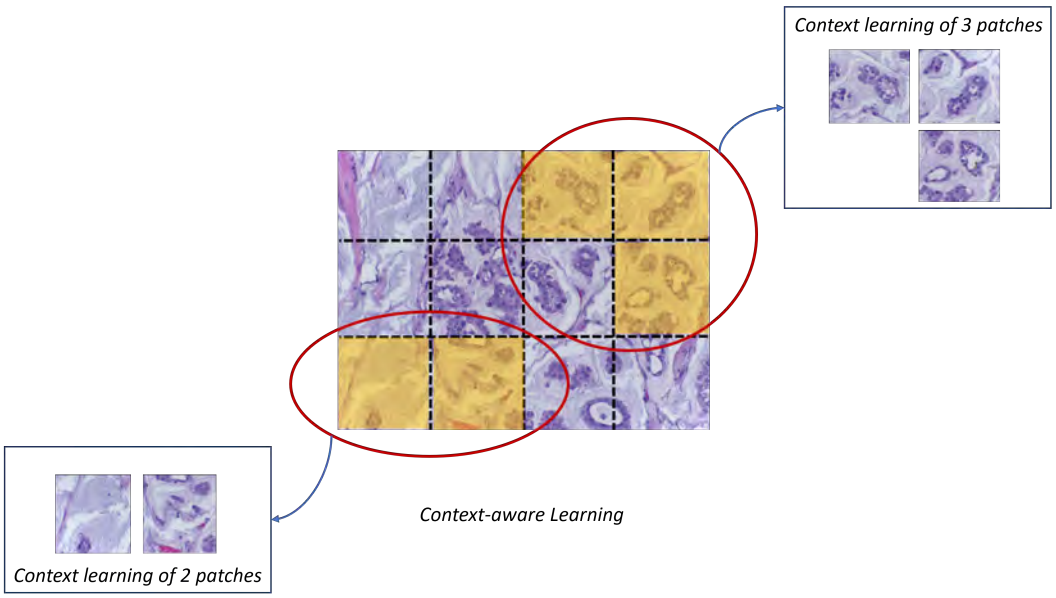


FIGURE 1.1: An example of context-aware learning. The figure illustrates the concept of context learning applied to a histopathology image divided into 12 patches. As stated above, most deep learning models learn local information of image patches without taking into consideration the spatial dependencies of these patches. This type of dependency is essential for enhancing the learning process conducted by deep learning models, and hence improving diagnosis performance. Our conceptualisation for context-aware learning includes learning multi-level combinations of patches. For instance, on the left side of the figure, context learning can be introduced between two patches of the bottom left corner of the image (Low-level context learning). In addition, a context learning of 3 patches can be applied to the 3 patches in the top right corner of the image. The context learning level could be designed by also including all 12 patches of the image (high-level context learning).

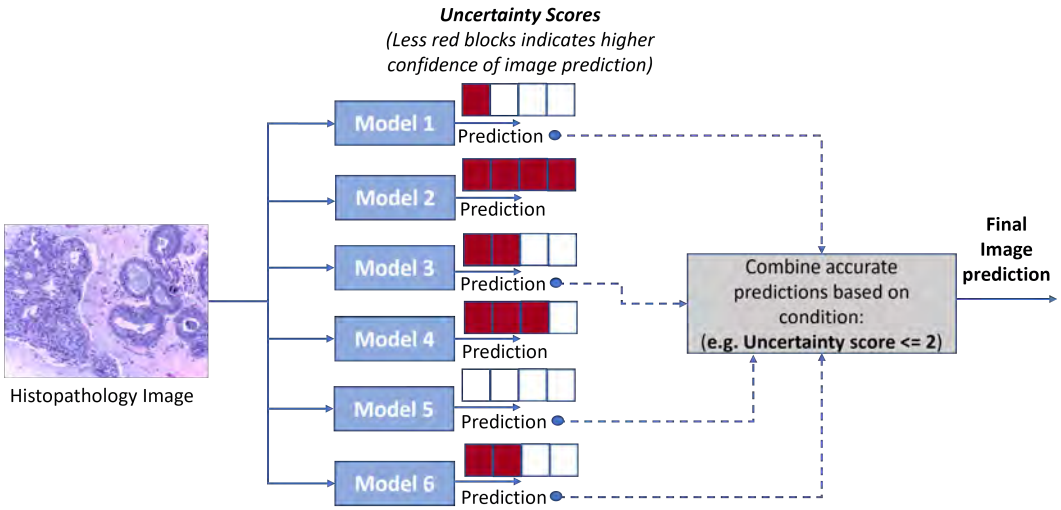


FIGURE 1.2: An example of an ensemble model consisting of 6 models which takes an input image and applies dynamic ensemble strategy. As shown in the figure, each model produces image prediction which is then used to identify the uncertainty score. Based on a predefined condition or threshold, models with low uncertainty scores are used for the final combination stage. Otherwise, the models are neglected. For simplification and clarification, we used a bar of red blocks to indicate the level of uncertainty in each model's prediction, and the models with an uncertainty score of less than or equal 2 are considered accurate models, which generate accurate predictions. For instance, Model 5 is considered a certain model, while Model 2 is considered a highly uncertain model. The accurate models' predictions are aggregated to produce the final image prediction. The dynamic process of model selection depends mainly on the input image and the measure of uncertainty introduced by each model in the ensemble architecture. In other words, the number of selected models in the combination stage varies depending on the input image and its uncertainty level.

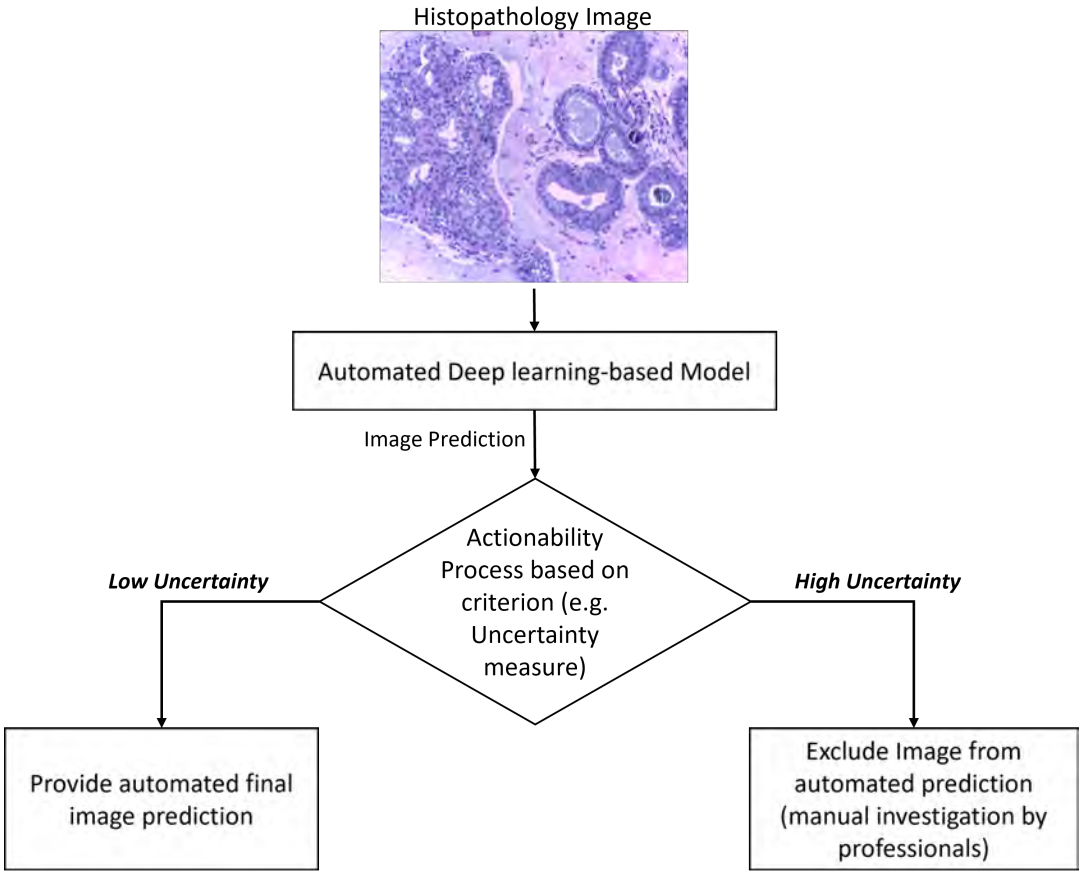


FIGURE 1.3: Overview of the actionability workflow. It is important to process predictions derived from an automated deep learning-based model to introduce actionability for medical images. For illustration, the figure depicts the workflow for an automated deep learning-based model that receives a digitised image as an input and generates image prediction (e.g. probability distribution including probability values for each class label). The generated image prediction is then used by an actionability stage, which serves as an active learner and takes the action of producing automated image prediction or excluding image from classification and returning it to a medical professional for further manual investigation. The action taken by the actionability stage depends on a particular criterion or condition.

1.4 Aim and Objectives

The aim of this thesis is to provide medical image classification systems that have design characterises of context-aware learning, trustworthiness feature using uncertainty quantification, generalisation and robustness using a novel dynamic ensemble strategy, and automated actionability by optimising uncertainty quantification.

The objectives of this research work have been established to fulfil the above aim. The objectives of this thesis are to:

1. develop a novel contextual information stage for automated diagnosis systems that improves diagnosis performance;
2. develop different uncertainty quantification techniques for measuring the level of uncertainty and randomness of input samples;
3. introduce generalisation and robustness by developing a novel ensemble learning strategy that combines different learning perspectives for automated diagnosis systems and provides an exclusion mechanism to exclude poor input samples that require investigation by medical professionals; and
4. develop a novel automated actionable technique for optimising uncertainty quantification that can be used by any deep learning model for the classification task.

1.5 Contributions

This thesis presents three contributions to achieving the above-mentioned objectives by introducing automated classification models of histopathological microscopic images accompanied by uncertainty measurements and an actionable model-agnostic method to optimise uncertainty quantification. The contributions of this thesis can be described as follows:

- **Entropy-based elastic ensemble of DCNNs model (*3E-Net*):** This model has been designed, developed, and implemented for the automated classification of invasive breast carcinoma histopathology samples into different grades. *3E-Net* introduces different image-wise learning perspectives in an ensemble technique and provides a measure for uncertainty quantification based on the prediction extracted from the input medical sample. The model introduces an exclusion mechanism which is beneficial for dealing with samples that are uncertain from the perspective of the deep learning model. The model starts by taking an input microscopic image and then divides the image into small patches. The extracted patches are then passed to a patch-wise network that acts as a feature extractor for extracting useful features of the input patches. The feature maps extracted from the patch-wise network are then used by a number of image-wise networks which are mainly developed to learn spatial dependencies between different image patches. Each prediction extracted from an image-wise network is inserted into an uncertainty quantification component based on Shannon Entropy [93]. This component generates an uncertainty value, which is mainly used to decide whether the image is certain. Having a low uncertainty score means that the model is accurate in the prediction of an input image, while a high uncertainty score means that the image is uncertain from the model prediction perspective. To introduce an ensemble of

all models, an elastic ensemble strategy is used that only takes into account certain models in the final image prediction [90].

- **Multi-level contextual and uncertainty aware model (*MCUa*):** This model has been designed, developed, and implemented for automated classification of breast cancer histopathology microscopic images. The model learns multi-level contextual information among feature maps extracted from patches of input sample and applies an intelligent mechanism to measure the level of ambiguity of image classification by implementing the Monte-Carlo dropout [29] technique. The usage of Monte-Carlo dropout results in different versions of image prediction, which aids in identifying variations in the list of image predictions and as a result helps in introducing a measure of uncertainty. The model works by taking an input image which is resized into multiple scales. The multiple image scales are then divided into smaller patches, which are passed into multiple feature extractors to learn salient features from different learning perspectives. The extracted feature maps from feature extractors are then inserted into multi-level context-aware networks which learn different levels of contextual information between image feature maps. The lowest level learns contextual information between multiple pairs of feature maps (two feature maps), while the highest level learns contextual information among all feature maps in a particular image. Finally, the Monte-Carlo method is applied to measure the uncertainty of the model's prediction where it is applied to the context-aware networks by activating the dropout layers during the testing phase of the architecture. This process aids in generating a list of predictions with some variations based on the randomness of dropping some of the neurons in context-aware networks. The standard deviation of the list of predictions is used as a measure of uncertainty to identify the level of randomness introduced in the image prediction. Based on the uncertainty measures introduced by the Monte-Carlo for all context-aware networks, a dynamic ensemble technique is utilised to identify the confident models to combine them towards the final image prediction. This is done using a predefined threshold that decides which models to be selected for final image predictions. If a particular image has no certain model from the group of all models in the ensemble architecture, then this image is excluded from the final image classification and returned to medical professionals for further investigation [91].

- **Actionable uncertainty quantification optimisation in deep learning architectures (*AUQantO*):** This method has been designed, developed, and implemented for optimising uncertainty quantification of deep learning models for medical image classification. The method relies on using different uncertainty quantification methods accompanied by the use of optimisation methods to minimise single- and multi-objective functions. The objective functions aim to minimise the number of excluded images that are highly uncertain and in need of manual investigation by medical professionals. The method is model- and dataset-agnostic, which means that it is a pluggable component that can be used by any deep learning model that generates probability distribution (predictions). The method starts with taking an input image into a deep learning model for image classification. Then, using two uncertainty quantification methods (Shannon Entropy and Monte-Carlo), the uncertainty measurement is calculated. The uncertainty value is then checked against a threshold value which is optimised using single- and multi-objective functions that mainly minimise the number of excluded images in a dataset [89].

1.6 Publications

The following papers have been either submitted or published towards building the work conducted in this thesis:

- **Senousy Z.**, Abdelsamea M., Mohamed M. M., and Gaber M. M., 3E-Net: Entropy-based Elastic Ensemble of Deep Convolutional Neural Networks for Grading of Invasive Breast Carcinoma Histopathological Microscopic Images, Entropy, 2021, MDPI. [**Impact Factor = 2.738, h5-index = 58**, ranked 13th in top publications of Physics & Mathematics (general)]
- **Senousy Z.**, Abdelsamea M. M., Gaber M. M., Abdar M., Acharya U. R., Khosravi A. and Nahavandi S., MCUa: Multi-level Context and Uncertainty aware Dynamic Deep Ensemble for Breast Cancer Histology Image Classification, IEEE Transactions on Biomedical Engineering, IEEE press [**Impact factor: 4.756, h5-index: 74**, ranked 3rd in top publications of Biomedical Technology]
- Abdelsamea M. M., Zidan U., **Senousy Z.**, Gaber M. M., Rakha E., Ilyas M., A Survey on Artificial Intelligence in Histopathology Image Analysis, WIREs Data Mining and Knowledge Discovery, 2022 [**Impact factor: 7.558, h5-index: 54**]
- **Senousy Z.**, Gaber M. M., and Abdelsamea M. M., AUQantO: Actionable Uncertainty Quantification Optimization in Deep Learning Architectures for Medical Image Classification [**Under review in Applied Soft Computing, Elsevier**]

1.7 Thesis Organisation

The rest of this thesis is organised in the following manner.

Chapter 2 provides a detailed background and theoretical explanation of the important methods and techniques required to build our automated grading and classification systems. The concepts of machine learning, deep learning, neural networks, and issues of neural networks and how to deal with are introduced in the chapter. Moreover, a detailed description of DCNN layers has been presented. Finally, an explanation of transfer learning strategies along with the utilised pre-trained deep learning models for our contributions have been introduced.

Chapter 3 reviews the related work of histopathology image classification models and methods which utilised context-aware learning in building models and used uncertainty measurements. Furthermore, different models used as benchmarks for comparison against our contributions have been explained in different sections of the chapter. Finally, a detailed description of deep learning applications has been introduced in the context of our contributions along with chapter discussion and summary.

Chapter 4 presents our entropy-based elastic ensemble *3E-Net* model for classifying grades of histopathology images as the first contribution in this thesis. The chapter introduces the experimental methodology of the model starting from building a single patch-wise network as the first stage and image-wise networks as the second stage to establish an ensemble model. Also, the chapter goes through the development of entropy-based method for introducing uncertainty measurements to the input image samples. Furthermore, we present our metrics for evaluating the

performance of the model and comparing our work against state-of-the-art models from the literature. We evaluated our model on two datasets for breast cancer microscopic histopathology images.

Chapter 5 goes through the multi-level context and uncertainty-aware model for automated classification of histopathology images as our second contribution. The chapter introduces the experimental methodology of context-awareness between image patches by presenting in details how we build contextual information among feature maps generated from image patches and the use of Monte-Carlo dropout for applying uncertainty quantification to the model. The chapter presents different evaluation metrics to prove how effective our model in outperforming models from literature. We evaluated our model on two datasets for breast cancer microscopic histopathology images.

Chapter 6 introduces our last contribution in this thesis which is an actionable uncertainty quantification method to optimise deep learning architectures for medical image classification. The chapter presents experimental methodology for the method and how it can be applied to any deep learning model which generates probability distribution (prediction) for an input sample. The method works by measuring uncertainty for input image and it decides whether to exclude image from final classification based on the measured uncertainty compared against a threshold which is optimised using single and multi-objective functions to minimise the number of excluded images in a particular dataset. Furthermore, we used four deep learning models from literature evaluated using two medical datasets to show how effective our model in optimising the number of excluded images.

Chapter 7 presents a discussion of the contributions in this thesis and reflection on research aim and objectives. The chapter concludes the work with highlighting a few important observations to be considered as future directions for research.

Chapter 2

Background

In the previous chapter, we presented an introduction for the thesis that included a problem statement along with the aims and objectives of the research work and a general explanation of the contributions of this thesis. This chapter covers the necessary background and concepts required to comprehend our developed methods (contributions) presented in chapters 4, 5, and 6.

2.1 Overview of Deep Learning

Deep learning is a subset of machine learning which uses multiple stages of non-linear information processing in hierarchical structures for pattern categorisation and feature learning. It has also been linked to representation learning, which incorporates a hierarchy of features or concepts where higher-level concepts are determined by lower-level concepts and where the exact lower-level notions aid in defining higher-level concepts. Prior to deep learning techniques, machine learning approaches employed the structure of shallow neural network architectures. These types of architectures mainly contain only one non-linear layer, which applies feature transformations with no usage of multiple layers of non-linear features. The relatively basic design of these shallow learning models comprises a single layer (hidden layer) capable of converting raw data input or attributes into a problem-specific feature space [25].

To further explain the idea of a shallow architecture, we can take neural networks as an example. The neural network is one of the common architectures to present the approach of machine learning. Neural networks consist of three layers: input, hidden, and output layer. The input layer processes the raw input data inserted into the architecture, while the output layer generates the architecture output in the form of a prediction. The hidden layer is responsible for applying non-linear transformations between the input and output layers. In deep learning, we can have deep architectures, such as deep neural networks (DNN), which have multiple hidden layers instead of only one. DNN is a type of deep learning algorithm which involves training artificial neural networks with multiple layers (more than one hidden layer), allowing them to learn highly complex representations and abstract features of input data. Figure 2.1 presents the workflow for a DNN architecture that takes a medical tissue image as input (two-dimensional (2D) unstructured data). Then, a non-linear transformation is applied for the input to extract features using N hidden layers, which are stacked together. Each hidden layer passes feature information to the next hidden layer until an output in the form of prediction is generated to decide the final class of the input image as tumour or no tumour.

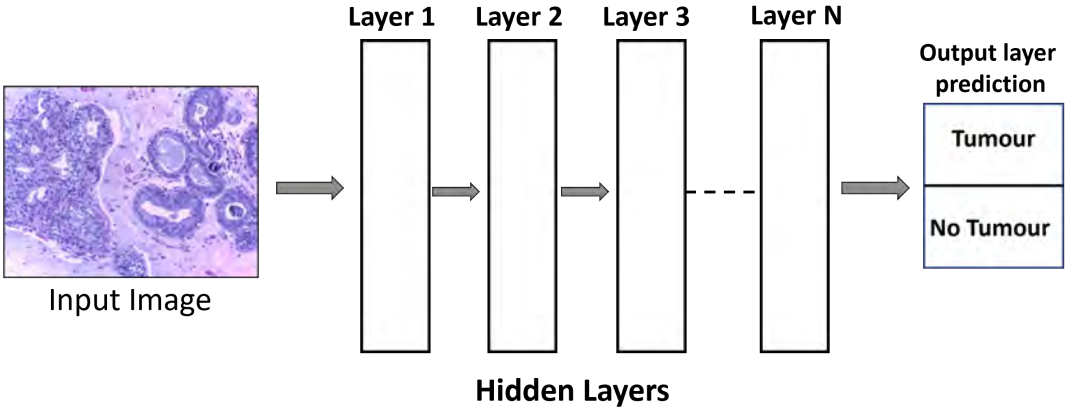


FIGURE 2.1: Generic scope of a DNN architecture which includes input data, a number of hidden layers to represent the deep architecture, and an output layer to represent two-class label predictions.

2.2 Deep Neural Networks (DNNs): A Deeper Look

In this section, we explain in detail the building blocks to form a DNN and the learning process conducted to train a DNN on training samples. Moreover, we listed the common problems encountered while training DNNs and how to fix them.

2.2.1 DNN: Building Blocks

At this stage, it is known that machine/deep learning generally is about having shallow/deep architectures which map inputs (such as images) to targets (such as class labels). Shallow architectures represent a machine learning approach with one single hidden layer, whereas deep architectures represent a deep learning approach that can have two or more hidden layers. In this subsection, we aim to have a deeper look at DNNs and introduce in detail the building blocks to establish a deep architecture of input, hidden, and output layers. A common DNN with an input layer, multiple hidden layers, and an output layer is called a feed-forward deep neural network (FF-DNN) or, in other words, a fully connected, dense, multi-layer perceptron (MLP). The smaller unit in a FF-DNN is called a neuron which mimics the biological neuron of human's brain. A neuron is considered the basic unit of a neural network. It works by taking an input value, performing some mathematical calculations and generating one output value. Then, for instance, to build a single hidden layer, it has to include a number of artificial neurons. The human brain learns by altering the strength of interconnected neurons as a result of frequent stimulation with the same signals. In artificial neural networks, we can refer to the strength of connecting linkages between various neurons as weights. Figure 2.2 represents a single neuron (a perceptron model), which is the main building block of DNNs.

As can be seen from Figure 2.2, the simple schematic representation for the perceptron model is comprised of: (1) input data values, (2) weights, (3) bias, (4) aggregation point, and (5) activation function (which acts as a threshold unit). The figure presents a number of inputs (x_1 to x_3), weights associated with the inputs (w_1 to w_3), b which is bias, f is the activation function applied to the weighted aggregation of the inputs and $\hat{y}$ is the output of the neuron. Weights in neural networks are parameters that help in transforming the input values within the hidden layers of the network. In other words, a weight determines how much of an impact the

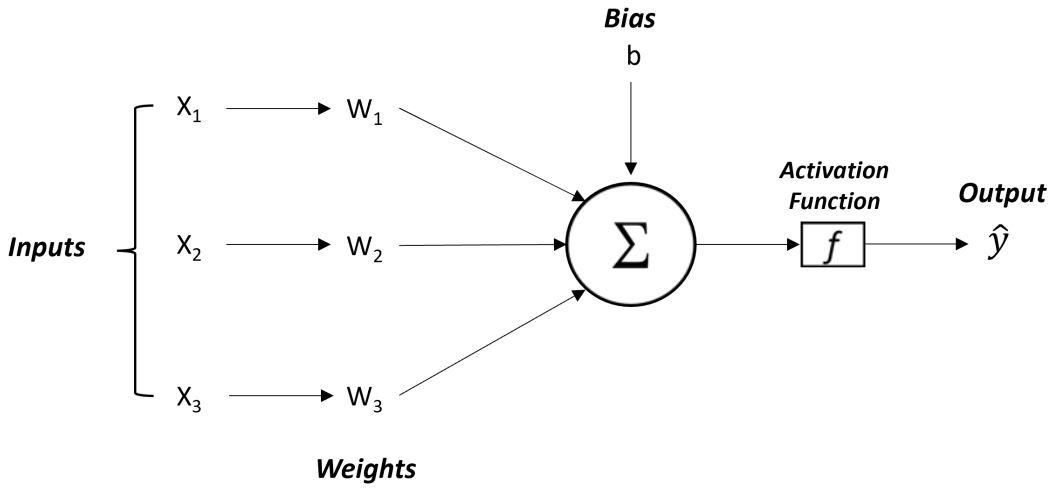


FIGURE 2.2: A single neuron is the basic unit of a neural network. It mimics the functionality of the biological neuron and learns the weights of the input data values to build a linear decision boundary that can differentiate between two different classes.

input has on the output. Bias is an additional parameter in neural networks which is used to control the output and the weighted sum of the neuron's inputs. The last important part of a perceptron model is the activation function, which is responsible for identifying whether or not a neuron should be triggered. This implies that it will use simple mathematical operations to determine whether the neuron's input to the network is essential or not throughout the prediction phase. The purpose of an activation function is to introduce non-linearity to the neural network. We can represent the following mathematical formula to depict the full operation:

$$\hat{y} = f\left(\sum_{i=1}^{N} x_i w_i + b\right), \tag{2.1}$$

where N represents the number of input samples. x and w depict the input and weights of the neuron. f represents the activation function used by the neural network. There are different types of activation functions that are used for hidden layers. The most common types are sigmoid, tanh, and Rectified Linear Unit (ReLU). Figure 2.3 represents the three common types of activation functions used in neural networks.

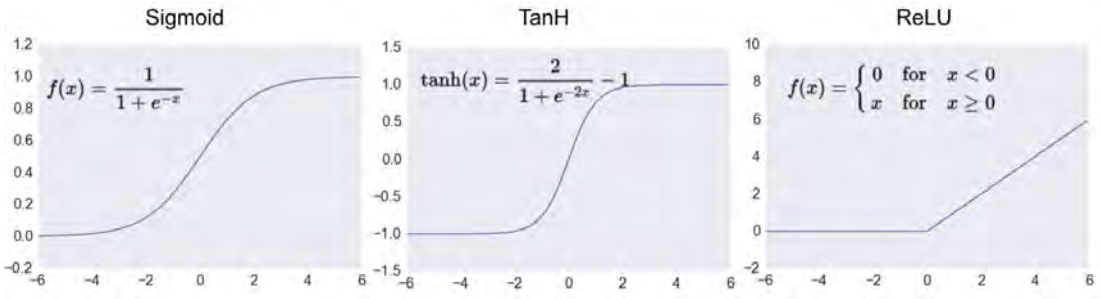


FIGURE 2.3: Common activation functions of neural networks.

Now, it is possible to represent a DNN architecture that includes input layer,

hidden layers, and output layers using neurons. Figure 2.4 presents the DNN architecture where each unit (neuron) is a simple function, but together they can create complex classification boundaries. It can be seen from the figure that each layer is formed using a group of neurons.

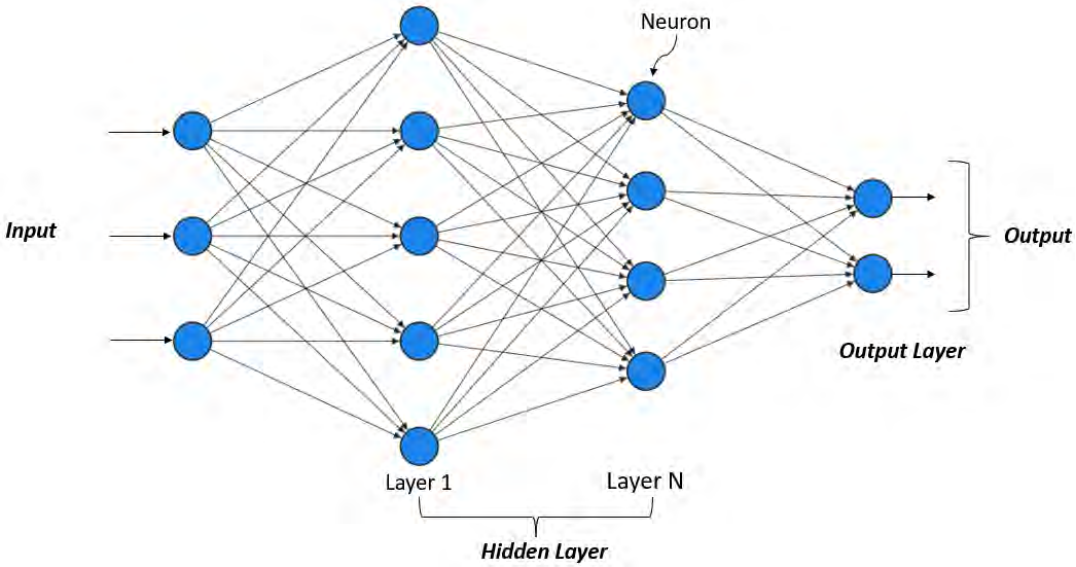


FIGURE 2.4: DNN architecture from a detailed scope including neurons where a group of neurons form a layer.

2.2.2 DNN: Learning Process

A DNN performs mapping from input to output through a series of data transformations. These transformations are learnt through various input samples as an example for training. Weights and bias are the learnable parameters in a DNN. Before learning begins, a teachable DNN randomises both the weight and bias values. Both parameters are modified as training progresses until the required values and the correct output are reached. The amount to which the two parameters impact the input data differs. Simply described, bias is the distance between the predicted value and the desired value. The discrepancy between the network's output and its desired output is made up of biases. A low bias indicates that the network is generating fewer assumptions about the output form, while a high bias value indicates that the network makes more assumptions about the output form. A model with high bias cannot capture the key features of dataset samples and cannot perform effectively on new data. As stated above, weights can be represented as the strength of the linkage between neurons. The degree of effect that a variation in the input has on the output is determined by its weight. A low weight value will have little effect on the input, whereas a higher weight value will have a greater impact on the output. Figure 2.5 presents how a transformation implemented by a layer in a DNN is parameterised by its weights.

To be able to manage the output of DNN, it is crucial to observe and monitor how much it differs from the desired output (ground truth prediction). This process is the responsibility of the loss function. The loss function computes the distance score, which is the difference between the network's prediction and the true target (the expected network's output). This distance score reflects how well the network performed on a specific task. The main objective in deep learning is to utilise this

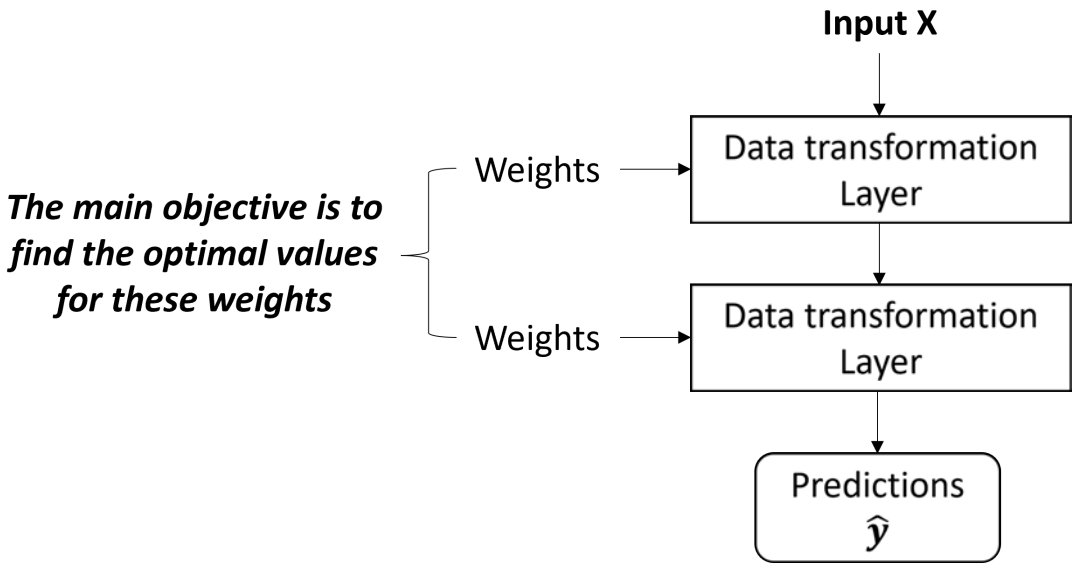


FIGURE 2.5: Data transformation of DNN layers is parameterised by network weights. Adapted from [21].

score as a feedback signal to tune the values of weights slightly in a way to enhance the performance of the network and reduce the loss value. To alter weights and achieve a minimum loss score, an optimiser is utilised, which uses Backpropagation algorithm as the fundamental approach responsible for weight correction and loss function minimisation. Figure 2.6 presents how the loss value is utilised as a feedback signal to modify the weights of the network. Learning rate is one of the important parameters used by optimiser which needs to be tuned during the learning process of a DNN. It controls how fast the model is adapted to the problem (e.g. classification problem). A small learning rate requires more updates until reaching the minimal point, which gives the minimal loss value, while a high learning rate indicates rapid changes which lead to divergent behaviour. One of the common optimisers used for DNN is the Adam optimiser [48].

As stated before, DNN weights are given random values at the beginning of the training process, resulting in a sequence of random transformations. The output of DNN is far from optimum and the loss function reflects this. However, when the network processes more examples, the weights are modified slightly in the correct track, and the loss score starts in decreasing. This is the learning loop (training iterations), which produces weight values that minimise the loss function after a sufficient number of iterations. A trained network must have outputs that are as close to the desired target as possible [21]. One learning loop is called an epoch where it can be defined as the unit which indicates one complete forward and backward pass through the entire dataset. It is considered one of the important hyperparameters during training of a neural network.

In summary, two essential phases occur during the neural network training process. The first is a feed-forward phase that describes the movement of information in the forward direction from the input layer passing through data transformation (hidden layers) until generating prediction from the output layer. During feed-forward propagation, the activation function can be considered as a gate between the input inserted into the network neurons and the output to the upcoming layer. The second is the backpropagation phase, which aims to minimise the loss function by adjusting network weights and biases. This leads to minimising the distance between the

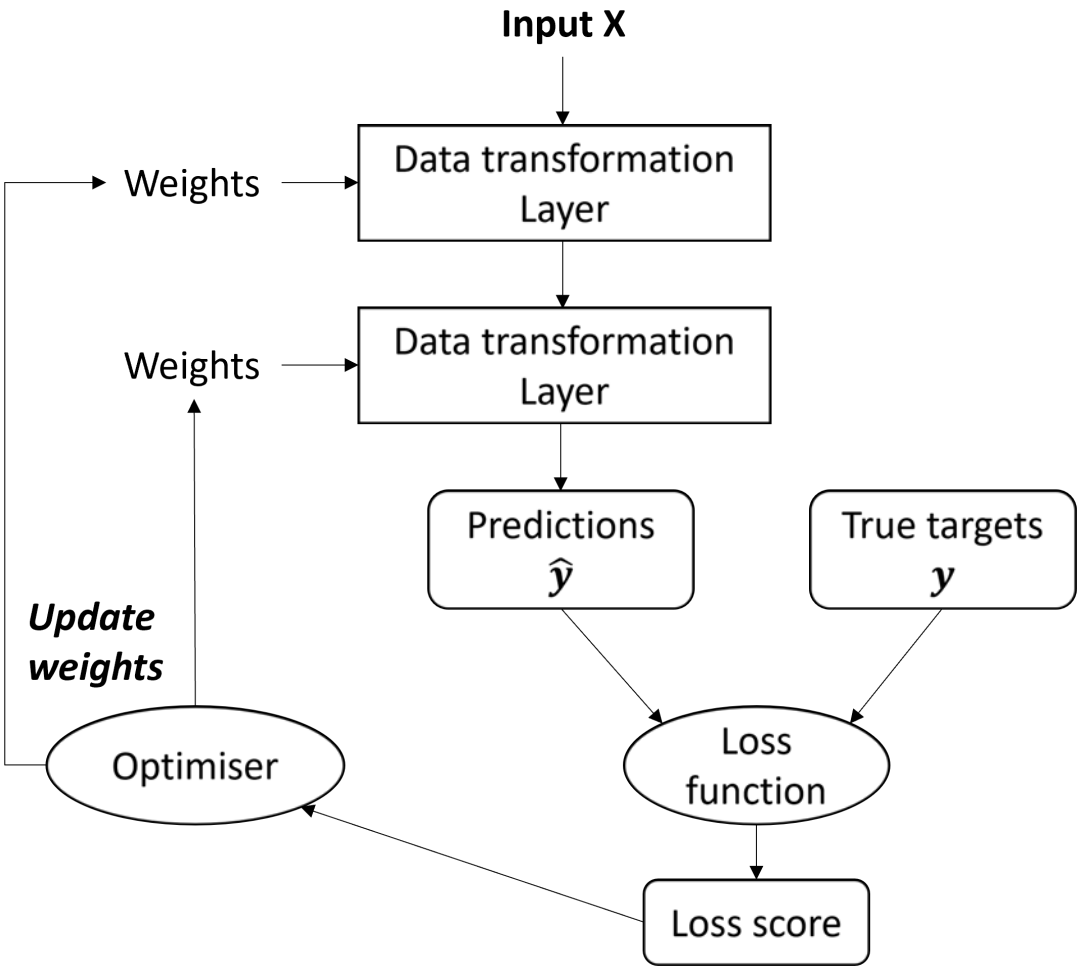


FIGURE 2.6: A deep learning flow of optimising a network using an optimiser with the loss function score used as a feedback signal. Adapted from [21].

actual output and the target (desired output).

2.2.3 Overcoming the Challenges of DNNs: Strategies for Improvement

When we train a DNN, we optimise the weights and biases so that the network can perform a mathematical mapping of input values to output values based on a given objective. Aside from the given objective, we also want to build a model that can generalise successfully from training samples to other samples from the problem domain. In other words, once the training process is done for a neural network, we aim to have a model which performs well on samples that have not been seen as well as the training samples. However, we can encounter a case where the performance on unseen samples is much worse, although the performance on training samples was good. This case is called overfitting.

Overfitting is a problem in DNNs where the model performs well on the training data but poorly on unseen data from the problem domain due to excessive memorisation of training data rather than generalisation. Overfitting occurs for two major reasons: (1) the training data samples contain noise and volatility and (2) the model has high complexity. The training data comprises noise and random volatility. A

high-complexity model can detect randomness and fluctuations in the data and understand them as underlying concepts and patterns. These noise and volatility are unique to the training set. When the model encounters new data, the wrongly learnt patterns and concepts no longer apply to the new data, and performance suffers significantly. This notion also applies to neural networks and any other machine learning model. That is, overfitting limits the ability of neural networks to generalise.

Underfitting is another problem that may be encountered during training DNNs. It is the opposite of overfitting. Underfitting indicates that a model cannot perform well on either the training samples or new unseen samples from the problem domain. The reason for this is due to the complexity of the model once again. However, the complexity is too low this time for the neural network to learn the mathematical mapping of input values to output labels. If the model is too simple to fit given data, the performance of that model will be poor.

Practically, a phenomenon known as a bias-variance trade-off will be encountered. This trade-off indicates that increasing the complexity of neural network model results in a smaller bias error on one hand, but a bigger variance error on the other. Therefore, it is crucial to find the optimal model with suitable parameters that result in the best bias-variance and to be located somewhere between overfitting and underfitting (see Figure 2.7).

Overfitting and underfitting are undesirable cases that occur while building deep learning models. Overfitting is by far the most prevalent problem in DNNs and a significantly greater concern. This concern comes because evaluating deep learning models on training samples is a bit different from evaluating the model on unseen samples, which is what we really care about (testing set). There are different techniques that can be utilised to avoid overfitting. Good practice can be by reducing the number of neurons in the DNN to make the model less complex. Also, using regularisation techniques such as L1, L2, and dropout which are three prominent and effective methods that we examine below.

Regularisation is a collection of techniques to reduce the complexity of DNN models during the training phase and to avoid overfitting. Simply, regularisation works by adding a penalty term to the loss function in order to lower the complexity of a DNN model and hence avoid overfitting. The key difference between L1 and L2 regularisation techniques is in the penalty term. The L2 regularisation approach, also known as weight decay or Ridge Regression, is the most popular of all regularisation techniques. It adds squared magnitude of coefficient (sum of model weights squared) as a penalty term to the loss function. While L1 regularisation is named Lasso Regression and it adds an absolute magnitude of coefficient (sum of the absolute values of the weights). We can represent the formulation for the loss function using L1 and L2 regularisation as follows:

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} |w_i| \tag{2.2}$$

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} w_i^2 \tag{2.3}$$

where Equations 2.2 and 2.3 represent loss equation using L1 and L2 regularisation, respectively. *Error* represents the loss score between the prediction of DNN model $\hat{y}$ and ground truth prediction y. λ determines the amount of regularisation that is

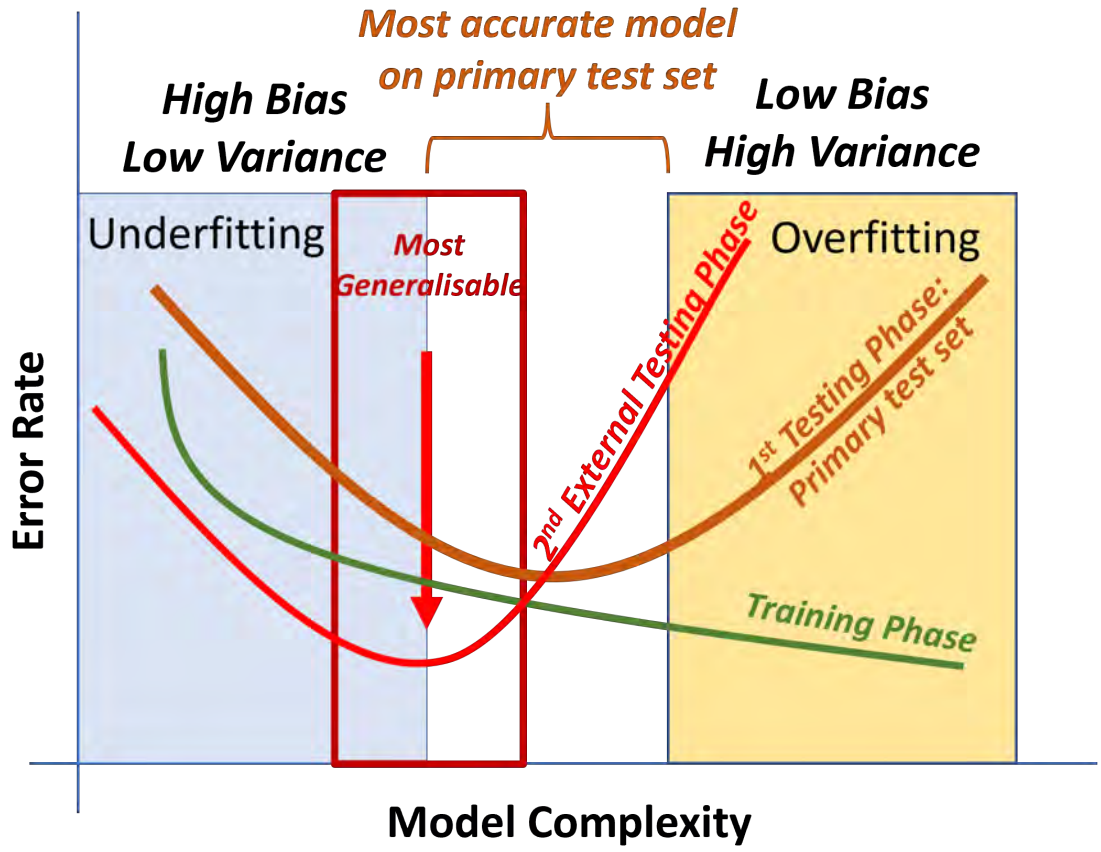


FIGURE 2.7: This diagram depicts the trade-off between bias and variance. Training data (green line) may not always accurately represent testing phase outcomes. Underfitting data is less variable, yet it has a high error rate and a high bias (blue box). Overfitting data, on the other hand, results in low bias and high variance (yellow box). The optimum region exists between overfitting and underfitting of data and may not be ideal until multiple testing trials have been performed (red line). Adapted from [80].

manually tuned and it has to be greater than zero to represent a loss function with regularisation.

Dropout is another popular technique to avoid overfitting. It applies a dropping (deactivating) operation to some random neurons from the DNN during the training phase. This process aims in reducing the complexity of the DNN model to avoid overfitting. Figure 2.8 represents a diagram of the neural network architecture which uses the dropout technique to reduce complexity.

Another important method to handle overfitting is data augmentation. In some particular cases, the dataset used is small in the number of samples and may cause overfitting. Data augmentation is an approach to greatly increasing the diversity of available data samples for training models without acquiring additional data. Increasing the number of samples without collecting new data is a good strategy as it alleviates the process of collecting new samples especially in the medical domain which is a bit difficult to collect. Data augmentation works by using the current samples available in a particular dataset, and it applies some operations to the current samples such as rotation by multiple degrees, flipping either horizontally and/or vertically, and applying colourisation operation.

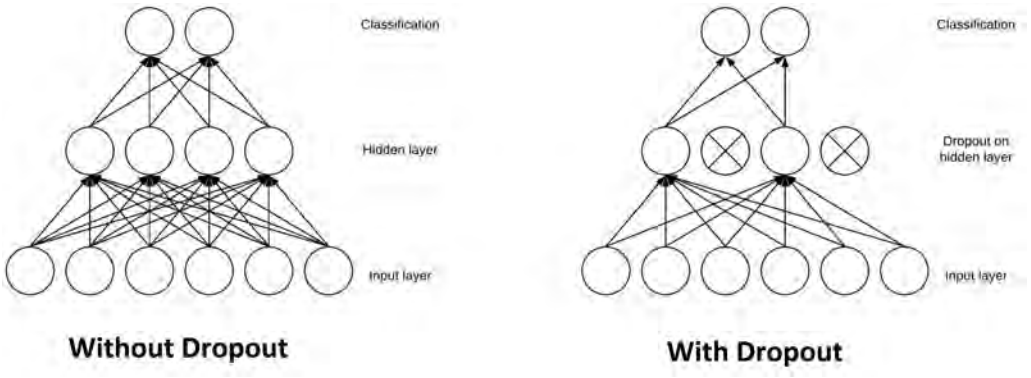


FIGURE 2.8: Neural network architecture before applying dropout (left) and after applying dropout (right).

Underfitting is a simple and uncomplicated process. As mentioned before, underfitting occurs due to a less complex model. Therefore, to reduce underfitting, we have to add more parameters to the neural network (weights and biases). These parameters can be added by increasing the number of neurons and layers in the network.

2.3 Deep Convolutional Neural Networks (DCNNs)

DNNs are used for a wide variety of applications. In addition to MLP, which is a class of FF-DNNs (explained above), there are different types of DNN that are commonly used for different tasks such as speech recognition, machine translation, text-to-speech processing, sentiment analysis, image processing, and computer vision (CV). For example, Recurrent neural networks (RNNs) are one of the common DNNs that can be used for any task relevant to text processing. Our focus in this research work is to develop novel deep learning models for the task of image classification (diagnosis systems). This task is one of the common tasks for CV and a popular type of DNNs which is performing very well for such tasks is Deep Convolutional Neural Networks (DCNNs) [52].

In this section, we cover all the concepts and building blocks for establishing DCNNs as well as explaining two common DCNNs that have been used to build our models in the upcoming chapters.

DCNN is a type of deep neural network that excels in processing input with a grid-like structure, such as images. A digital image is a two-dimensional representation of visual information created using discrete digital (pixel) values [31]. It consists of a grid-like arrangement of pixels where pixel values indicate how bright and what colour each pixel should be. DCNN takes input images and then presents importance in terms of learnable weights and biases to different parts in images (e.g. important objects). This aids in discriminating between multiple objects in images. DCNNs require substantially less pre-processing compared to pre-deep learning methods which were based on hand-engineered features. DCNNs can learn filters/characteristics with sufficient training. A DCNN design is similar to the connectivity structure of neurons in the human brain and was inspired by the organisation of the visual cortex. Particular neurons only react to stimuli in a small region of the visual field called the receptive field. A group of similar fields will encompass the full visual region if they overlap [32].

DCNN design typically consists of three types of layers: convolutional layer, pooling layer, and fully connected layer. Figure 2.9 presents a CNN design that takes an image as input and then it applies a series of convolution and pooling layers. Then, it flattens the features extracted from the last pooling layer to be inserted into a fully connected layer and softmax output layer. The softmax function is utilised as the activation function in the output layer of CNN models. It aids in generating probability distribution, which represents the model's prediction. At this point, the generic scope of a CNN is clear with its design and the layers used. Next, We explain in more detail how each layer works and what are the operations done.

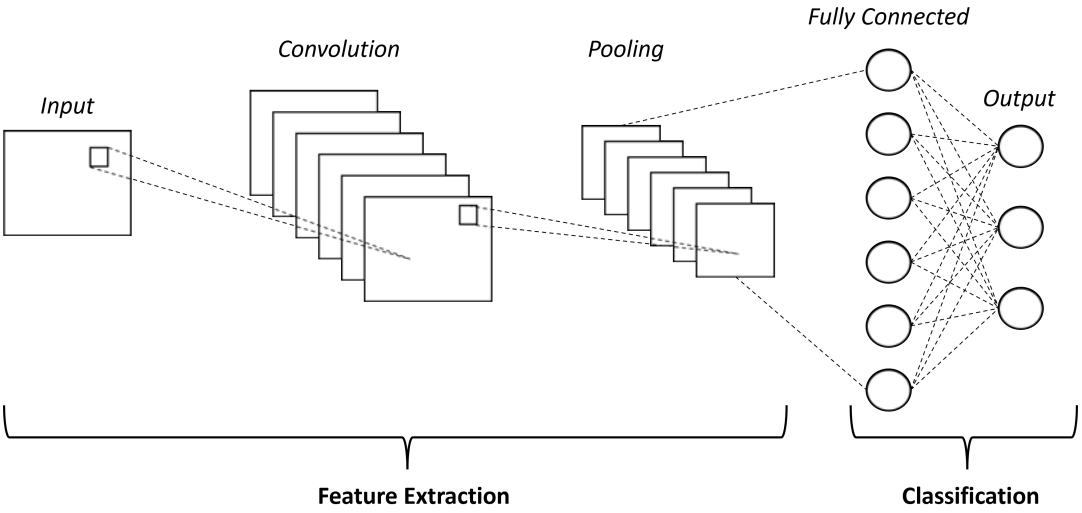


FIGURE 2.9: A CNN architecture consists of a convolutional layer, a pooling layer, and a fully connected layer followed by a softmax layer to generate image prediction. We can represent CNN for image classification into two stages: (1) the feature extraction stage, which takes input image and applies convolution and pooling operations to extract features from input image, and (2) the classification stage, which has a fully connected layer and an output softmax layer to generate the final classification of the image.

2.3.1 Convolutional Layer

The main building block of CNN is the convolutional layer. It handles most of the computational burden on the network. Convolution is an operation coming from the field of signal processing. In the field of deep learning, it essentially performs a dot product (matrix multiplication) between two matrices. The first matrix is a specific portion of the input image known as the receptive field and the second matrix is a set of learnable parameters known as filter or kernel. This matrix multiplication process (dot product) occurs by sliding the kernel over different image regions (sliding the kernel over the image's height and width). The kernel has a smaller spatial size compared to the image (smaller height and width) but it is equivalent to image depth. For instance, if an image has three (RGB) channels, the kernel's height and width will be smaller than the image's height and width, but the depth will span all three channels. Figure 2.10 presents an example of a 2D convolution operation that occurs when a convolution operation is applied on an input image.

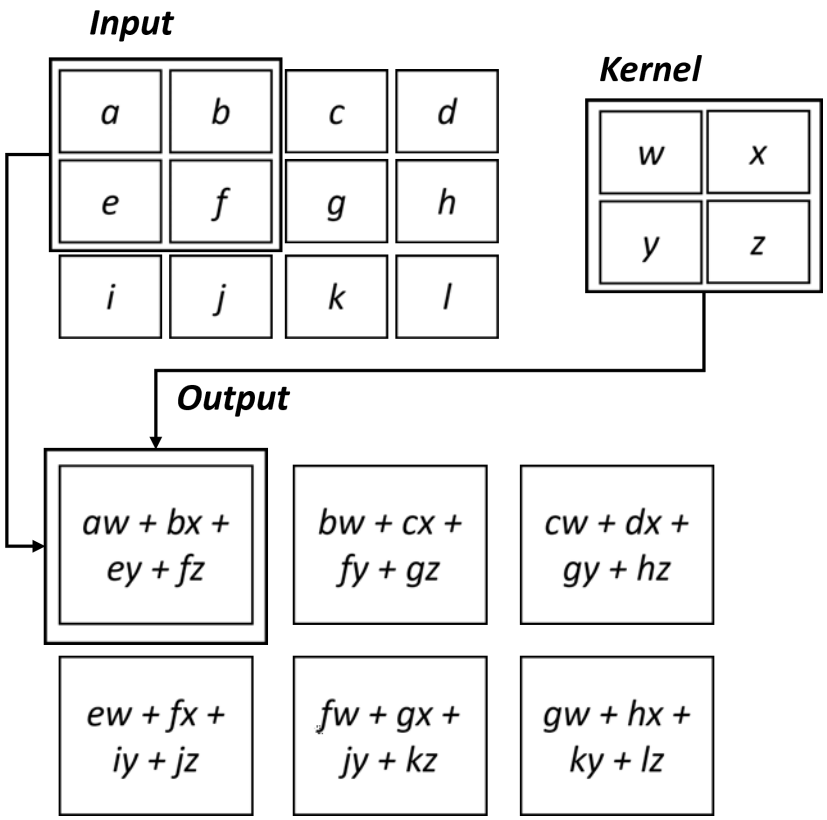


FIGURE 2.10: 2D Convolution operation occurs in the convolutional layer between different image regions (receptive fields) and the kernel. The output matrix is called a convolved feature map, which contains the salient features of different parts of the image. Adapted from [32].

The kernel moves across all image parts passing through image's height and width, generating an image representation of all receptive regions. The image representation generated is a two-dimensional description of the image known as an activation feature map, which contains the kernel's reaction at each spatial place in the image. A hyperparameter named stride is used to decide the stepping of the kernel over the image (sliding size).

The objective of the convolution operation is to extract high-level features from input images. Typically, a CNN design should have multiple convolutional layers. The first convolutional layer is mainly responsible for producing low-level features such as edges. Then, with the addition of more convolutional layers, the network captures high-level features.

The output size of a convolved feature map depends on the kernel size, the stride value, and another hyperparameter called padding. Padding refers to the insertion of empty pixels around the borders of an image to preserve the original size of an input image. This is done to allow the kernel to execute full convolutions on the image's edge pixels. There are two types of padding: valid padding and same padding. Valid padding requires setting the padding value to zero (which means no padding is applied), while same padding indicates that the padding is applied so that the input is equivalent to the output in size provided that stride value is assigned to one. The following formula is used to calculate the output size of the feature map ($Out_w \times Out_h$) given the size of the input image $W \times H$, kernel $f_w \times f_h$, stride s, and

padding p:

$$Out_w = \left\lfloor \frac{W - f_w + 2p}{s} \right\rfloor + 1 \tag{2.4}$$

$$Out_h = \left\lfloor \frac{H - f_h + 2p}{s} \right\rfloor + 1 \tag{2.5}$$

The motivation behind using the concept of convolution, instead of the normal process of having fully connected hidden layers as present in FF-DNN, lies in three crucial factors that motivated researchers to use DCNN: sparse interaction, shared parameters, and equivariance of representations. Basic neural networks apply matrix multiplication to represent the parameters' matrix characterising the connection between the input and output units. This implies that each output unit communicates with each input unit. On the other hand, CNNs have the feature of sparse interaction where the kernel (which has a smaller spatial size compared to the input image) aids in detecting salient information of hundreds of pixels from an image that has thousands of pixels. This indicates that we need to keep fewer parameters, which not only decreases the model's memory demand but also enhances the model's statistical efficiency [32].

Parameter sharing is another important aspect that helps in reducing the number of parameters and computational cost. The idea behind this is that features learned in one part of an image can also be applicable to another part of the same image. Hence, instead of having separate kernels for each region of an image, a single kernel is used to convolve the entire image. This leads to parameter sharing and the reuse of the same set of parameters throughout the image [32].

Another feature of CNNs that is related to parameter sharing is equivariance to translation. This refers to the ability of the layers in a CNN to adapt to translations in the input image. The shared parameters allow the network to be robust to small translations in the input, without having to learn separate parameters for each possible translation. This results in the network being equivariant to translations, meaning that if the input is shifted, the output also changes accordingly [32].

2.3.2 Pooling Layer

The next common stage after the convolutional layer is the pooling layer. The pooling layer is mainly responsible for lowering the spatial size of convolved feature map generated from the convolutional layer. This stage aids in reducing computational power necessary for processing data via dimensionality reduction. It is also beneficial for capturing dominating characteristics that are rotational and positional invariant, allowing the model to be efficiently trained.

There are two forms of pooling: maximum (max) pooling and average pooling. Max pooling retrieves the maximum value from the portion of the image contained by the kernel. While average pooling retrieves the average of all values from the region covered by the kernel. Figure 2.11 presents the two forms of the pooling layer. As can be seen in the figure, max pooling works by picking the maximum value of the receptive field (image portion covered by kernel - identified by different colours). While average pooling works by calculating average of all values covered in the receptive field.

The convolutional and pooling layers utilised in a DCNN can process the input image and capture image features. Moving forward to the upcoming stage, the

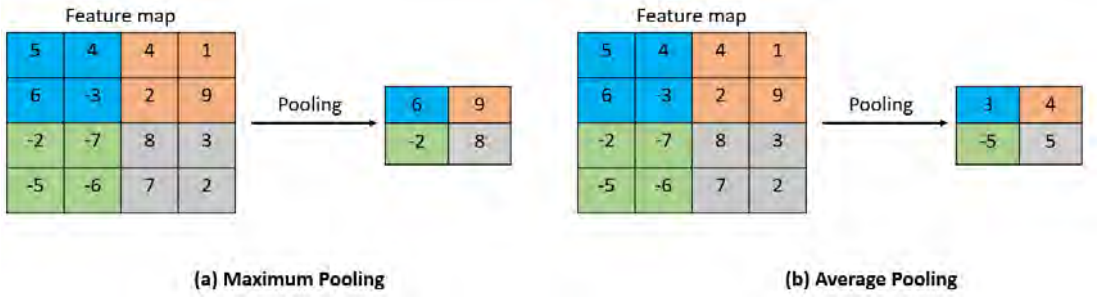


FIGURE 2.11: An example of maximum pooling and average pooling applied to feature maps produced from the convolutional layer by using a kernel size of 2×2 and a stride of 2.

output extracted from successive convolutional and pooling layers is flattened and inserted into a standard neural network block to classify an input image.

2.3.3 Fully Connected Layer

The final block in a DCNN design is the fully connected layer. The fully connected layer is mainly responsible for learning non-linear representations of high-level features as introduced by the output generated by the last combination of convolutional-pooling layers. This is done by flattening the shape of high-level features learnt in the previous stage (which is the shape that suits a standard neural network). The flattened form of the features is then connected to a fully connected layer, which helps to map the representations between input and output. When all the building blocks of DCNN are joined together, the network is able to discriminate between different levels of features in an input image and to classify the image into a certain class using softmax classification. Figure 2.12 depicts the last stage of a CNN design.

2.3.4 Non-Linearity Layers

As the convolutional layer is based on linear operation, non-linearity layers are commonly placed just after the convolutional layer to provide non-linearity to the output feature map. We introduced above the frequently used activation functions that can be used to introduce non-linearity (see Figure 2.3).

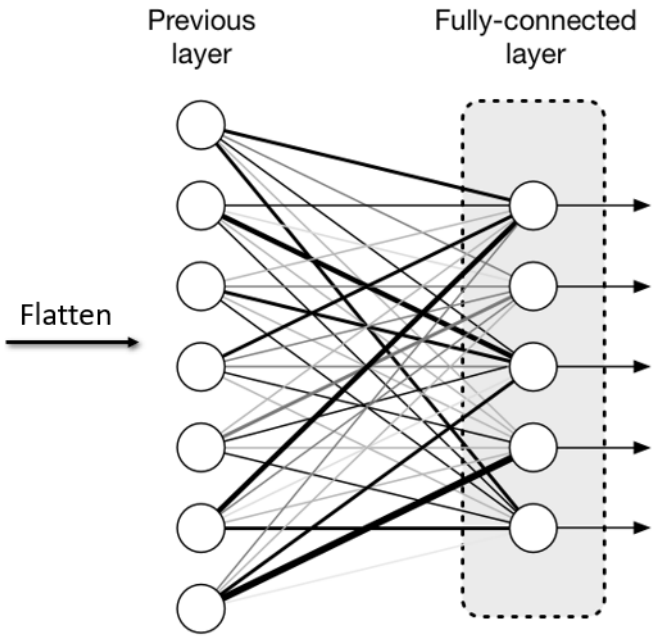


FIGURE 2.12: An example of the fully connected layer. As illustrated by the figure, the output generated from the last pooling layer is flattened and fed into a fully connected layer to map the generated features from the image into an output that presents the final image classification.

2.4 Transfer Learning using Pre-trained DCNN Models

Medical imaging datasets are limited in size and are categorised as small datasets. DCNN models that are trained on medical imaging samples are prone to overfitting due to the limited number of samples. In addition, training a DCNN model from scratch is time-consuming and can lead to inaccurate performance. Therefore, the transfer learning approach has been considered one of the important techniques that aid in building our developed models to generate accurate results and save time during the training process.

Transfer learning is a prominent approach in CV since it helps us to create accurate models in a timely manner. It is a machine learning approach in which we reuse a previously trained model as the foundation for a new model on a new problem. Instead of starting the learning process from scratch (e.g. using randomly initialised weights), transfer learning begins with patterns learnt in one domain for the purpose of addressing a problem in another domain. This allows us to build on existing knowledge rather than starting from scratch. In other words, transfer learning is a machine learning technique where a model which is trained on one task is reused on another related task [32]. For instance, in our situation, we use models trained on ImageNet dataset prepared for ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [86] for other domains, such as medical imaging datasets. The intuition behind this approach is that a model trained using large-scale images will have similar low-level features to images from any other domain. Therefore, a model trained on low-level features can be used as the starting point instead of applying training from scratch. This means that the general features learnt from large-scale images (one domain) can be utilised by small medical imaging datasets (another domain).

Transfer learning can be seen as an optimisation method that facilitates quick progress while modelling the second task from a model learnt on one task. It can

yield considerably greater performance than training models with a little amount of data from scratch [75]. Researchers prefer to start with a pre-trained model that already understands how to identify objects and has learnt general properties like edges and shapes in images. In CV, transfer learning is typically introduced through the usage of pre-trained models. A pre-trained model depicts the concept of transfer learning as they are a type of model that is trained on a large benchmark dataset that can be used to handle a problem comparable to the one we have at hand (medical image classification task). It is important to note that transfer learning can be used in case the features initially gained from a model trained on one task are generic and can be used on another task [122]. Various cutting-edge image classification techniques rely on transfer learning solutions [37, 52, 100].

2.4.1 Transfer Learning Strategies

As explained above, a DCNN typically has two main stages: (1) the feature extraction stage, which is responsible for generating salient features from input images. This stage is mainly constructed from successive convolutional and pooling layers, and (2) the classifier stage is responsible for generating image prediction based on the detected features. It consists of fully connected layers. Figure 2.13 presents a simplified version of a CNN architecture showing the two stages for an image classification task.

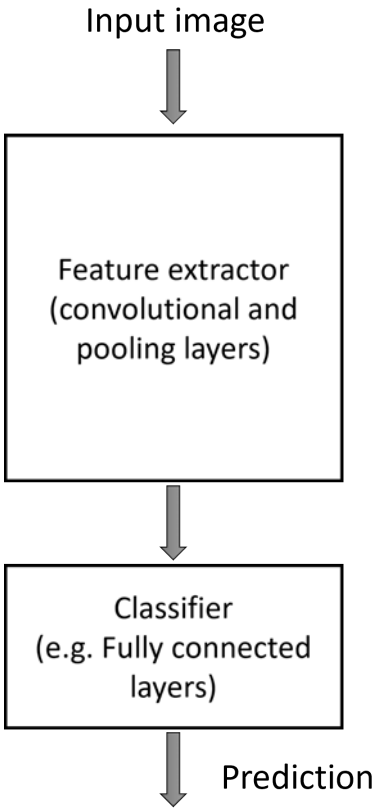


FIGURE 2.13: Workflow stages of CNN architecture.

As a part of re-purposing pre-trained DCNNs, it is important to highlight that DCNNs can learn hierarchical feature representations. This indicates that features learnt by the first convolutional layer are generic while features learnt by the last convolutional layer are specific to the problem domain. Generic features learnt by

lower convolutional layers can be transferred and reused across other problem domains.

To reuse a pre-trained model, it is crucial to change the number of neurons in the classifier's output layer to the number of classes of the classification task. Then, we have to follow a fine-tuning process based on one of the following strategies:

1. **Full training for the model:** In this strategy, the full architecture (feature extractor and classifier stages) of the pre-trained model is utilised for training on a particular dataset. It is recommended with this strategy to use a large dataset as the model will be trained from scratch.

2. **Partial training for the model:** In this strategy, some layers will be trained, while other layers will be left untrained (frozen). As mentioned earlier, the lower layers of the network present generic features, and the higher layers present specific layers. In this case, it is crucial to experiment with different scenarios of training some layers and freezing others. A frozen layer indicates that the weights in this particular layer remain unchanged during training. Based on experimentation, it is a good practice to keep more layers frozen when training a model with a high number of parameters on a small dataset. On the other hand, if we have a large dataset with few parameters, it is recommended to increase the model's complexity by adding more layers as overfitting is not going to be a problem.

3. **Freeze feature extractor:** In this case, we freeze all the convolutional and pooling layers that are responsible for the feature extraction stage. The fundamental concept, in this case, is to preserve the feature extractor in its original state (all the weights of the feature extractor remain unchanged). The classifier takes the output generated from the feature extraction stage. This scenario is ideal when computational capacity is limited, the used dataset is small, or the pre-trained model is solving a problem similar to the problem which was trained on before. Figure 2.14 presents a diagram with possible scenarios for applying fine-tuning to pre-trained models (transfer learning strategies).

2.4.2 Overview of the used Pre-trained DCNN Architectures

Here, we describe the pre-trained DCNN architectures that have been used as the backbone of our automated diagnosis systems. These architectures are based on the CNN design explained above (feature extraction and classifier stages). There are many DCNN architectures that showed promising performance for image classification task: Xception [22], VGG [100], ResNet [37], DenseNet [41], MobileNet [40], Inception [106], and EfficientNet [107].

Throughout time, DCNN architectures become deeper by adding more layers to learn more complex patterns, enhance performance, and make architectures robust for complex image recognition and classification tasks. However, it is found that increasing the number of layers (i.e. architectures become deeper) beyond a particular extent leads to a vanishing gradient issue where it becomes increasingly harder to train the architecture, and the architecture's accuracy begins to saturate and subsequently declines. The vanishing gradient problem occurs during the training of DNNs with gradient-based learning techniques. Stochastic gradient descent is one of the gradient-based learning methods. It is an optimisation method for finding the optimal model parameters that lead to a minimised error function to reflect the

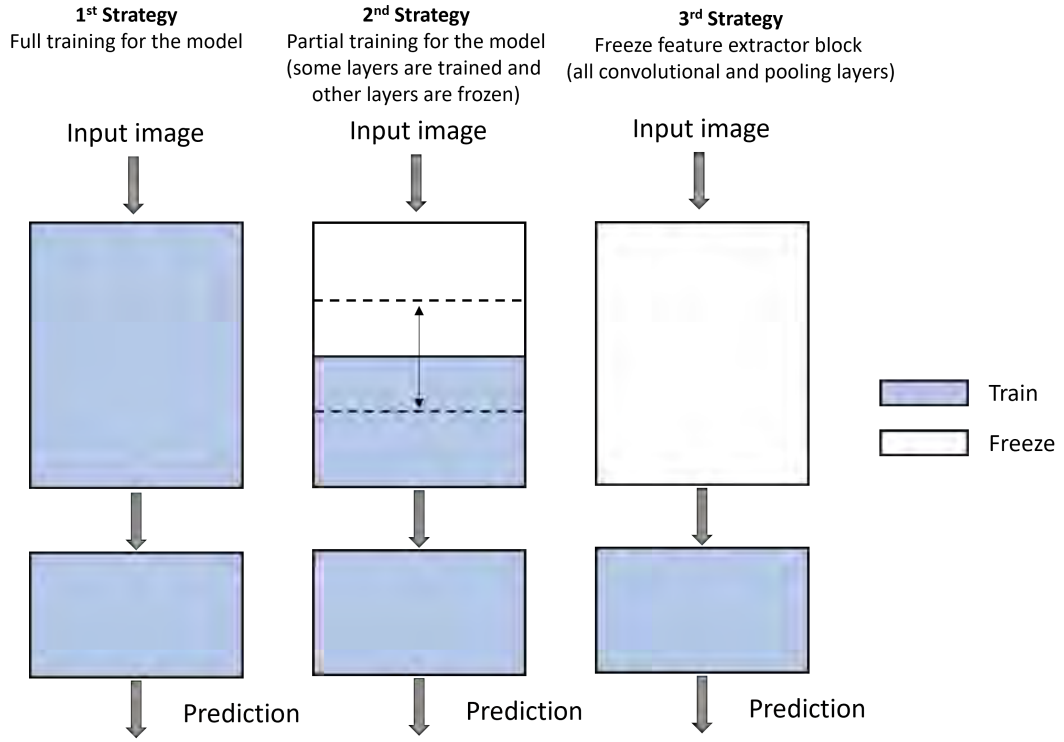


FIGURE 2.14: Transfer learning strategies for deep learning techniques.

optimal fit between predicted and ground truth outputs. In gradient-based learning techniques, specifically during backpropagation, the weights of a DNN are adjusted proportionally to the gradient value. The issue is that in some scenarios the gradient shrinks exponentially, thereby the weights are not updated and the learning stops [12]. In our research work, we utilised two popular DCNN architectures that showed good performance and the ability to solve the issue stated above:

1. **ResNet:** Deep Residual network is one of the popular DCNNs that has been introduced by researchers from Microsoft Research in 2015 [37] to fix the issue of vanishing gradient. Deep Residual Networks are nearly identical to networks that contain convolution, pooling, activation, and fully-connected layers piled one on top of the other. The identity connection (skip connection) between the layers is a unique modification to the basic network that makes it a residual network. The skip connection bypasses training from a few levels and links directly to the output. So instead of having layers learning the underlying mapping, we let the network fit the residual mapping. The residual block utilised in the network is shown in Figure 2.15. The identity connection is represented by the curving arrow that comes out from the input and goes to the summation point of the residual block.

   The benefit of including this sort of skip connection is that if any layer degrades the performance of the DCNN model, this will be bypassed by regularisation. As a result, very deep networks can be trained without the issues caused by vanishing gradients.

2. **DenseNet:** Densely connected convolutional neural network is another promising architecture proposed by Huang et al. [41] to alleviate the problem of vanishing gradients, improve feature propagation, and significantly reduce the

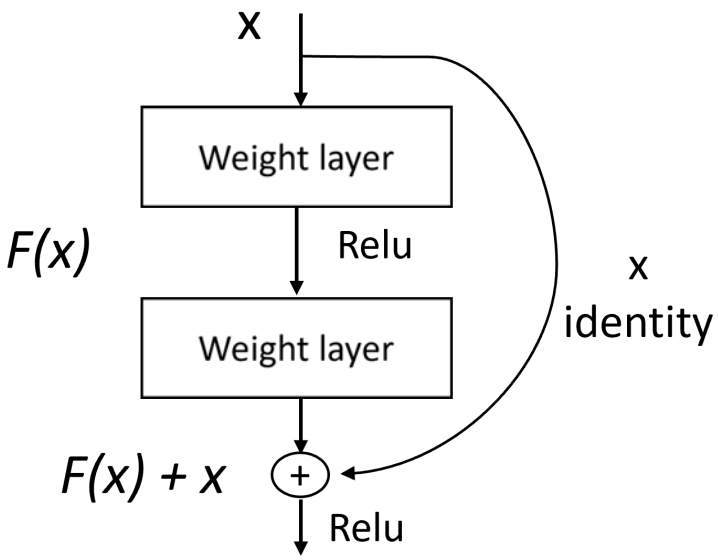


FIGURE 2.15: Building block in residual learning. Adapted from [37].

number of parameters. DenseNet works in a feed-forward style by connecting each layer to every other layer. Unlike traditional CNNs, which have L connections (one connection between a particular layer and the following layer), DenseNet has $L(L+1)/2$ direct connections. For a current layer, the feature maps generated from all previous layers are used as input, and the feature maps generated from the current layer are then used by all subsequent layers. Figure 2.16 presents a schematic layout of DenseNet. Unlike ResNet, which combines features using summation, the features combined from previous layers are concatenated together.

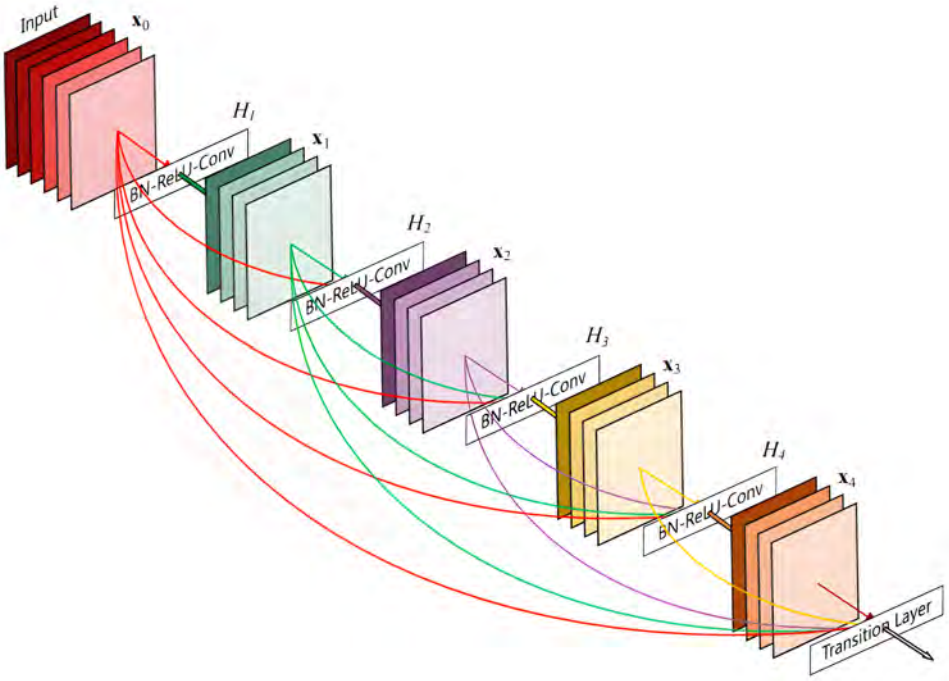


FIGURE 2.16: Schematic layout of DenseNet. Adapted from [41].

In our work, we utilised pre-trained ResNet-152 and DenseNet-161 networks which are particular versions of ResNet and DenseNet architectures, respectively. These pre-trained networks have been used for the feature extraction of an input image in our contributions presented in chapters 4 and 5.

2.5 Summary

In this chapter, we provided an overview of deep learning. We explained the building blocks, the learning process, and the problems of DNNs. Then, we explained DCNNs and its building blocks. Lastly, we presented transfer learning strategies and an overview of the utilised pre-trained DCNN architectures in our contributions. In the next chapter, we conduct a literature review of the recent methods applied to histopathology image analysis.

Chapter 3

Literature Review in Histopathology Image Analysis

In the previous chapter, we went through the main concepts of deep learning and DCNN architectures that are important building blocks in our work. Here, we describe the related work conducted in the field of histopathology image analysis.

3.1 Overview

DCNNs have shown tremendous performance for image classification and other related Computer Vision (CV) tasks. As mentioned in chapter 2, DCNN design mainly have two important stages: (1) feature extraction which is mainly responsible for the extraction of features from the input image, and (2) classifier which presents the final prediction (class label) of the input image. In this chapter, we thoroughly examine various image classification methods used in histopathology, which is the study of tissues at a microscopic level. The methods we cover are based on both single and ensemble learning techniques, and we delve into their key principles and techniques, analysing their performance in classifying histopathology images.

We then turn our attention to recent context-aware models that preserve contextual information among different regions of an input image, which can lead to improved accuracy in diagnosis. In addition, we review image classification models that incorporate uncertainty measures, providing a way to quantify the degree of uncertainty associated with a prediction. This information is particularly important in medical imaging, where a misdiagnosis can have severe consequences. We also examine the applications of deep learning in medical image analysis. The chapter concludes with a discussion and summary of the key findings and insights, including an examination of the research gaps in the field and potential solutions to address these gaps.

3.2 Histopathology Image Classification Methods

Recent methods have been proposed to enhance the performance of classification for histopathology image sections. For example, Nazeri et al. [73] proposed a two-stage CNN model for the classification of breast histopathology images. Their model has two stages: (1) patch-wise CNN, which takes image patches as input and extracts salient features, and (2) image-wise CNN, which generates an image-level class prediction for a particular input image. A work introduced by Koné et al. [51] presented a CNN hierarchy system that categorises images from general pathological groups, such as carcinoma and non-carcinoma, and then into the four normal, benign, in

situ carcinoma and invasive carcinoma. A transfer learning approach has been developed by Vesal et al. [109] for the classification of breast histopathology images. They used pre-trained versions of Inception-V3 and ResNet-50 CNN architectures for extracting features from input image patches and classification. The CNN architectures are pre-trained on the ImageNet dataset. Gupta et al. [35] proposed a sequential framework for breast histopathology image classification which uses multi-layered deep features that are generated from a fine-tuned DenseNet. Their framework captures different magnifications for image regions that can include discriminative features at different levels. A sample is taken from a certain deep layer if it passes a particular confidence threshold.

Shi et al. [98] devised a pairwise-based deep ranking hashing (PDRH) technique for skeletal muscle and Lung cancer histopathology image classification and retrieval that can extract features from images while also learning their binary representations. Furthermore, their model retains the intra-class relevance order for image retrieval while preserving the inter-class difference for image classification. To accomplish these aspects, they created a pairwise matrix based on image labels and their associated relevance order within the same class. Then, they developed an objective function to learn binary image descriptions. Lastly, they created a supervised deep learning approach using the created objective function to learn features and associated binary codes. To avoid the lack of annotated training samples in histopathology images, Xia et al. [114] proposed a framework based on GoogLeNet DCNN for the classification of histopathology data with minimal training datasets. They suggested a method for classifying tumour patches from Whole Slide Images (WSIs). Their method employs WSI from different cancer types in plenty to train a CNN to discover the image representation of cellular components, which can then be transferred and tuned for tumour identification in WSI histopathology images in scarcity. Roy et al. [85] developed a patch-based classifier (PBC) that uses CNN to classify breast histopathological images. The suggested method operates in two modes: one patch in one decision (OPOD) and all patches in one decision (APOD). The proposed PBC predicts the patch's class label using OPOD mode. If the class label for all extracted patches matches the class label for the image, the result is considered an accurate classification. In the other mode, APOD, the class label of each extracted patch is extracted as in OPOD, and the class label of the image is determined using a majority vote process.

The work introduced by Li et al. [61] proposed a histopathology image classification model which extracts small- and large-sized patches from image and applies ResNet-50 to extract different levels of patch features (considering features at the cell and tissue levels). They designed a patch screening method and a CNN which filters out patches of a particular image that lack enough and useful information that relate to image label and picks discriminative patches. The screening method used is based on a clustering algorithm that groups image patches according to their phenotypes. They employed SVM for final image-wise classification. The work introduced by [58] has presented a self-interpretable invasive breast cancer diagnosis approach due to the variability of cancer progression as well as the variety of benign tissue generative lesions in breast cancer histopathology images. The method employs contrasting characteristics and features between normal and malignant images in a weak-supervised style with limited annotation information, generating a probability map of anomalies to evaluate its reasoning. To discover the main structural patterns among normal image patches, a fully convolutional auto-encoder is utilised. A one-class support vector machine and a one-layer neural network are employed to recognise and evaluate patches that do not share the features of this

normal population. Sun et al. [104] suggested a deep learning approach for categorisation of liver histopathology images that utilises global labels only. Due to the complexity of features and the scarcity of annotated training instances for liver histopathology images, patch-level features are collected and employed by transfer learning in conjunction with multiple-instance learning to produce image-level features for final classification. Alom et al. [4] proposed a hybrid architecture named Inception Recurrent Residual Convolutional Neural Network (IRRCNN) that has been used for the classification of breast cancer histopathology images. Their model combines Inception-V4, ResNet, and Recurrent Convolutional Neural Network (RCNN) introduced in [63]. The IRRCNN model consists of stacks that has convolutional layers, inception recurrent residual units (IRRU), transition blocks, and a softmax output layer.

Rui et al. proposed DenseNet121-AnoGAN [68] for the classification of breast histopathology images as benign or malignant. Their proposed architecture has two stages: (1) patch screening method, which applies screening to mislabelled patches with unsupervised anomaly detection utilising generative adversarial networks (AnoGAN) and (2) multi-layered feature extraction from discriminative patches using (DenseNet). Another work proposed by Hirra et al. [39] employed Deep Belief Network (DBN) to construct a patch-based deep learning system called Pa-DBN-BC to identify and classify breast cancer histopathology images. The Pa-DBN-BC model includes four main stages: preprocessing, patch production, DBN, and classification. To extract features from input image patches, an unsupervised pre-training and supervised fine-tuning phase is performed. The patches of the histopathology images are then categorised using logistic regression. The model's output is presented in the form of a two class probability distribution where probability values differentiate between cancer samples (positive) or non-cancer samples (negative). Jiayun et al. [55] developed a multi-resolution multiple instance learning (MIL) model for fine-grained grade prediction that employs significant feature map representations to detect suspicious image areas. Their model can be trained end-to-end utilising slide-level annotations only, rather than region- or pixel-level annotations. The model is evaluated on WSI large-scale prostate biopsy dataset.

Sornapudi et al. [101] introduced a Deep Learning (DL)-based nuclei detection technique, which is based on collecting localised information through super-pixels generation using a basic linear iterative clustering algorithm and training with a CNN. Their framework detects nuclei and classifies them into one of squamous epithelium cervical intraepithelial neoplasia (CIN) grades. The work introduced by Li et al. [56] proposed a DCNN architecture based on Xception network for fine-grained classification in breast cancer histopathology images. Their architecture has three stages. First, they integrated multi-class recognition and verification tasks of image pairs into the representation learning process. Second, a piece of prior knowledge is developed during the feature extraction process, where the variance in feature outputs between different sub-classes is significantly large while the variance within the same subclass is minimal. Finally, the feature extraction method incorporates prior knowledge that histopathological images with various magnifications belong to the same classification. Awan et al. [10] introduced a metric called Best Alignment Metric (BAM) to measure the shape of the glands in colon cancer. They showed a correlation between the glandular shape metric and class grade of the tumour class. Their model is based on a DCNN for detecting gland boundaries and a support vector machine (SVM) classifier is used for deciding the grade of cancer. Arvaniti et al. [8] presented a DL approach for automated gleason grading of prostate cancer tissue micro-arrays with hematoxylin–eosin staining. Their system

was trained using detailed Gleason annotations.

The work presented by [2] proposed a classification model for breast cancer images. They utilise a patch selection method to classify histopathology images using transfer learning. They first extract patches from WSIs then Efficient-Net is used to generate patch features which are then used to by SVM classifier for final classification. Kanavati et al. [45] introduced a deep learning model for identifying transbronchial lung biopsy (TBLB) WSIs as adenocarcinoma (ADC), squamous cell carcinoma (SCC), small-cell lung cancer (SCLC), or non-neoplastic. Their approach is made up of a CNN and a Recurrent Neural Network (RNN) to acquire patch predictions and aggregate patch predictions into a single WSI classification, respectively. Jiahui et al. [54] proposed that integrating pixel-level and image-level annotation can lead to even greater improvements. In computational pathology, this is problematic since the high resolution of WSIs makes end-to-end classification model training challenging. To address this, they created a hybrid supervised learning approach for gastric histopathology high-resolution images that included sufficient image-level coarse annotations as well as a few pixel-level fine labels. With the use of coarse image-level labels, this strategy can enhance generated pixel-level pseudo-labels when used for training patch models.

Due to challenges of achieving acceptable classification performance with few labelled samples, [112] proposed a deep transferable semi-supervised domain adaptation model (HisNet-SSDA) for classification of histopathology WSIs. Their method is based on information being transferred from a highly labelled source domain to a partially labelled target domain via semi-supervised domain adaptation. To begin, a pre-trained network known as HisNet is used to extract high-level characteristics from randomly selected patches in the source and target domains. The properties of the two domains are then matched using a multiple-weighted loss functions criterion with a new manifold regularisation term in semi-supervised domain adaptation. Finally, the estimated probabilities of the sampled patches are added together to get the final image-level categorisation. A multi-layer hidden conditional random fields (MHCRFs)-based cervical histopathology image classification (CHIC) model [53] is employed with a weakly supervised learning approach to detect good, intermediate, and poorly differentiated stages of cervical cancer. Their strategy begins with the extraction of deep learning features from histopathological image patches. The collected features are then utilised to produce patch-level classification probabilities via neural network, support vector machine, and random forest classifiers. Effective classifiers are then chosen to create unary and binary potentials. Finally, using the potentials generated, the MHCRF model predicts image-level classification results. The work introduced by Xiang et al. [115] indicated that WSI analysis can be performed efficiently by merging data at both local and regional levels. They expressed local information by auto-encoding the visual signals in each patch of WSI into a latent embedding vector, while regional information was represented by a down-sampled WSI with several scales. The WSI label is then predicted using a Dual-Stream Network (DSNet), which takes as input updated local patch embeddings and multi-scale thumbnail images. This input helps in training their model by using image-level annotations only.

Despite the success of single CNNs, several computer vision challenging problems (such as the limited availability of training images, high-level of noise, and high variability of the morphological architecture of region of interest in images) still persist. Multiple CNN models (ensemble models) are required to introduce diversity in learning and cope with complicated cases.

Due to the challenging problems stated above concerning histopathology images, researchers proposed the adoption of the ensemble approach. This approach is based on the combination of multiple DCNN models with different learning perspectives, which consequently improves diagnosis accuracy. For example, Chennamsetty et al. [20] proposed an ensemble of three CNNs trained on various pre-processing normalisation settings. The goal of this work was to demonstrate that no single architecture or pre-processing setting can deliver superior performance. The work introduced by Yang et al. [120] suggested an Ensemble of Multi-Scale Network (EMS-Net) CNN ensemble model for classifying hematoxylin-eosin stained breast histopathology images. EMS-Net can extract features at several scales by employing several pre-trained CNN models and selecting the best subset of fine-tuned deep models. Kassani et al. [46] developed an ensemble of deep learning models for automated binary categorisation of breast histopathology images. The suggested model is built using three pre-trained CNNs: VGG19, MobileNet, and DenseNet. The ensemble model was utilised to extract features, which were then inserted into a multi-layer perceptron classifier for the classification task.

Nguyen et al. [74] developed a feature concatenation and ensemble approach to combine numerous CNNs with varying depths and architectures to increase the accuracy of biomedical image classification, including images of cervical cancer. The proposed model comprises three pre-trained transfer learning models (Inception-v3, ResNet152, and Inception-ResNet-v2) plus a fourth model that operates as a multi-feature-extractors model. This feature extraction module concatenates the three feature maps taken from the three basic models into a larger feature vector. The four feature maps are produced using an ensemble learning approach (three from the base models and one from the multi-feature descriptor). Marami et al. [69] proposed an automated classification method for identifying the micro-architecture of tissue structures in breast histopathology images. Their proposed architecture is based on ensembling multiple inception networks which are trained using different data subset sampling and image perturbation. Their inception network is modified by using adaptive pooling, which increases the practical utility of their trained network, as it can be applied to images with minor scale changes from the input training images. Hameed et al. [36] introduced an ensemble model for the classification of non-carcinoma and carcinoma breast cancer histopathology images. They used different models based on pre-trained VGG16 and VGG19 architectures. Then they followed an ensemble strategy by taking the average of predicted probabilities. The work introduced in [30] proposed an ensemble of DCNNs for multi-class classification and textural segmentation of histopathology colorectal cancer tissues. The work presented by Xue et al. [117] introduced an Ensemble Transfer Learning (ETL) framework to classify cervical histopathology images that are well, moderate, or poorly differentiated. They used TL structures based on Inception-V3, Xception, VGG-16, and Resnet-50. Then, to improve classification performance, a weighted voting EL technique is implemented.

Guo et al. [34] proposed a hybrid CNN based on GoogLeNet to merge local and global information. They introduced a patch-level CNN to capture local information and an image-level CNN to generate global features. Both local classification scores and global classification scores are combined to pick the optimal match for each input image. They also introduced a bagging approach and a hierarchical voting strategy. To apply this, multiple models are trained initially using various data splitting and sampling methods, and then image classification is done via hierarchical (patch-level and image-level) voting. Similar work introduced by Zhu et al. [126] proposed a hybrid CNN for the classification of breast histopathology images by assembling

multiple compact CNNs. Their architecture is built using a local model, which captures local information from image patches extracted from the input image, and a global model, which takes a down-sampled version of the input image and extracts global information. The predictions generated from both models are then weighted to generate the final prediction. In the work introduced by Hongdou et al. [121], they proposed a parallel structured framework which uses CNN and RNN in a parallel fashion style for the classification of breast histopathology images. They extract image features using both networks, and then using a perceptron attention mechanism they merge the features extracted from the two networks. The work proposed by Vang et al. [108] presents a deep learning model that performs patch-level categorisation using Inception-V3. The model works by extracting patch-level predictions using Inception-V3, then these predictions are then inserted into an ensemble fusion architecture which includes majority voting, gradient boosting machine (GBM), and logistic regression. A fully convolutional auto-encoder is then used to identify the dominant structural patterns among normal image patches. To recognise and assess patches that do not share the characteristics of the normal population, a one-class support vector machine and a one-layer neural network are utilised.

All the mentioned work using ensemble learning has shown different methods to improve the performance of diagnosis using the standard ensemble approach which combines all the models in the ensemble architecture. However, they lack the usage of contextual information strategy which is of importance as it is used to build spatial dependencies among different image regions specifically with high-resolution histopathology images. This approach helps to improve the performance of diagnostic models.

3.3 Context-aware Methods for Histopathology Image Classification

In histopathology image analysis, the importance of learning contextual information using DCNN has been introduced for the image classification task. The contextual information aids in preserving the spatial dependencies of a particular image region to cover a large tissue region (i.e. the surroundings of a region). The results shown from different studies indicate that contextual information plays a vital role in reducing anomalies in heterogeneous tissue structures.

Ruqayya et al. [9] presented a two-stage context-aware technique consisting of two major steps: (1) a patch-based deep CNN based on ResNet-50 to extract significant features from image patches, (2) a separate SVM classifier to perform image-based classification using the features generated from overlapping patches. Their model is developed to acquire contextual information between image patches. Similarly, Ehteshami et al. [27] proposed a context-aware stacked CNN model to classify breast WSIs. They built their model in two stages: first, they trained a CNN to preserve the cellular-level information from image patches, and then placed a Fully Convolutional Network (FCN) on top of that to allow for the merging of global inter-dependency of structures to encourage predictions in neighbouring regions. Likewise, Huang et al. [42] proposed a deep fusion network to capture the spatial relationship among histology image patches. This is by adopting a residual network to learn visual features from cellular-level to large tissue organisation. Consequently, a deep fusion network has been developed to model the inconsistent construction of distinctive features over patches and rectify the predictions of the residual network. Yan et al. [119] proposed a hybrid model by integrating convolutional and recurrent

deep neural networks for breast cancer histology image classification. It considers the short-term and long-term spatial correlations between image patches using a Bidirectional Long Short-Term Memory (LSTM) network. This is done by extracting feature representations from image patches of a histopathology image and then feeding the extracted features into the bidirectional LSTM to preserve the spatial correlations among the feature representations.

The work presented in [113] developed a weakly supervised technique for the classification of WSIs of lung cancer. They used patch-based FCN for discriminatory block retrieval and also introduced context-aware feature selection and aggregation to generate an all-encompassing holistic WSI description. Shaban et al. [92] proposed a two-stacked CNN for colorectal image classification. This is done by integrating a larger context by using a context-aware neural network. To make a final prediction, the model transforms the local representation of a histopathology image into high-dimensional features and then combines the features by perceiving their spatial arrangement. Zhou et al. [125] introduced a new cell-graph convolutional neural network (CGC-Net) for grading of colorectal cancer. Their network transforms each large histopathology image into a graph, with each node represented by a nucleus within the input image and cellular associations denoted as edges among these nodes based on node similarity. The network uses local features of the nuclei and spatial dependencies of the nodes to enhance the model performance.

The work presented by Pati et al. [77] has suggested a hierarchical cell-to-tissue graph (HACT) representation to improve the structural description of histopathology tissue. Their method consists of two sorts of graph. First, a low-level cell-graph that shows cell morphology and interconnections. Second, a high-level tissue-graph that captures the morphological properties and spatial distribution of tissue sections. Moreover, their method captures cells-to-tissue hierarchies that integrate the relative spatial patterns of cells in relation to tissue distribution. A hierarchical graph neural network (HACT-Net) is also developed to translate HACT presentations into histopathological breast cancer subtypes. HistoGAN has been suggested by Xue et al. [118] to improve the categorisation of histopathology images. It employs conditional Generative Adversarial Networks (GAN) to generate realistic histopathology image patches based on class labeling. They devised a synthetic augmentation method that includes HistoGAN-generated synthetic image patches selectively rather than increasing the training set directly using synthetic images. The framework maintains the quality of synthetic augmentation by selecting synthetic images based on the reliability of their assigned labels and their feature similarity to actual labelled images. They showed that employing HistoGAN-generated images with selective augmentation increases classification performance considerably.

The work introduced by Li et al. [59] developed HCRF-AM (Hierarchical Conditional Random Field based Attention Mechanism) for gastric histopathology image categorisation. The HCRF-AM model consists of two components: an Attention Mechanism (AM) and an Image Classification module (IC). In the AM component, an HCRF model is developed to capture attention regions. The IC component trains a CNN using the provided attention regions and then uses an ensemble learning approach based on probability distribution to produce image-level results from the patch-level output of the CNN. Chen et al. [19] created an interactive WSI diagnosis method for thyroid frozen sections based on pathologists' selected suspicious spots. Their technique is based on producing feature patterns for suspect locations by acquiring and fusing patch features using DNNs. The feature representations are then used to evaluate four classifiers and three supervised hashing algorithms for region classification and retrieval. Sharma et al. [95] proposed Cluster-to-Conquer (C2C),

an end-to-end architecture that separates a WSI's patches into k-groups, selects k's patches from each grouping for training and applies an adaptive attention method to provide final slide prediction. They demonstrated that partitioning a WSI into clusters helps model training by presenting it to a range of discriminative properties extracted from patches. The study published in [97] proposed a Deformable Conditional Random Field (DCRF) model to learn the offsets and weights of neighbouring patches of WSIs in a spatially adaptive manner. They also employed adaptive adjusted offsets in a WSI to locate patches with more robust feature representations rather than overlapping patches.

Shao et al. [94] developed a framework named correlated MIL to solve the issue of ignoring the connection between distinct instances. They designed a Transformer-based MIL (TransMIL) which focused on both morphological and contextual information. The proposed TransMIL can handle unbalanced/balanced and binary/multiple classifications. Another work proposed by [18] provided a broad approach for autonomously diagnosing various types of WSIs using unit stochastic selection and attention fusion. A single unit on a histopathology slide could be a cell on a cytopathology slide. Their method starts with training a unit-level CNN to achieve two goals: building feature extractors for the units and determining the non-benign probability of each unit. Then, based on CNN's findings, they use a stochastic selection approach to choose a small segment of non-benign units known as Units Of Interest (UOI). After then, the attention mechanism is employed to combine the UOI representations into a fixed-length description for the WSI's diagnosis. Campanella et al. [15] introduced a DL technique based on MIL that uses the provided diagnoses only as labels for training, avoiding costly and time-consuming pixel-by-pixel manual annotations. MIL is utilised in the built framework to train DNNs, resulting in tile-level feature representation. These representations are then used by RNN to integrate the information throughout the whole slide and provide the final classification result. The study reported in [111] proposed the use of Graph Convolutional Networks (GCNs) to model the spatial organisation of cells as a graph in a weakly-supervised approach for grade classification in tissue microarrays (TMA). Changjiang et al. [123] introduced a framework that integrated features from different magnifications of WSIs to achieve classification and localisation of colorectal cancer using only global labels. Haoyuan et al. [17] presented the IL-MCAM framework for colorectal histopathology image classification, which is based on attention processes and interactive learning. The framework is divided into two stages: automatic learning (AL) and interactive learning (IL). The AL stage contains three independent attention mechanism channels and CNNs to extract multiple channel features for classification. For the IL stage, the system employs an interactive approach to continuously include misclassified images into the training set, hence improving the model's classification performance.

The context-aware approach proved to be improving the performance of DL-based models for histopathology image diagnosis. However, nowadays, it is crucial to enhance trust in models by introducing a measure of confidence in the developed models. Therefore, clinical practice needs to present models that measure the uncertainty of samples' predictions and as well aid in increasing the level of reliability of automated diagnosis systems.

3.4 Uncertainty Quantification for Medical Image Analysis

As an important initial step to explainable classification and segmentation models, it is required to measure the uncertainty of the predictions obtained by machine learning and deep learning methods [1]. A few recently proposed image segmentation and classification approaches have adopted an uncertainty quantification component for medical image analysis. For example, Simon et al. [33] used a measure of uncertainty in a CNN-based model using an instability map to highlight zones of equivocalness. Fraz et al. [28] proposed a framework for micro-vessel segmentation with an uncertainty quantification component for H&E stained histology images. A calibration approach [62] has been designed in a way to preserve the overall classification accuracy as well as improve model calibration. It provides an Expected Calibration Error (ECE), which is a common metric for quantifying miscalibration. Their approach can be easily attached to any classification task and showed the ability to reduce calibration error across different neural network architectures and datasets. Mobiny and Singh [72] proposed a Bayesian DenseNet-169 model, which can activate dropout layers during the testing phase to generate a measure of uncertainty for skin-lesion images. They investigated how Bayesian deep learning can help the machine–physician partnership perform better in skin-lesion classification. In another research, Raczkowski et al. [79] proposed an accurate, reliable and active Bayesian network (ARA-CNN) image classification framework for classifying histopathology images of colorectal cancer. The network was designed based on residual network and variational dropout.

The methods presented in this section lack either standard ensemble or the elasticity of ensemble of multiple DNN models based on uncertainty measures. A dynamic ensemble based on a measure of confidence in image predictions is crucial to increase the trust in an automated diagnosis system by (1) making sure that only models with a pre-defined degree of confidence contribute to the final image prediction, and by (2) flagging out cases that are hard to classify confidently by the model for further inspection (excluding the untrustable samples from the perspective of uncertain predictions). Table 3.1 presents characteristics of all the medical image analysis methods presented in this chapter.

TABLE 3.1: Characteristics of histopathology image classification methods.

Method	Object type	Feature type	Multi-scale	Ensemble	Uncertainty	Accuracy
Two-Stage CNN [73]	Breast	Local & global	No	No	No	95%
Hybrid CNN [34]	Breast	Local & global	No	Yes	No	87.5%
Hierarchical CNN [51]	Breast	Contextual	No	No	No	96%
Inception and ResNet [109]	Breast	Local	No	No	No	97.5%
Multi-layered framework [35]	Breast	Local	Yes	No	No	94.71%

Continued on next page

Table 3.1 – continued from previous page

Method	Object type	Feature type	Multi-scale	Ensemble	Uncertainty	Accuracy
PDRH [98]	skeletal muscle & Lung	Local	No	No	No	97.49%
Parallel CNN-RNN [121]	Breast	Local	No	Yes	No	89%
multiple compact CNNs [126]	Breast	Local & global	No	Yes	No	87.2%
Ensemble Fusion Architecture [108]	Breast	Local	No	Yes	No	87.5%
Patch-level tumour classification [114]	Lymph node metastases & Prostate	Local	No	No	No	84.3%
PBC [85]	Breast	Local	No	No	No	90%
Patch screening method [61]	Breast	Local & global	Yes	No	No	88.89%
Pattern Mining Autoencoder [58]	Breast	Local	No	No	No	76%
Liver Cancer classifier [104]	Liver	Local	No	No	No	98%
IRRCNN [4]	Breast	Local	No	No	No	96.76%
GCNs [111]	Prostate	Contextual	No	No	No	96.59%
DenseNet121-AnoGAN [68]	Breast	Local	No	No	No	99.38%
HACT-Net [77]	Breast	Contextual	No	No	No	62.89% (F1-Score)
HistoGAN [118]	Cervical & Lymph node metastases	Contextual	Yes	No	No	[94.8%, 82.1%]
Pa-DBN-BC [39]	Breast	Local	No	No	No	86%
Multi-resolution model using MIL [55]	Prostate	Local & global	Yes	No	No	92.7%
HCRF-AM [59]	Gastric	Contextual	No	Yes	No	91.4%
Super-pixels nuclei detection [101]	Cervical	Local	No	No	No	95.97%

Continued on next page

Table 3.1 – continued from previous page

Method	Object type	Feature type	Multi-scale	Ensemble	Uncertainty	Accuracy
Multi-task Xception [56]	Breast	Local	No	No	No	94.8%
BAM-DCNN-SVM [10]	Colorectal	Local & global	No	No	No	91%
Unit Stochastic Selection [18]	Thyroid frozen, Colonoscopy, and Cervical	Contextual	No	No	No	[98.3%, 85.1%, 83%]
Global labels framework [123]	Colorectal	Contextual	Yes	No	No	94.6%
MobileNet gleason grade classification [8]	Prostate	Local & global	No	No	No	75%
Context CNN [92]	Colorectal	Contextual	No	No	No	95.7%
CGC-Net [125]	Colorectal	Contextual	No	No	No	97%
Inception Ensemble [69]	Breast	Local	No	Yes	No	84%
VGG Ensemble [36]	Breast	Local	No	Yes	No	95.29%
Tumour detection Ensemble [30]	Colorectal	Local	No	Yes	No	96.16%
MIL + RNN [15]	Breast & Prostate	Contextual	No	No	No	[99.1%, 93% (AUC)]
EfficientNet + SVM [2]	Breast	Local & global	No	No	No	96.99%
Patch fusing + classification [19]	Frozen section thyroid	Contextual	No	No	No	96.1%
Hybrid supervised learning approach [54]	Gastric	Local	No	No	No	97.05%
Cluster-to-Conquer (C2C) [95]	Breast & gastrointestinal disease	Contextual	No	No	No	[91.12% (AUC), 86.2%]
HisNet-SSDA [112]	Colon	Local	No	No	No	94.32%
MHCRFs [53]	Cervical	Local	No	No	No	93%
DSF-Net [42]	Breast	Contextual	No	No	No	95%

Continued on next page

Table 3.1 – continued from previous page

Method	Object type	Feature type	Multi-scale	Ensemble	Uncertainty	Accuracy
Hybrid LSTM [119]	Breast	Contextual	No	No	No	91.3%
DSNet [115]	Breast & Lung	Local & global	Yes	No	No	[93%, 69.6%]
CNNs Ensemble (different Normalisation settings) [20]	Breast	Contextual	No	Yes	No	87%
EMS-Net [120]	Breast	Local	Yes	Yes	No	91.75%
TransMIL [94]	Breast & Lung & Renal	Contextual	No	No	No	[94.66%, 88.37%, 88.35%]
VGG19, MobileNet, and DenseNet Ensemble [46]	Breast	Local	No	Yes	No	95%
Three CNNs and multi-feature-extractors model [74]	Cervical	Local & global	No	Yes	No	93.04%
ResNet-50 + SVM [9]	Breast	Contextual	No	No	No	83%
Wide ResNet + FCN [27]	Breast	Contextual	No	No	No	81.3%
DCRF model [97]	Colorectal	Contextual	No	No	No	94.68%
IL-MCAM [17]	Colorectal	Contextual	No	No	No	98.98%
MILD-Net [33]	Colon	Local & global	No	No	Yes	91.4% (F1-Score)
FABnet [28]	Oral	Local & global	No	No	Yes	96.3%
Expected Calibration Error (ECE) Method [62]	4 medical datasets	Local	No	No	Yes	[94.1%, 88.41%, 87.96%, 71.06%]
Bayesian DenseNet-169 [72]	Skin lesion	Local	No	No	Yes	90%

Continued on next page

Table 3.1 – continued from previous page

Method	Object type	Feature type	Multi-scale	Ensemble	Uncertainty	Accuracy
Bayesian network (ARA-CNN) [79]	Colorectal	Local	No	No	Yes	92.44%
ETL framework [117]	Cervical	Local	No	Yes	No	97.03%

3.5 Applications of Deep Learning for Medical Image Analysis

DL-based methods have been very effective by introducing applications for multiple purposes that showed tremendous performance. It is important to highlight here two types of applications that are used for medical image analysis: diagnosis applications and grading applications. Diagnostics of medical images are considered the basic level of differentiation between different class boundaries. For instance, cancer images could be divided into benign and malignant images. The DL-based automated diagnosis applications which use this type of approach work on extracting features from input images and then these features are used to differentiate between the morphological structures of tissues in benign samples and malignant ones. Similarly, *MCUa* diagnosis model [91] works in differentiating between different classes of breast cancer samples: normal tissue, benign lesion, in situ carcinoma, and invasive carcinoma (multi-class classification task). *MCUa* is considered as a diagnosis application which can be applied to either binary or multi-class classification tasks. On another level, as a more challenging task, DL-based automated grading applications are mainly based on differentiating between different grades of cancer. This type of grading application is important to identify suitable treatment plans based on the level of cancer and how far is the cancer spreading in a particular body's organ. For instance, the *3E-Net* grading model [90], which is our first contribution in the thesis, works on differentiating between (i.e. classifying) different grade levels of invasive breast carcinoma samples (three grades).

3.6 Discussion

The related work mentioned in this chapter showed different methods to improve classification performance. Single architectures have been proposed to introduce a simple design for a deep learning model. However, these single architectures lack learning diversity, which is paramount for improving classification performance to higher levels. Single architectures usually show poor performance when compared to standard ensemble architectures [35, 51, 73, 98, 104, 109, 114]. The existence of multiple models in an ensemble structure with various learning strategies aims to have different analysis perspectives for the features extracted from input data. This consequently helps in generating more accurate classification decisions and enhances performance. Ensemble learning is one of the strategies used to combine predictions from different learners. This method aims to introduce diversity in learning. However, combining all learners in an ensemble architecture may show sub-optimal

results when we have some learners that are less confident about a particular prediction [20, 30, 69, 74, 120, 126]. Therefore, the elastic ensemble is an effective strategy for using only the most confident learners for the final model prediction.

A drawback noticed in single and ensemble architectures is that they lack multi-level learning for contextual information, which builds spatial dependencies among different image regions. This type of information helps to build a better contextual vision for the whole image instead of working on separate image regions. The context-based methods mentioned in this chapter showed improvement in the performance. However, they lack an uncertainty quantification method that is crucial for clinical practice and to show how confident a generated prediction is. Although uncertainty quantification-based methods mentioned in this chapter showed the importance of presenting uncertainty measures, they either lacked the usage of elastic ensemble which is pivotal for model actionability (decision making for uncertain images) or applying contextual information for the input medical samples.

The drawbacks presented in the related work mentioned in this chapter motivated us on taking advanced steps and reduce the gap of the challenges presented in medical image analysis. This has been done by providing effective automated grading/diagnosis applications which feature: (1) usage of diverse learning strategies for input images by utilising multiple image scales, (2) diversity in the extraction of image features using multiple pre-trained CNNs, (3) applying multi-level contextual information for different image regions, and (4) the usage of uncertainty quantification using different methods.

3.7 Summary

In this chapter, we reviewed different histopathological image analysis methods from different perspectives (i.e. single architectures, ensemble architectures, context-based architectures, and uncertainty-based methods). The challenges we found in the related work as described in Section 3.6 motivated us to develop robust and effective systems that are beneficial for clinical practice. First, we developed *3E-Net* model for classifying invasive breast carcinoma images into different grades. *3E-Net* includes the development of elastic ensemble of deep learning models which learn contextual information among input image patches and uncertainty quantification using Shannon Entropy. Second, we developed *MCUa* model for the classification of breast cancer histopathology images. This model includes the development of advanced multi-level context-aware models for learning multi-level contextual information among image patches and a flexible uncertainty quantification method using MC dropout. *3E-Net* and *MCUa* have been used for two different applications: grading and diagnosis, respectively. They filled the gap of introducing models that have diversity in contextual learning and usage of uncertainty measure. Moreover, they introduced robust design and actionability by having elasticity in the ensemble architecture developed and an exclusion mechanism that excludes images based on their high uncertainty. Third, we developed an automated actionable method for optimising deep learning models. This method aids in introducing automated actionability for the developed deep learning models when dealing with uncertain samples. In the next chapter, we present our first contribution *3E-Net* model [90]. The chapter includes *3E-Net* model's methodology and the comprehensive experimental study conducted.

Chapter 4

3E-Net: Entropy Elastic Ensemble Model for Classifying Grades of Invasive Breast Carcinoma Images

In the previous chapter, we reviewed the literature work conducted in the field of medical image analysis and we discussed the drawbacks of the proposed methods in the literature and how those drawbacks motivated us to develop the contributions presented in this thesis. In this chapter, our first contribution, 3E-Net for classification of invasive breast carcinoma histopathology image grades is explained in detail. The chapter includes the datasets used, the model methodology and the comprehensive experimental study conducted. Findings reported in this chapter have been published in [90].

4.1 Overview

Automated grading systems using deep convolution neural networks (DCNNs) have proven their capability and potential to distinguish between different breast cancer grades using digitised histopathological images. In digital breast pathology, it is vital to measure how confident a DCNN is in grading using a machine-confidence metric, especially with the presence of major computer vision challenging problems such as the high visual variability of the images. Such a quantitative metric can be employed not only to improve the robustness of automated systems, but also to assist medical professionals in identifying complex cases. In this chapter, we present Entropy-based Elastic Ensemble of DCNN models (3E-Net) for classifying grades of invasive breast carcinoma microscopy images which provides an initial stage of explainability (using an uncertainty-aware mechanism adopting entropy). Our model has been designed in a way to (1) exclude images that are less sensitive and highly uncertain to our ensemble model, and (2) dynamically grade the non-excluded images using the certain models in the ensemble architecture. We evaluated two variations of 3E-Net on an invasive breast carcinoma dataset and we achieved grading accuracy of 96.15% and 99.50%.

The chapter is organised as follows. Section 4.2 presents detailed introductory of the background and the developed work. Section 4.3 discusses, in detail, the architecture of our developed 3E-Net model. Section 4.4 describes the dataset used, our experimental results, and discusses our findings. Section 4.5 discusses the impact of 3E-Net and the motivation for the upcoming contribution.

4.2 Introduction

Breast cancer is a major public health concern around the world, where its prevalence rate is the second-highest rate for women (excluding lung cancer) among all forms of cancer [99]. The study of histopathological images remains the most commonly used tool for diagnosing and grading breast cancer, even with the substantial advances in medical science. Early diagnosis can dramatically improve the effectiveness of therapy. The symptoms and signs of breast cancer are numerous, and the diagnosis encompasses physical analysis, mammography, and confirmed by core needle biopsy tissue (CNB) from the suspicious breast area. The sample tissue extracted from the CNB process demonstrates the cancerous cells and the grade of cancer associated with them. Pathologists typically look for certain characteristics that can help them predict disease prognosis during the visual inspection of the biopsy specimen of the tissue (i.e. what is the likelihood of cancer spreading and growing?).

For tumour grading, pathologists usually use the Nottingham scoring system that depends on morphological changes including glandular/tubular formation, nuclear pleomorphism, and mitotic count [3]. Due to the high visual variability of the samples in terms of their morphological structure, visual qualitative grading assessment is a time-consuming and laborious process [83]. In the context of histopathological image analysis, grading of invasive breast cancer provides many challenging problems. First, there are variations in subjective criterion evaluation between observers when it comes to diagnosis/grading. Second, it is difficult to capture the proper combination of features and the morphological heterogeneity within the tumour regions [50, 83]. Such challenges usually lead to substantial effort and exhaustive manual qualitative study from pathologists. Thanks to computational pathology which helped in alleviating this burden in recent years. In computational pathology, deep learning (DL) approaches have made tremendous progress and achieved outstanding results, leading many researchers to provide automated and unbiased solutions for several different histopathological image analysis applications including breast cancer grading and tissue classification [65]. Deep convolution neural networks (DCNNs) are the most commonly used type of DL approaches, demonstrating outstanding performance in extracting image salient features for the different computational pathology applications [96].

Despite the prevalence of DCNNs in several histopathological image analysis applications including grading, the ability of a single DCNN model to obtain discriminatory features is constrained and usually results in sub-optimal solutions [9, 42, 73]. As a consequence, an ensemble of DCNN models has been proposed to conserve the description of histopathological images from recognisable perspectives to a more precise classification [120]. More importantly, to the best of our knowledge, previously proposed DCNN-based diagnosis tools lack a preliminary measure of uncertainty, which is an initial important step towards an explainable computational pathology. Developing an uncertainty quantification component can contribute to the recognition of multiple regions of ambiguity that may be clinically instructive. It also allows pathologists and medical professionals to rate images that should be prioritised for pathology annotations. Despite the existence of DCNN models and their high potential in minimising the workload burden from pathologists, a limited number of microscopy images would require pathologists' assistance.

In this chapter, we introduce a novel Entropy-based Elastic Ensemble of DCNN

models (*3E-Net*) [1] for the automated classification of grades of breast carcinoma using histopathological images. *3E-Net* has an elasticity capability in allocating different classifiers (e.g. DCNNs) for each particular image. To put it differently, the term "elastic" implies that the number of classifiers selected in the ensemble architecture to contribute to the final image prediction can differ for each image. Our model is supported by an uncertainty quantification component which helps pathologists to refine annotations for developing more robust DCNN models that can meet their needs. Conversely, in this work, we first extract patches from the input image. Then, we designed a patch feature extractor network (i.e. pre-trained and fine-tuned DenseNet-161 [41]) to learn salient features from image patches. The extracted feature maps are then fed into multiple image-wise CNN models which are designed to capture multi-level spatial dependencies among the patches. Eventually, an uncertainty-measure ensemble-based component is introduced to select the most certain image-wise models for the final image grading. The performance of our model is evaluated on the Breast Carcinoma Histological Images dataset [26], which consists of 300 high-resolution hematoxylin-eosin (H&E) stained breast histopathological images, divided into three invasive grades.

The contributions of this chapter are summarised as follows: (1) a novel uncertainty-aware component adapted by an entropy formula to measure how confidence DCNN models of our automated breast cancer classification system on input images. This uncertainty-aware mechanism assists pathologists in identifying the complex and corrupted images which are hard to be graded by automated systems; (2) an automatic exclusion of poor histopathological images for manual investigation; (3) a new elastic ensemble mechanism is developed using most certain DCNN models, where each input image will be classified by a pool of models, but only confident ones contribute toward the final prediction using a dynamic ensemble modeling mechanism; and (4) quantitative and qualitative analysis study have been conducted using our automated system on breast carcinoma dataset. To the best of our knowledge, this is the first attempt to introduce an entropy-based uncertainty quantification metric to achieve an elastic-based ensemble of DCNN models in automated grading of invasive breast carcinoma from histopathological microscopic images.

4.3 *3E-Net* Model

In this section, we describe, in detail, our *3E-Net* model. Given a histopathological image section with a high-resolution (1280×960 pixels) as an input, the main target is to classify the image into one of three invasive grades of breast cancer: grade 1, grade 2, or grade 3. As illustrated by Figure 4.1, our model consists of several DCNNs which are designed and implemented based on the input size of the image and the number of patches extracted from the image. First, the input image is divided into many smaller patches which are then inserted into a pre-trained and fine-tuned DCNN which acts as patch-wise feature extractor network. Second, the extracted feature maps are fed into image-wise networks which encode different levels of contextual information. As a final and prominent step, the final image predictions (i.e. grades) from image-wise models are then inserted into an elastic ensemble stage which is mainly based on measuring the uncertainty of predictions in each model. This uncertainty measure of predictions is designed using the Shannon entropy [93] which measures the level of randomness in the model's final prediction. More precisely, Shannon entropy values of different models in our ensemble architecture were

[1] The code is available at https://github.com/zakariaSenousy/3E-Net-Model.

used to select the most accurate/certain models (i.e. the models which have a small entropy value) to improve the elasticity capability of *3E-Net* in allocating different classifiers and improving diversity. Using a pre-defined threshold, only models with a high degree of certainty are included in the final elastic ensemble of the image.

4.3.1 Patch-wise Feature Extraction

Due to the scarcity of annotated training data in the medical field, transfer learning [122] has emerged as a prominent approach to cope with the problem. Transfer learning is a mechanism that uses machine learning models (e.g. CNNs) which are pre-trained on large datasets (e.g. large-scale images of ImageNet dataset) to be adapted and used in different domain-specific tasks (e.g. breast cancer grading). In such mechanisms, the network configuration is preserved, and the pre-trained weights are used to configure the network for the new domain-specific task. During the fine-tuning stage, the initialised weights are continuously updated, allowing the network to learn hierarchical features relevant to the desired task. Fine-tuning is effective and robust for various tasks in the medical domain [9, 119, 120].

As stated earlier, the patch-based paradigm proved to be effective when it comes to high-resolution histopathological images [9, 73, 119, 120]. In this work, we utilise a pre-trained and fine-tuned DenseNet-161 to act as feature extractor networks for image patches. DenseNet-161 has demonstrated a superb performance for ILSVRC ImageNet classification task [24]. Moreover, DenseNet-161 has shown a great success in several histopathological image analysis pipelines [16, 43, 49, 57, 60, 76, 82, 120, 124]. In order to supply the patch-wise feature extractor network with image patches, we extract a number of patches k based on the following equation [73]:

$$k = \left(1 + \left\lfloor \frac{W - w}{s} \right\rfloor\right) \times \left(1 + \left\lfloor \frac{H - h}{s} \right\rfloor\right) \tag{4.1}$$

where W and H are width and height dimensions of the input image, respectively. While, w and h are width and height dimensions for the image patch, respectively and s is the stride used over the input image.

To improve variety (in the training data) and alleviate overfitting for the patch-wise feature extractor network, we extracted and used partially overlapped patches. Furthermore, we applied data augmentation techniques by transforming each patch using rotation and reflection operations. For example, random color alterations introduced by [66] has been applied to each patch as it aids in minimising the visual diversity of the patches. Our model learns rotation, reflection, color invariant characteristics, and makes pre-processing color normalisation [67]. The patch-wise feature extractor network is then trained using categorical cross-entropy loss based on image-wise labels. The loss equation is defined as:

$$\mathcal{L}(y - \hat{y}) = -\sum_{i=1}^{c} y_i \log \hat{y}_i \tag{4.2}$$

where y_i and $\hat{y}_i$ represent the ground truth label and the prediction of each class i in c classes, respectively.

4.3.2 Image-wise Grading

Once the feature extraction is accomplished, feature maps are fed into multiple image-wise networks to encode multi-level contextual information. The main purpose of the image-wise network is to grade images based on local and contextual

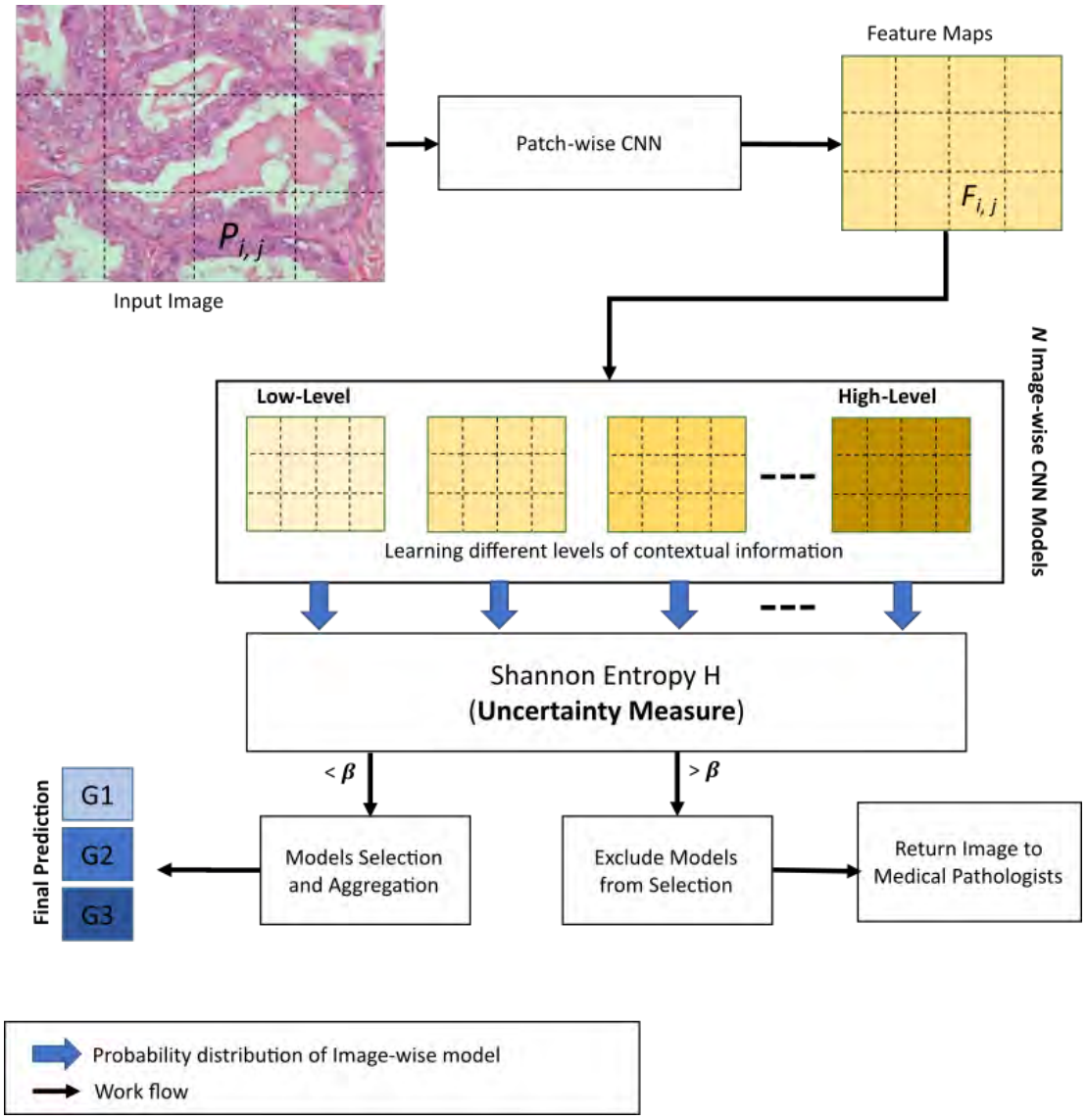


FIGURE 4.1: Overview of *3E-Net*. The model starts by taking a histopathological image section as input. Several small patches are extracted from the image where $P_{i,j}$ is one of the extracted patches. All patches are then fed into a patch-wise CNN for feature extraction, where $F_{i,j}$ is one of the extracted feature maps. Feature maps are then inserted into N image-wise CNN models to learn multiple levels of spatial dependencies information. Finally, Shannon entropy H is adopted in our uncertainty-aware component to measure the sensitivity of the input image to the N image-wise models. According to a pre-defined threshold β, the most certain models were selected for final grading prediction. In case of having zero certain models, the input image is returned to medical professionals for manual exploration and further investigation.

features captured from image and spatial dependencies information between different patches, respectively.

During the training stage of an image-wise network, we extract non-overlapping patches from the input image, where they are used to form newly concatenated feature maps that are designed based on neighboring feature maps only. This criterion

helps in building the intended contextual information. In our model, we build various image-wise networks that are based on multi-levels of contextual information. Each patch in the image has its own feature map. The number of image-wise network models depends on the number of feature maps extracted from the image and the possible formed shapes of neighbor feature maps. The contextual levels have low-level context which builds contextual feature maps among 2 original neighboring feature maps only, and high-level context builds contextual feature maps among all the original feature maps extracted from the image. For instance, having q feature maps extracted from the input image helps in generating image-wise models which learn contextual information among 2 feature maps (low-level) to q feature maps (high-level). Furthermore, for each level of contextual information (except for the highest level), a number of image-wise models can be generated based on different shapes of the neighbor feature maps. The formation and concatenation of any two or more feature maps can have different shapes. Likewise in the patch-wise network, the data augmentation process is applied to dataset images by applying rotation, reflection, and color alterations. Also, categorical cross-entropy loss is used in the training process against the corresponding image-level labels.

Image-wise CNN is composed of two blocks for context-aware feature learning. Each block has two 3×3 convolutional layers followed by a 2×2 convolution layer with a stride of 2 for down-sampling. Batch normalisation and ReLU activation function were attached after each layer. Batch normalization helps to stabilise the training process and improve convergence, while the ReLU activation function replaces negative activations with zeros, making the network more robust. The first block uses 64 channels, while the second block uses 128 channels. The number of channels represents the number of filters used in the convolutional layers. A 1×1 convolutional layer is used after the feature learning blocks and before the classifier to obtain the spatial average of feature maps. As a final block for classification, the network ends with 3 fully connected layers and a log softmax classifier. The softmax activation function is defined as:

$$S\left(z_i\right) = \frac{e^{z_i}}{\sum_j^c e^{z_j}} \tag{4.3}$$

where z_i represents output element i of the last fully connected layer.

4.3.3 Elastic Ensemble using Uncertainty Quantification

In this section, we describe our elastic ensemble of the constructed image-wise models. As a crucial step in this work, we transform the standard ensemble-based model into an elastic ensemble model which dynamically selects models based on the uncertainty of models as a measuring factor. In other words, for each image, a dynamic number of models is selected and combined towards the final image prediction. To measure uncertainty for our ensemble model, we adopted Shannon entropy for each image-wise model. Shannon entropy is a mathematical concept that measures uncertainty or information content in a probability distribution. This method is often used to quantify uncertainty or randomness in deep learning models. A high entropy value indicates a more uncertain or random prediction, while a low entropy value indicates a more certain or precise prediction. It can be used to measure the diversity of the model's output, making it a useful tool for improving reliability of Deep learning models. The formula for Shannon entropy is represented as:

$$H(X) = H(p_1, \ldots, p_c) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{4.4}$$

where $H(X)$ represents Shannon entropy for input image X and $p_1, \ldots, p_c$ is probability distribution for image X on c class categories.

During the testing stage, the input image is classified using all the image-wise models in an ensemble-based model. Each model generates the grade classification of the image in the form of a probability distribution for c class categories. Then, these probability distributions are evaluated using Shannon entropy (based on an uncertainty threshold value (β)) to measure uncertainty. According to the calculated uncertainty measure, a dynamic number of image-wise models will be selected for each image.

The selection process of image-wise models in the elastic ensemble process works by comparing the Shannon entropy measure evaluated for a particular model against a pre-defined threshold value β, as defined in the experimental study. If the entropy value is less than β, then the model will be chosen and included in a list of chosen models for a particular image. In the end, each image in the dataset should have a dynamic number of chosen models to produce the final prediction. In case of having images with zero chosen models, we prioritise these images for pathology annotating by medical professionals. After selecting the most certain image-wise models, the class predictions of these models are aggregated to produce the final class prediction distribution.

Algorithm 1 provides a detailed description of *3E-Net* model. The input image is divided into smaller patches. Then, using patch-wise CNN, many feature maps are extracted. These feature maps are then inserted into image-wise CNN models. Each image-wise model produces a probability distribution of the input image. In the end, the Uncertainty-aware component is utilised to measure the level of uncertainty for each image-wise model's prediction. The models with uncertainty values less than a threshold β are chosen and their predictions are aggregated for final classification grade $\hat{y}$. If the input image has no chosen models, medical professionals are involved in the final grading decision.

Algorithm 1: *3E-Net* Model

Input: histopathological Input Image X
Output: Image Label $\hat{y}$ or Uncertainty Decision

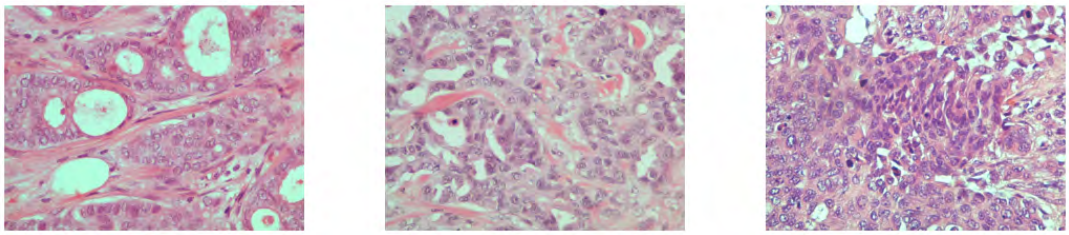
/* Image X is inserted into a function to extract smaller patches represented as $P_{m,n}$ where m and n are the patch row and column index in the image X, respectively. */
$P_{m,n} = PatchExtraction(X)$
/* Extract feature maps from image patches */
for $i \in m$ **do**
    for $j \in n$ **do**
        $F_{i,j} = PatchWiseFeatureExtraction(P_{i,j})$
V = [ ] // Empty List to store models' predictions
/* Insert Feature maps $F_{m,n}$ into N Image-wise Models */
for $i \in N$ **do**
    /* Image-wise Models learn different levels of contextual information */
    $Pred_i = ImageWiseModel_i(F_{m,n})$
    $V.append(Pred_i)$
/* Elastic Ensemble using Shannon Entropy */
selectedModels = [ ] // Empty list to store the most certain models' predictions
for $i \in N$ **do**
    $UncertaintyValue = -\sum_{i=1}^{c} V_i \log_2 V_i$ // where c is the number of classes
    /* Check if Uncertainty measure is less than a pre-defined threshold β */
    if $UncertaintyValue < \beta$ **then**
        /* Append prediction of the model which has small uncertainty */
        $selectedModels.append(V_i)$
/* Check if a dynamic number of models are chosen for final prediction */
if $selectedModels \neq 0$ **then**
    /* Aggregate the probability distributions of selected models and produce the final image grade label */
    $G = Aggregate(selectedModels)$
    $\hat{y} = \arg\max G$
/* If no models are chosen, this means all models in the ensemble architecture are uncertain about the final grade of the input image */
else
    /* Uncertainty decision */
    Exclude image from grading
    Return image to medical pathologists

4.4 Experimental Study

We evaluated the performance of our work on the Invasive Breast Carcinoma dataset. The utilised dataset in this experimental study has 300 images which all are used for training the ensemble model using 5-fold cross-validation. Cross-validation enables us to overcome the limited availability of annotated images, making sure that the model is well-trained. For training patch-wise networks, we used microscopy patches extracted from training images. These patches are augmented using rotation, flipping, and colourisation methods. Similarly, in image-wise networks, the same training process is conducted, but using the image-level dataset instead of patches. In the experimental study, we designed and implemented two standard ensemble models. First, the baseline ensemble model which has DenseNet-161 as the patch-wise feature extractor CNN will be denoted by Standard Ensemble Model (Version A). Second, we applied a modification by using the patch-wise CNN introduced in [73] as the feature extractor of the ensemble model. The modified ensemble model will be denoted by Standard Ensemble Model (Version B). Finally, our contribution has two *3E-Net* models: *3E-Net* Version A & *3E-Net* Version B, where we apply elastic ensemble approach to the standard ensemble models.

4.4.1 Dataset Description

Breast carcinoma histological images [26] were used for this work. The dataset contains cases of breast carcinoma histological specimens collected in the department of pathology, "Agios Pavlos" General Hospital of Thessaloniki, Greece. The dataset is composed of 300 H&E stained breast histopathological microscopy sections with the size of 1280 $\times$ 960 pixels. The dataset is mainly categorised into three grades of invasive carcinoma: grade 1, grade 2, and grade 3 (See Figure 4.2).



(a) Grade 1 (b) Grade 2 (c) Grade 3

FIGURE 4.2: Three H&E stained breast histopathological microscopy images from different invasive carcinoma grades.

The categories are divided as 107 images for grade 1, 102 images for grade 2, and 91 images for grade 3. These images are associated with 21 different patients with invasive ductal carcinoma of the breast. The image frames are from tumour regions taken by a Nikon digital camera connected to a compound microscope with a 40X magnification objective lens.

4.4.2 Hyperparameter Settings

As we have DenseNet-161 as the patch-wise feature extractor of the baseline ensemble model (Standard Ensemble Model (Version A)), we extracted patches of size 224 $\times$ 224 from the input image. Consequently, a number of 20 non-overlapped patches can be generated (where the original size of the input image is 1280 $\times$ 960) to extract

high-level contextual information. However, due to the limited GPU memory, we down-sampled the input images to a smaller scale of 896 $\times$ 672.

For training data extraction, we set the stride to s = 112 to extract partially overlapped patches for both versions (A & B). This stride value helps in increasing the training patch samples for patch-wise CNN and prevents the network from overfitting. We applied data augmentation by rotating the training patches by 90 degrees with horizontal and vertical flipping. To fine-tune the patch-wise CNN for Standard Ensemble Model (Version A) to our grading task, we modified the number of output neurons from 1000 to only 3 (as we have three grades). We used Adam optimiser [48] for minimising the cost function and we set the learning rate to 0.0001 for 5 training epochs and batch size to 32 for both patch-wise CNNs in versions A & B.

The extracted feature maps from patch-wise CNN are then inserted into image-wise models. For training image-wise model, we extracted non-overlapped patches from the new image scale giving us 12 patches by using s = 224. This means that we have a total number of 12 feature maps represented as a matrix of size (3 $\times$ 4) (as shown in Figure 4.1) to be used for the training process of image-wise models. Different levels of contextual information have been learned by combining all the original feature maps to form multi-level contextual feature maps. For example, the lowest-level contextual feature maps are generated by combining 2 neighboring feature maps while the highest-level contextual feature maps are generated by combining the 12 feature maps of the image. As mentioned earlier, different shapes of neighbor feature maps can be generated from each contextual level (except for the high-level as we combine all the 12 feature maps). Once the different levels of contextual feature maps are constructed, a number of DCNNs will be set up to learn the multi-level contextual information. This results in an arbitrarily chosen number of 17 image-wise models to form our ensemble architecture. Image-wise CNNs are trained on augmented image-level samples by applying rotation of 180 degrees with flipping. The remaining settings are the same as patch-wise CNN except that each image-wise CNN is trained for 10 training epochs and a batch size of 8.

Finally, we design and implement an elastic ensemble approach (*3E-Net* Versions A & B) for the standard ensemble models. This is accomplished using Shannon entropy to measure the uncertainty of the 17 image-wise models. Each input image can have a dynamic number of models less than 17 based on the pre-defined β which excludes the models with high uncertainty values. We used a wide range of β values from 10^{-8} to 2 to demonstrate the capability of *3E-Net* versions to provide high performance.

4.4.3 Quantitative Evaluation

We adopted accuracy, precision, recall, and F1-score metrics to evaluate the performance of our model. Precision is the classifier's capability to not mark a result as positive if it is negative, the classifier's recall is its ability to locate all positive samples, and F1-score can be expressed as the harmonic mean of the precision and recall. The accuracy, precision, recall, and F1-score were determined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.5}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.6}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.7}$$

$$F1 - score = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \tag{4.8}$$

where TP and TN represent the correct predictions by our elastic ensemble models for the occurrence of a certain class or not, respectively, while FP and FN are the incorrect model predictions for all cases.

Performance of Standard Ensemble-based Models

Tables 4.1 and 4.2 illustrate precision, recall, F1-score and grading accuracy of standard ensemble of DCNNs (i.e. ensemble of the total 17 models) for Version A and Version B, respectively. Tables 4.1 and 4.2 show that both ensemble models can effectively differentiate grade 2 from the two other grades (grade 1 and grade 3). Moreover, Version A and Version B have achieved an average precision of 93.04% and 90.98%, respectively, while they achieved average grading accuracy of 93% and 90.68%, respectively.

TABLE 4.1: Grading performance (mean) of standard ensemble model (**Version A**) on Invasive Breast Carcinoma dataset using 5-fold cross-validation.

Grade	Precision	Recall	F1-score	Accuracy
Grade 1	89.86%	90.65%	90.25%	93.00%
Grade 2	99.05%	99.05%	99.02%	99.33%
Grade 3	90.05%	89.00%	89.51%	93.67%
Total	**93.04%**	**93.00%**	**93.01%**	**93.00%**

TABLE 4.2: Grading performance (mean) of standard ensemble model (**Version B**) on Invasive Breast Carcinoma dataset using 5-fold cross-validation.

Grade	Precision	Recall	F1-score	Accuracy
Grade 1	85.83%	88.83%	87.21%	90.68%
Grade 2	98.09%	95.14%	96.48%	97.68%
Grade 3	89.04%	87.89%	88.39%	93.00%
Total	**90.98%**	**90.68%**	**90.72%**	**90.68%**

Performance of *3E-Net* Models

To evaluate the performance of the uncertainty-aware component, we further investigate the grading accuracy of the elastic ensemble approach. Moreover, for a fair

comparison with the standard ensemble-based models, we introduced two new metrics: (1) Weighted Average Accuracy (WAA), which measures the average of grading accuracies for the 5 folds in the dataset weighted by the number of the included images in each fold; and (2) Abstain percentage (AP): measures the percentage of the excluded images to the total number of images in the dataset. The formulation of the two metrics are determined as follows:

$$WAA = \frac{1}{\sum_{i=1}^{t} d_i} \sum_{i=1}^{t} Accuracy_i * d_i \tag{4.9}$$

$$AP = \left(\frac{\sum_{i=1}^{t} R_i}{DS} \right) \times 100 \tag{4.10}$$

where d_i and $Accuracy_i$ represent the number of included images and grading accuracy in fold i over a total number of t folds, respectively, R_i is the count of the excluded images in fold i, and DS is the total number of images in the dataset

Table 4.3 demonstrates the capability of our elastic ensemble approach in providing higher grading accuracies for both *3E-Net* model variations (Version A & B) when compared to the standard ensemble models. Moreover, such improvement in the grading accuracies indicates that the excluded images are difficult to classify by the DCNN models, where a manual investigation is required for such images. It can be noticed that *3E-Net* models achieve the highest accuracies of 96.15% ($\beta = 5 \times 10^{-7}$) and 99.50% ($\beta = 5 \times 10^{-6}$) for Version A and Version B, respectively. As illustrated by Table 4.3, the other threshold β values yield grading accuracy of $\sim$ 95% for Version A and $\sim$ 99.40% for Version B.

TABLE 4.3: WAA and AP of *3E-Net* Model variations (Version A & Version B) on different β values.

Model	β	**Accuracy**	**AP**
3E-Net (Version A)	5×10^{-7}	**96.15%**	4.67%
	9×10^{-7}	95.82%	4.33%
	5×10^{-6}	94.86%	2.67%
	10^{-5}	94.56 %	2.00%
3E-Net (Version B)	5×10^{-6}	**99.50%**	33.00%
	10^{-6}	99.43%	42.00%
	9×10^{-7}	99.42%	43.00%
	5×10^{-7}	99.38 %	46.33%

Figure 4.3 depicts AP of the excluded images from the dataset over different values of β for *3E-Net* models (Version A & Version B). The curves show that AP decreases when we increase β. Also, starting from $\beta = 0.75$, the number of excluded images reaches zero for both models. Figure 4.4 depicts the ROC curves for both model versions using the standard and elastic ensemble-based approaches, see also Figure 4.7 for the confusion matrices obtained by our models.

Figure 4.5 and Figure 4.6 demonstrate the output visualisations of multiple filters applied to the first and last convolutional layers of the patch-wise network of the

standard ensemble model (version B). Note how the feature maps are distinctive in terms of their morphological structures.

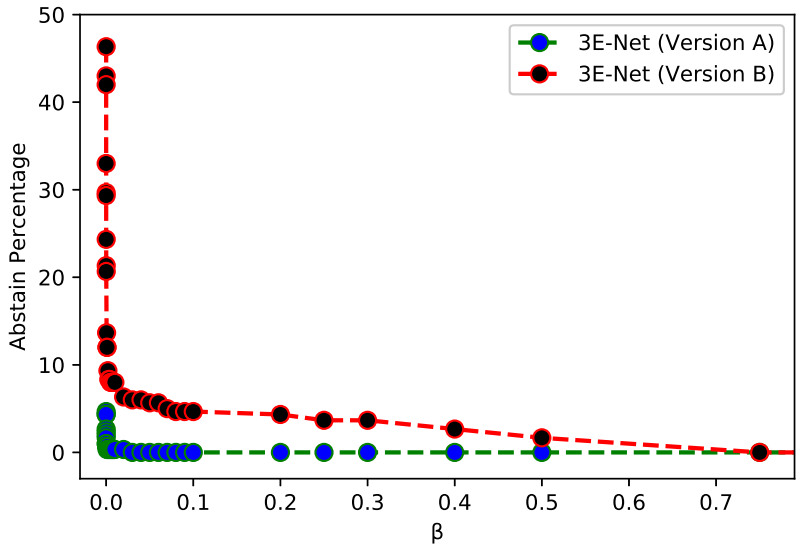


FIGURE 4.3: AP of excluded images for *3E-Net* Version A (Blue) and *3E-Net* Version B (red) over a range of threshold β values using elastic ensemble on Invasive Breast Carcinoma Dataset.

Comparison with different Methods

To demonstrate the effectiveness of our solution, we applied ablation study by comparing the performance of a state-of-the-art single DCNN model, standard ensemble-based models, and our elastic ensemble approach. In Table 4.4, we compare our *3E-Net* models with the state-of-the-art models in digital breast pathology, namely DCNN+SVM model [9], deep spatial fusion CNN model [42], two-stage CNN model [73], and ensemble of multi-scale networks (EMS-Net) [120]. As demonstrated by Table 4.4, our *3E-Net* model outperformed both the recent models in the literature and the standard ensemble models.

Performance of *3E-Net* on BreakHis Dataset

To confirm the effectiveness of *3E-Net* model, we applied *3E-Net* model (version A) on the Breast Cancer Histopathological Database (BreakHis) [102]. BreakHis has a total number of 7909 breast cancer histopathological images taken from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). The dataset is divided into 2480 benign and 5429 malignant microscopic images with a resolution of 700 x 460 pixels. We use 40X magnification images which has 625 benign and 1370 malignant samples.

Here, we down-sampled the images to around 80% of the original scale (448 x 336). This image scale produces 6 image-wise CNNs to be used in the ensemble process. We also used the same hyperparameter settings except for patch-stride values, where we used s = 28 for training the backbone network (DenseNet-161) and s = 112 for training the 6 image-wise CNNs. Finally, as the BreakHis dataset contains only two classes (benign or malignant), we fine-tuned DenseNet-161 by updating

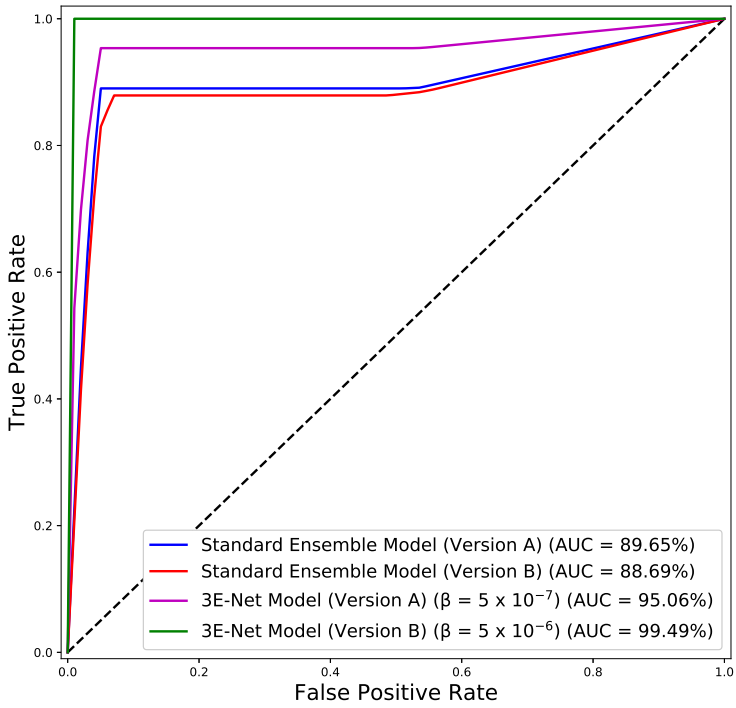


FIGURE 4.4: ROC curves for the standard and elastic versions of our models (A & B).

TABLE 4.4: Comparison between different methods on Invasive Breast Carcinoma Dataset using 5 fold cross-validation.

Method	Precision	Recall	F1-score	Accuracy
DCNN + SVM [9]	87.64%	87.38%	87.38%	87.38%
Deep Spatial Fusion CNN [42]	92.67%	92.65%	92.62%	92.65%
Two-stage CNN [73]	93.07%	92.69%	92.70%	92.69%
EMS-Net [120]	93.04%	93.00%	93.00%	93.00%
Standard Ensemble Model (Version A)	93.04%	93.00%	93.01%	93.00%
Standard Ensemble Model (Version B)	90.98%	90.68%	90.72%	90.68%
3E-Net (Version A) ($\beta = 5 \times 10^{-7}$)	**96.23%**	**96.15%**	**96.16%**	**96.15%**
3E-Net (Version B) ($\beta = 5 \times 10^{-6}$)	**99.54%**	**99.50%**	**99.50%**	**99.50%**

the number of neurons from 1000 to only 2 neurons in the last fully connected layer. As shown in Table 4.5, our model has proved to be effective on both standard and elastic ensemble. We applied 5-fold cross-validation and achieved a classification accuracy of 99.80% using standard ensemble technique. Also, the results show the validity of our novel elastic method of *3E-Net* on different *beta* values by improving

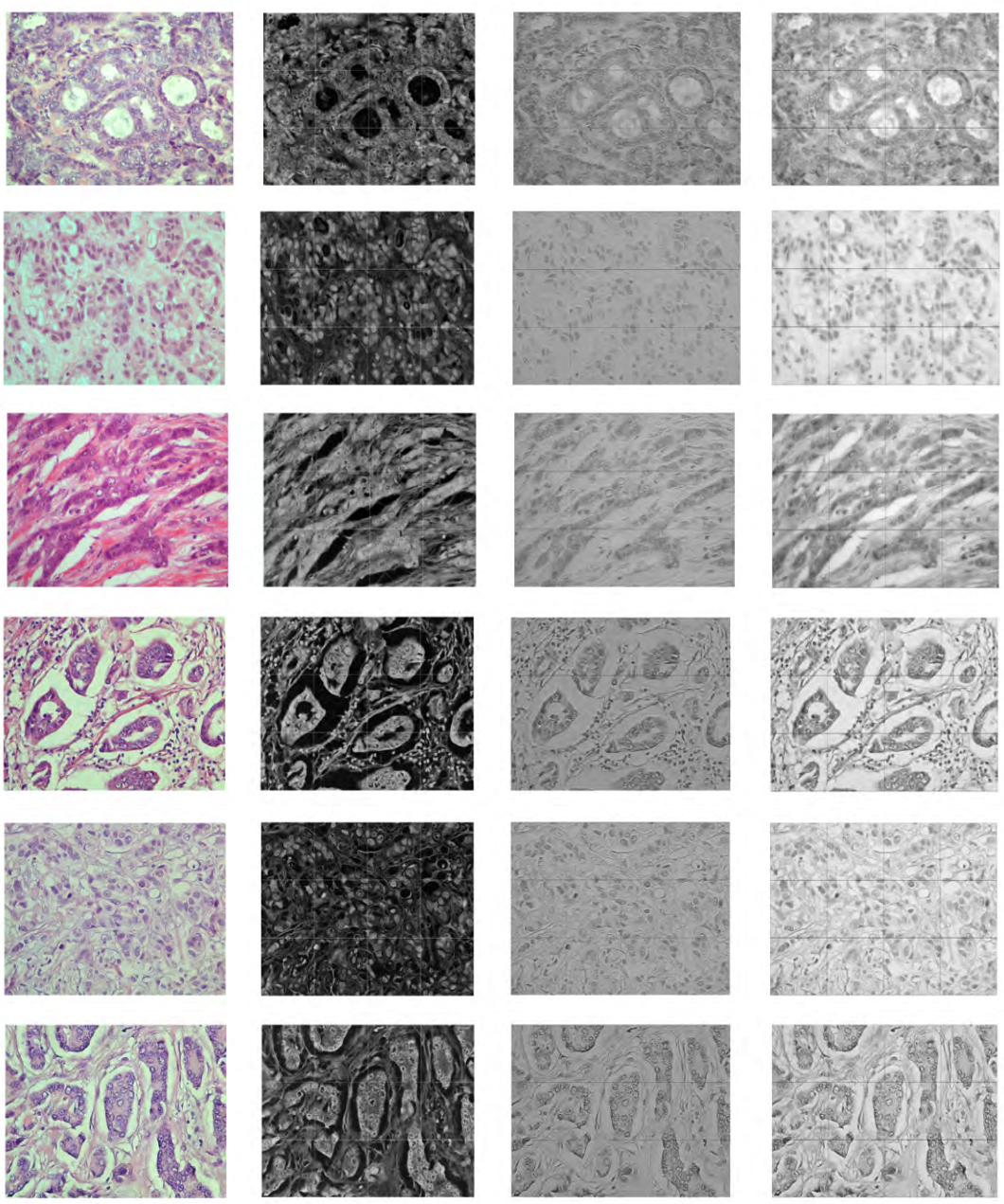


FIGURE 4.5: Examples of feature maps obtained by multiple filters learned within the first convolutional layer of the patch-wise network of standard ensemble (version B). The colored image is the original, while the gray-scale images are the output maps.

the performance, where an accuracy of 99.95% has been achieved on ($\beta = 9 \times 10^{-6}$).

4.4.4 Qualitative Evaluation

To quantitatively evaluate the performance of our model on the excluded images, we set β to a high value to find images that are less sensitive and highly uncertain to the 17 image-wise models in the ensemble of DCNN models. Figure 4.8 shows the images, for which all the image-wise models in the ensemble agree on the uncertainty decision based on the high uncertainty values resulted from these models. Figure 4.8(c) shows two images from the selected excluded images which are agreed

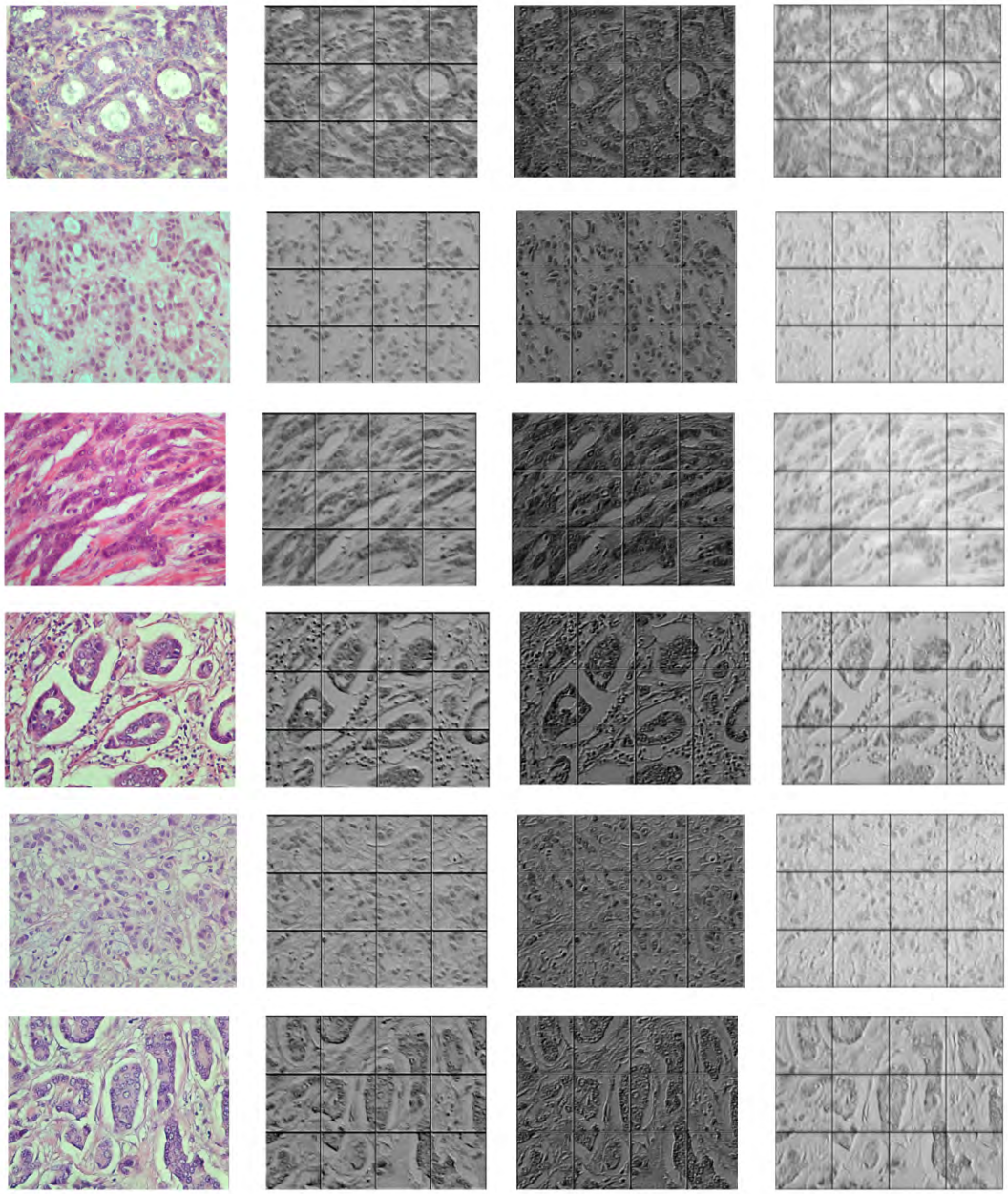
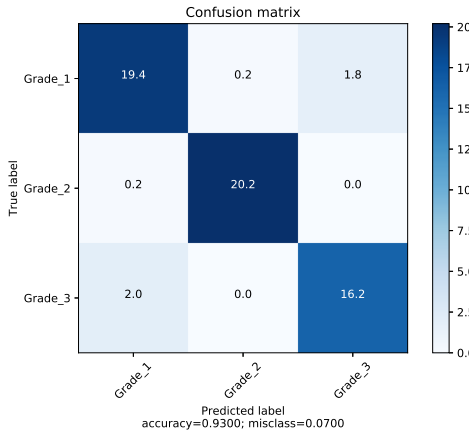


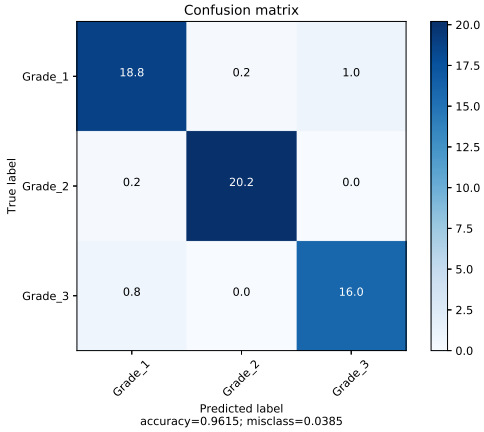
FIGURE 4.6: Examples of feature maps obtained by multiple filters learned within the last convolutional layer of the patch-wise network of standard ensemble (version B). The colored image is the original, while the gray-scale images are the output maps.

on their uncertainty by both *3E-Net* model variations (Version A and Version B). Moreover, it can be noticed that the highly uncertain images come from grade 1 or grade 3, which proves trustworthy of our results in Tables 4.1 and 4.2 to show how it is slightly hard to differentiate between grade 1 and grade 3.

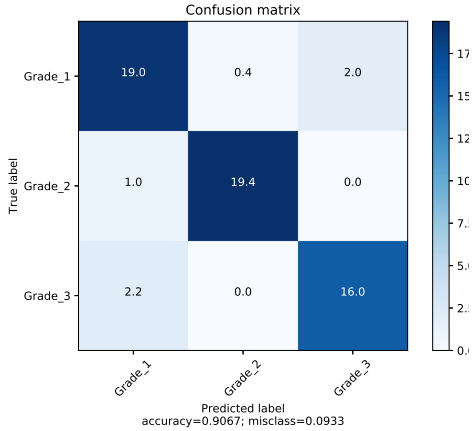
Based on the sample of the excluded images shown in Figure 4.8, we returned to a domain expert to further investigate the possible reason behind the high uncertainty of the excluded images. The uncertainty may be due to usage of datasets from heterogeneous populations [47], or reduced sample size used in the study [88]. In this regard, additional information depending on the staining of specific biomarkers for breast cancer grading such as Ki67 [64] could be used to resolve the diagnostic



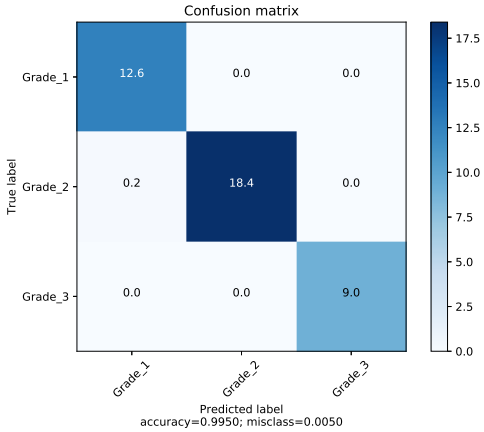
(A) Standard Ensemble Model (Version A)



(B) *3E-Net* (Version A) ($\beta = 5 \times 10^{-7}$)



(C) Standard Ensemble Model (Version B)



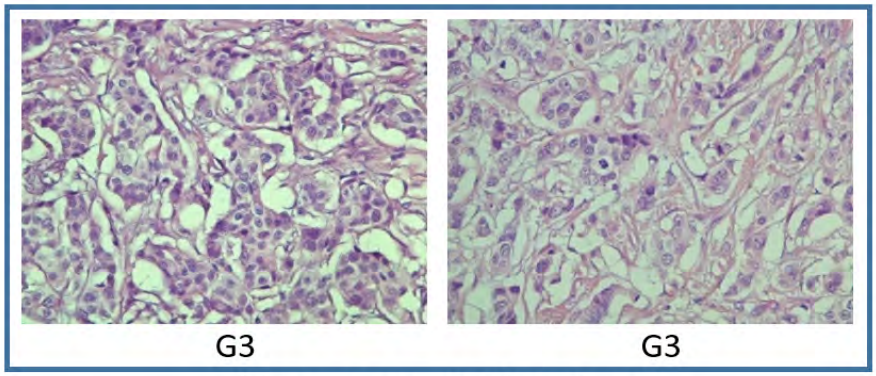
(D) *3E-Net* (Version B) ($\beta = 5 \times 10^{-6}$)

FIGURE 4.7: Confusion matrices for our developed models.

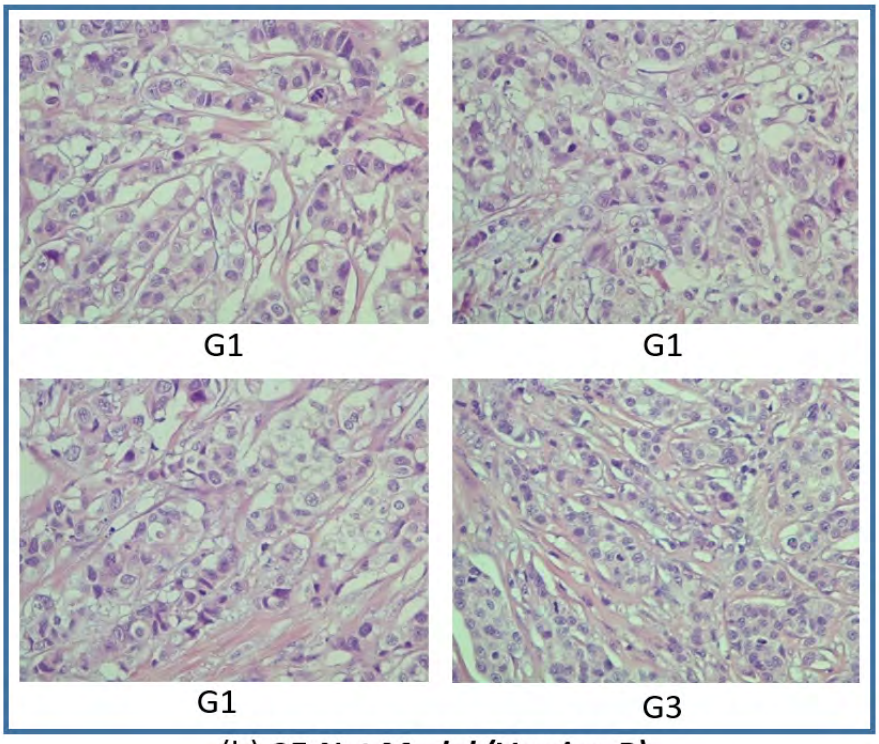
TABLE 4.5: Performance (mean) of standard and elastic ensemble models (**Version A**) on BreakHis dataset using 5-fold cross-validation.

Model	β	**Accuracy**	**AP**
Standard Ensemble Model	NA	**99.80%**	NA
3E-Net Model	9×10^{-6}	**99.95%**	1.10%
	5×10^{-4}	99.90%	0.50%
	3×10^{-2}	99.85%	0.10%

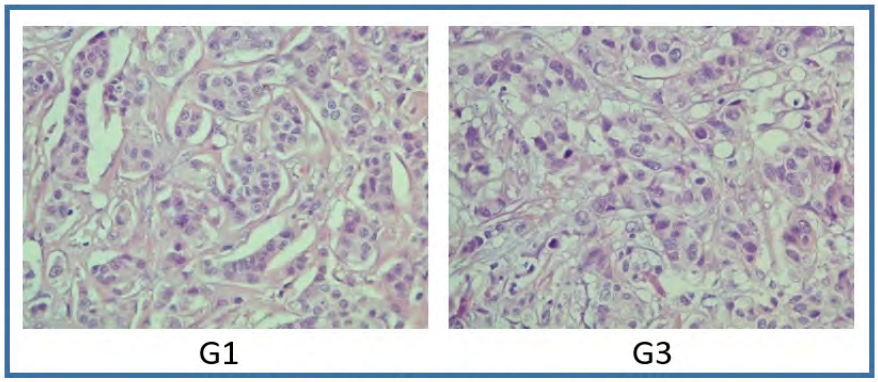
uncertainty in CNN.



(a) ***3E-Net Model (Version A)***



(b) ***3E-Net Model (Version B)***



(c) ***3E-Net (Versions A & B)***

FIGURE 4.8: Highly uncertain excluded images from the grading process of our dynamic ensemble-based models. The excluded images come from three perspectives: (a) *3E-Net* Model (Version A), (b) *3E-Net* Model (Version B), and (c) Versions A & B combined. Each image in the figure has a caption that presents the ground truth label (G1: grade 1 and G3: grade 3).

4.5 Summary

In this chapter, we introduced *3E-Net* model to classify invasive breast carcinoma using histopathological images into three grades: grade 1, grade 2, and grade 3. Our model has the capability to learn different levels of contextual information using image patches through various image-wise CNN models. Moreover, our ensemble model has been designed in a way to measure the level of randomness (using a novel entropy-based formula) in the input images and quantify the challenges in grading images. We evaluated *3E-Net* on Invasive Breast Carcinoma Dataset from 'Agios Pavlos' General Hospital of Thessaloniki, Greece. Our elastic ensemble model has two variations that achieved a grading accuracy of 96.15% and 99.50% in the five-fold cross-validation on training images and outperformed standard ensemble-based models and a state-of-the-art method. *3E-Net* demonstrated its capability in excluding the uncertain microscopy images to be investigated and explored by medical professionals.

3E-Net proved its effectiveness in terms of performance and the selection of uncertain images to be excluded and investigated by medical professionals. The effectiveness of *3E-Net* came from its ability to be used for grading task (i.e. classify different grades) which is a very challenging task. The grading task usually includes fuzziness and overlap between class boundaries making it a difficult task. Moreover, *3E-Net* features a simple strategy of building an automated model for grading/diagnosis. This simplicity of design make *3E-Net* a light-weight model that can be applied to resource constrained environment where it can perform better in terms of space and time requirements (e.g. a tablet PC). This helps hospitals to have models developed on portable devices which ease the process of grading and introducing immediate treatment plans for patients based on fast inference results from *3E-Net*.

In the next chapter, we present, our second contribution, *MCUa* model for classification of breast cancer histopathology images. *MCUa* has the characteristic of capturing different size/scale variations of nuclei objects in histopathology images by introducing multi-scale input and multi-architecture usage for feature extraction. This characteristic aids in providing diversity in scales and features, which consequently enrich the automated model by detailed information to enhance the performance. In addition, *MCUa* introduces a flexible uncertainty-aware component which generates multi-scalar predictive probability distributions for measuring uncertainty of image prediction. This aids in introducing a reliable method which accurately select the uncertain samples.

Chapter 5

MCUa: Multi-level Context and Uncertainty aware Model for Classification of Breast Cancer Images

In the last chapter, we explained in detail *3E-Net* for classifying grades of invasive breast carcinoma histopathology images as the first contribution of the thesis. *3E-Net* showed simplicity in the design making it suitable to be deployed for resource constrained environments (portable edge devices) as it generates inference in short time. Moreover, *3E-Net* demonstrated high performance for a very challenging task such as grading task which is usually characterised by high overlapping between class boundaries. *3E-Net* combines simplicity in design and efficiency for difficult tasks making it an important advancement for medical image analysis. In this chapter, we introduce *MCUa* model for classification of breast histopathology microscopic images where it utilises a deep learning strategy to build an automated diagnosis system. Unlike *3E-Net*, *MCUa* has the ability to (1) capture different size/scale variations of nuclei objects in histopathological images by introducing multi-scale input and multi-architecture usage for feature extraction, (2) learns multi-scale and multi-level context-aware information, and (3) provide an uncertainty-aware component based on Monte-Carlo (MC) dropout [29], which generates predictive probability distributions instead of a single scalar. All these characteristics help to provide a highly developed system that benefits clinical practice. Findings reported in this chapter have been published in [91].

5.1 Overview

Breast histology image classification is a crucial step in the early diagnosis of breast cancer. In breast pathological diagnosis, Convolutional Neural Networks (CNNs) have demonstrated great success using digitised histology slides. However, tissue classification is still challenging due to the high visual variability of the large-sized digitised samples and the lack of contextual information. In this chapter, we introduce a novel CNN, called Multi-level Context and Uncertainty aware (*MCUa*) dynamic deep learning ensemble model. *MCUa* model consists of several multi-level context-aware models to learn the spatial dependency between image patches in a layer-wise fashion. It exploits the high sensitivity to the multi-level contextual information using an uncertainty quantification component to accomplish a novel dynamic ensemble model. *MCUa* model has achieved a high accuracy of 98.11% on

a breast cancer histology image dataset. Experimental results show the superior effectiveness of the presented solution compared to the state-of-the-art histology classification models.

The chapter is organised as follows. Section 5.2 presents detailed introductory of the background and the research work. In section 5.3, we discuss in details the architecture of our model. Section 5.4 describes our experimental results obtained. Finally, Section 5.5 discusses our findings by presenting the summary of our work and introducing few potential future research directions.

5.2 Introduction

Breast cancer is the driving sort of cancer in women, coming about in 1.68 million modern cases and 522,000 passings in 2012 around the world. It has been accounted for 25.16% of all cancer cases and 14.71% of cancer-related passing [103]. Precise determination of breast cancer is pivotal for suitable treatment and prevention of further progression. A few symptomatic tests have been utilised, counting physical examination, mammography, magnetic resonance imaging (MRI), ultrasound, and biopsy. Histology image examination resulted from biopsy considered as a crucial step for breast cancer diagnosis. In the diagnosis process, pathologists evaluate the cellular areas of hematoxylin-eosin (H&E) stained histology images to decide the predominant type of breast tissues, including normal tissue, benign lesion, in situ carcinoma, and invasive carcinoma. Histology images are large in size with a complex morphological structure. Therefore, identifying carcinoma regions based on the manual investigation conducted by medical professionals is a challenging and time-consuming process.

Traditionally, histology imaging in clinical practice is focused primarily on pathologists' manual qualitative analysis. However, there are three main issues with such practice. One, there is shortage of pathologists around the world, especially in developing countries and small hospitals. This scarcity of resources and unequal allocation is a pressing issue which need to be addressed. Second, the pathologist's extensive scientific expertise and long-term diagnostic experience determine whether the histopathological diagnosis is accurate or not. This subjectivity may cause in a slew of diagnostic errors. Finally, pathologists are vulnerable to fatigue and inattention while reading the complex histology images. In order to address these issues, it is crucial to establish automated and precise histological image classification tasks. Thanks to the advancement of computer aid diagnosis (CAD) frameworks that have made the difference in reducing the workload and improved the detection accuracy [44].

There are two challenging perspectives in the classification of H&E stained breast histology images. First, there are colossal intra-class fluctuations and inter-class likenesses in microscopy images, e.g., the difficult mimics from benign which has a comparative morphological appearance with carcinoma. Figure 5.1(a) shows benign and carcinoma microscopy images with a similar morphological structure, in terms of the nuclei distribution. Second, in histology image analysis, structural and contextual information is usually lost due to the sectioning process of high resolution images into small patches. To put it differently, the loss of structural and contextual information is due to the fact that histological image is divided into sections and dealing with only local representation of image patches makes it difficult to preserve the spatial dependencies of different image patches. Therefore, learning contextual information is crucial by integrating important information from different image parts

and hence improving the classification performance. Figure 5.1(b) depicts the shape of image patches used as an input to patch-based deep convolutional neural network (DCNN) models. Several different feature engineering [11, 110] and feature learning [6, 71, 73] models have been previously developed to classify digitised breast histology tissues. Feature Learning showed great success in addressing numerous issues within the field of digital pathology, including the above-mentioned challenging problems. As of lately, deep convolutional neural networks (DCNNs) have been broadly recognised as one of the foremost capable tools for histology tissue classification. In spite of their predominance, a single DCNN model has constrained capacity to extract discriminative features and results in lower classification accuracy [9, 73]. Hence, an ensemble of DCNN models has been developed to memorise the representation of histology images from distinctive view-points for more precise classification [120]. However, accommodating contextual information in the architecture of CNNs is a requirement to cope with the huge size of histology images [9, 27]. Consequently, ensemble CNNs should allow for the contextual representation to be learned. Moreover, despite the prevalence of DCNN models in providing high classification performance and alleviating the workload encountered by pathologists, a number of histological images might need assistance in diagnosis by professional medical expertise due to their complexity. Such images have to be excluded from automated image classification and to be presented for pathologists for manual investigation. Consequently, we introduce an uncertainty quantification method which measures the level of image prediction's randomness using DCNN models. This approach aids in the identification of various ambiguous regions which can be clinically useful. It also helps pathologists and medical practitioners to prioritise images for annotations.

In this chapter, we present a novel dynamic ensemble CNN with terming Multi-level Context and Uncertainty aware (*MCUa*) model[1] for the automated classification of H&E stained breast histology images. First, we resize input images into two different scales to capture multi-scale local information. Then we designed patch feature extractor networks by extracting patches and feed them to pre-trained fine-tuned DCNNs (*i.e.* DenseNet-161 and ResNet-152). Unlike the work conducted in [120], the extracted feature maps are then used by our context-aware networks to extract multi-level contextual information from different pattern levels. Finally, a novel uncertainty-aware model ensembling stage is developed to dynamically select the most certain context-aware models for the final prediction. We evaluated the performance of our model on BreAst Cancer Histology Images (BACH) challenge dataset [7], which consists of 400 high-resolution H&E stained breast histology images and divided into four categories, namely normal, benign, in situ carcinoma, and invasive carcinoma. *MCUa* model alleviates the bias that might be caused during the traditional workload of histological image analysis by introducing an automated image classification model which captures the spatial dependencies among patches of high-resolution images. Additionally, it presents a measure of uncertainty which helps in providing a more robust predictions using a dynamic ensemble mechanism that improves the diversity of the model by coping with different network architectures and multi-level contextual information. This can be achieved by (1) introducing effective pre-trained and fine-tuned DCNN models which learn to explore hierarchical discriminative features and differentiate between different class categories and (2) learn spatial dependencies among image patches to preserve contextual information between feature maps.

[1]The code is available at https://github.com/zakariaSenousy/MCUa-Model.

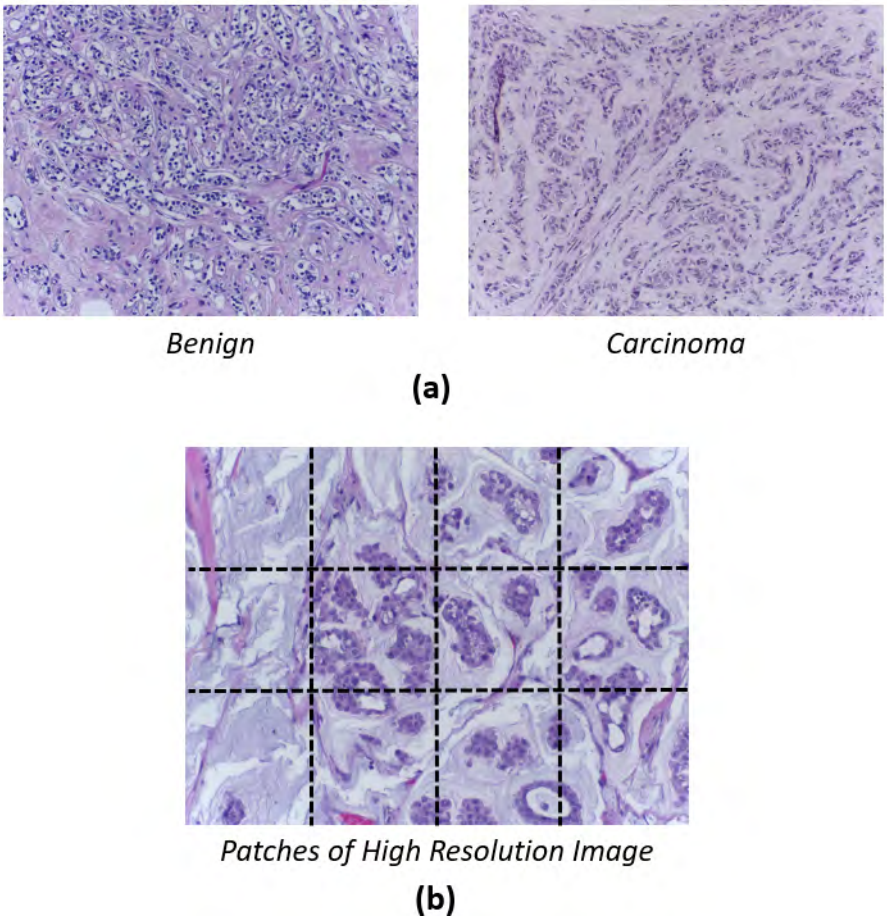


FIGURE 5.1: (a) An example of similar morphological structures between benign and carcinoma sections. (b) Patches of a high section, which are used by DCNN models to learn the spatial dependencies information.

The contributions of this chapter are summarised below:

- introduced a multi-scale input and multi-architecture stage for feature extraction which exploits the granularity in encoding multi-level features and increase the diversity of the extracted features. Multi-scale and multi-architecture mechanism helps in capturing different sizes and scales for nuclei and tissue structures;
- developed a novel context-aware model to learn multi-scale and multi-level contextual information by encoding the spatial dependencies among patches in histology images;
- introduced a novel dynamic ensemble strategy by selecting the most certain models for each particular image based on an uncertainty-aware mechanism. The developed mechanism has been designed by measuring the level of randomness of all models in the ensemble architecture, and consequently a dynamic number of accurate models is chosen and combined to obtain the final prediction; and

- conducted a thorough experimental study on the BACH image dataset, and obtained better performance than state-of-the-art computational pathology models.

5.3 *MCUa* Model

In this section, we describe our Multi-level Context and uncertainty aware (*MCUa*) dynamic deep learning ensemble model in details. As illustrated in Figure 5.2, the *MCUa* model consists of an arbitrary number of multi-level context-aware models, where each model consists of two components: a) a patch-wise feature extractor component, to extract the most prominent features from image patches; and b) a context-aware component, aims at capturing the spatial dependencies among the extracted patches. *MCUa* starts by taking the original image and then resizing the image to m scales to get various and integral visual features from the multi-scale image feature. A number of patches are extracted from each image scale to be inserted into a pre-trained feature extractor. Several salient feature maps are extracted from the pre-trained feature extractor, which are then inserted to multi-level context-aware models. Each context-aware model has a certain contextual information level that can be learned from a group of feature maps. As a final stage, MC-dropout is applied to each context-aware model to produce a measure of uncertainty. This is done by applying a number of test passes for each input image through the context-aware network. Each test pass produces a class probability for the image, using this information, we calculate the mean and standard deviation to provide image class label and uncertainty measure, respectively. A dynamic process of model selection, based on an uncertainty measure value and a pre-defined threshold, is utilised to pick up the most certain models and then produce the final class label.

5.3.1 Multi-scale Feature Extraction

Multi-scale image feature extraction is pivotal for having diverse and complementary visual features in H&E stained breast histopathological microscopy. To extract multi-scale features, we first resize the original image to different scales. Then, image patches are extracted from each scale using a sliding window of size $p_w \times p_h$ and a stride s. Therefore, the total number of patches extracted from the resized image can be represented by

$$a = \left(1 + \left\lfloor \frac{I_W - p_w}{s} \right\rfloor\right) \times \left(1 + \left\lfloor \frac{I_H - p_h}{s} \right\rfloor\right), \tag{5.1}$$

where I_W and I_H are width and height dimensions of the resized image, respectively.

The images at the different scales are then divided into partially overlapped patches using different stride values for training and testing data extraction. This increases the level of locality information and the number of training patches. Moreover, to increase the diversity of training data and, at the same time, alleviate the overfitting of DCNN models, several data augmentation methods have been applied. For instance, each patch has been transformed using a rotation operation and with/without vertical reflections. Also, random color perturbations recommended by [66] has been applied to each patch to alleviate the high visual variability of the patches. The data augmentation process makes our model learn rotation invariant, reflection invariant, color invariant features and make pre-processing color normalisation [67].

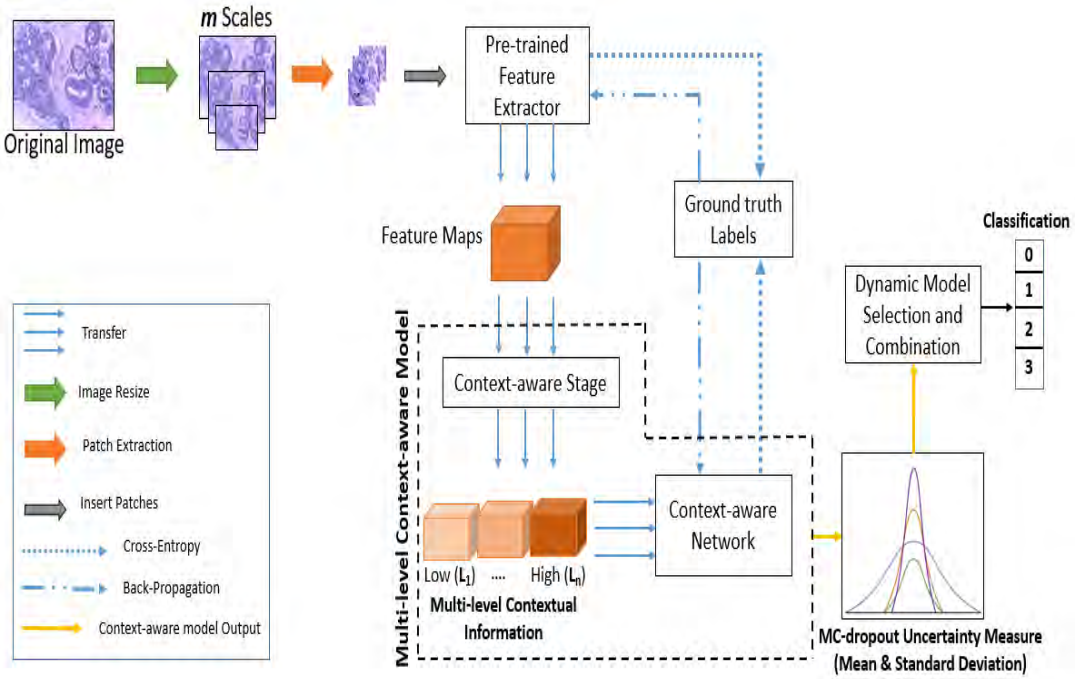


FIGURE 5.2: Overview of *MCUa* model. Our model takes the original image and resizes it into multiple scales. For each scale, several patches are extracted which are then inserted into patch-wise networks (i.e. pre-trained DCNNs) to extract salient features. The extracted features are then inserted into multi-level context-aware models to learn spatial dependencies between image feature maps. Context-aware models work on extracting contextual information between feature maps based on different levels (*L*1 to *Ln*). L_1 is considered as low level context which builds contextual information among two original feature maps, while *Ln* is considered as high level context which builds contextual information among all the original feature maps extracted from the image. Finally, a dynamic model selection is applied to select the most certain models based on uncertainty quantification and a combination of selected models is applied to produce the final prediction. For each image, a number of test passes is applied using MC-dropout to produce a list of probability distributions which are then used to generate mean and standard deviation. The mean is used for identifying the class label of a single model, while standard deviation is utilised for measuring the level of randomness and uncertainty. In a dynamic way, each image in the dataset has a number of accurate models which are chosen based on low value of uncertainty determined using a pre-defined threshold. These selected models' mean predictions are aggregated for final class prediction.

5.3.2 Fine-tuning the Backbone Networks

The pre-trained DCNN models (namely, ResNet-152 [37] and DenseNet-161 [41]) are fine-tuned to be used as the backbone feature extractors of *MCUa* model. We adapted the pre-trained DCNN models to a four-category image classification problem, by modifying the number of neurons in the last fully-connected layer from 1000 neurons (where ResNet-152 and DenseNet-161 are pre-trained models on ImageNet [86]) to only 4 neurons. Consequently, the fine-tuned DCNN models can take input of microscopy image patches (i.e. augmented versions of the patches extracted from

resized versions of microscopy images) and produce an output of softmax probabilities belonging to the 4 cases (Normal, Benign, In situ carcinoma, and Invasive carcinoma) of the BACH dataset.

During the fine-tuning process, we used Adam optimiser [48] to minimise the categorical cross-entropy loss function which is presented in equation 4.2. Softmax activation function (as shown in equation 4.3) is applied to the DCNN model's predictions of the last fully connected layer.

Once the training of the pre-trained DCNN models is accomplished, the last convolutional layer is used to construct our feature space or to extract a number of feature maps (equivalent to the number of patches in each image).

5.3.3 Multi-level Context-aware Models

To capture the spatial dependencies among image patches, *MCUa* has been designed in a context-aware fashion to learn different possible multi-level contextual information. Here, we used the output of feature extractor of each pre-trained DCNN model and fed it into several multi-level context-aware models. The level of contextual information learned by *MCUa* is determined by a pattern of neighborhood criteria. More precisely, we encode the spatial relationship information among patches based on the neighborhood of patches that form some random shape. In other words, our context-aware models have been designed based on a pattern tuple $P_{g,S_i} = (g, S_i)$, where g is the number of patches used in the context-aware process and S_i is the set of shape indices (where each index i is associated to a unique set of shape indices). To identify a shape index, the starting patch and $g-1$ directions should be specified. Figure 5.3 clarifies an example of how different pattern levels work to extract contextual information. For instance, P_{2,S_1} has a value of $g = 2$ and two shapes. Moreover, P_{4,S_2} has a value of $g = 4$ and a set of shapes where the shapes are built using a number of feature maps (e.g. 3, 6, 5, and 4). More precisely, the process of building contextual information for the shape index represented in P_{4,S_2} works by identifying the starting feature map location (i.e. feature map number 3), then all the possible directions in the matrix of the feature maps has to be defined, where direction 1 is for the *down* direction to pick feature map number 6, then directions 2 and 3 are for the $left$ directions to pick feature maps 5 and 4, respectively, (Please see Figure 5.3). Each feature map utilises the pattern tuple mechanism to bring the spatial dependencies information with other neighboring feature maps.

The feature patterns have been designed by taking into account the image-level labels for the final classification during the minimisation of loss function. Context-aware networks are mainly image-wise networks which take concatenated feature maps generated from the original neighbor feature maps (extracted from the input image). These concatenated feature maps are fed into context-aware networks to classify the images based on local and contextual features extracted from images. Context-aware networks are trained against image-level labels. More precisely, we minimise the loss function of different patterns of feature maps inserted as an input to the final class label associated to the image as an output.

Each Context-aware CNN consists of a sequence of 3×3 convolutional layers followed by a 2×2 convolution with stride of 2 for down-sampling. Batch normalisation and ReLU activation function were used at the end of each layer. To obtain the spatial average of feature maps, a 1×1 convolutional layer is used before the classifier. The network ends with 3 fully connected layers and a log softmax classifier. The full description of the architecture is presented in section 4.3.2.

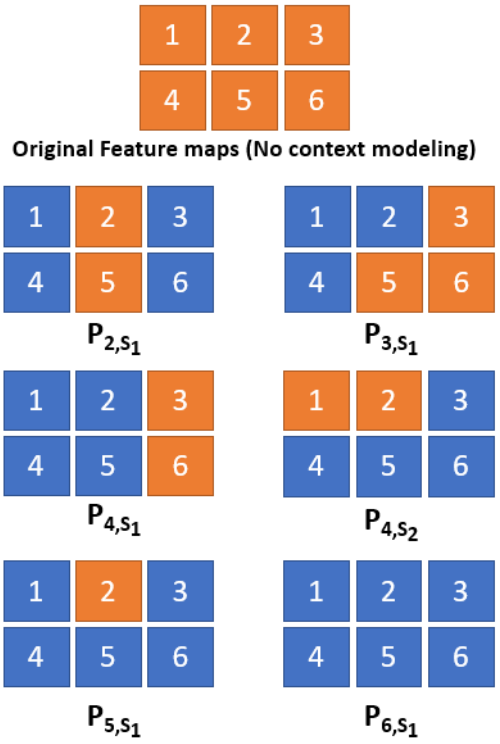


FIGURE 5.3: The extraction process of the contextual information (i.e. context modeling) with different pattern levels using six feature maps. The original feature maps (highlighted in orange) are used to encode different levels of contextual information. For instance, P_{2,S_1} represents the contextual information of a pattern that is composed of 2 neighbor feature maps, while P_{4,S_1} and P_{4,S_2} represent the process of building contextual information for four neighbor feature maps with different set of shapes, respectively. The blue highlighted feature maps represent the maps chosen to build contextual information.

During the training of *MCUa*, a partly overlapped patches are extracted from the image by using different stride values. The stride value for each scale is chosen to increase the number of patches and hence improve the contextual representation of *MCUa*. We found in our experiments that using high stride decreases the accuracy for a single context-aware model on a validation set.

The context-aware CNN has been trained using categorical cross-entropy loss and learns to classify images based on the local features of image patches and spatial dependencies among the different patches. Like pre-trained DCNN, data augmentation has been applied.

In algorithm 2, we describe the implementation flow of a single context-aware model. We start by resizing the image to multiple scales to extract smaller patches. We then pass the extracted patches to a pre-trained DCNN model to extract feature maps. After that, we iterate over each feature map and get the indices of all possible feature maps that can build possible pattern of neighborhood relationships. The related feature maps are then concatenated and inserted in a set which holds all the concatenated feature maps. Finally, we pass the concatenated feature maps set to the context-aware CNN. This is to learn spatial dependencies among the related feature maps and produce the network output. As a consequence, the feature maps will be fed into the log softmax function to produce the probability distribution of the

image.

Algorithm 2: Single Context-aware Model

Input: Original image X to be classified

Output: class label $\hat{y}$

$X^{'} \leftarrow X$ // resize original image to m scales

// extractPatches is a function which takes image and patch dimensions as an input and outputs n patches $\{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$

$\{x^{(1)}, x^{(2)}, ..., x^{(n)}\} \leftarrow$ extractPatches($X^{'}$, p_w, p_h)

// featureExtractor network takes n patches as an input and outputs n feature maps $\{fm^{(1)}, fm^{(2)}, ..., fm^{(n)}\}$

$\{fm^{(1)}, fm^{(2)}, ..., fm^{(n)}\} \leftarrow$ featureExtractor($\{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$)

$F \leftarrow \{fm^{(1)}, fm^{(2)}, ..., fm^{(n)}\}$ // store all the extracted feature maps to set F

$T, Y \leftarrow \{\}$ // define and initialise two empty sets

for $fm \in F$ **do**

  // getPatternIndices is a function which takes a feature map fm and returns the indices of all possible neighbor feature maps which form a pattern

  $\{i_1, i_2, ..., i_n\} \leftarrow$ getPatternIndices(fm)

  // concatenate fm with possible neighbor feature maps which form a pattern

  $Y \leftarrow fm \parallel k_{i_1} \parallel k_{i_2} \parallel ... \parallel k_{i_n}$; where $k_{i_1}, k_{i_2}, ..., k_{i_n} \in F \wedge k_{i_1}, k_{i_2}, ..., k_{i_n} \neq fm$

  // append the newly concatenated feature maps Y to T

  $T \leftarrow T \parallel Y$

end

// contextAwareCNN is the network responsible for learning the spatial dependencies of all feature maps and their formed patterns

$O \leftarrow$ contextAwareCNN(T)

// logSoftmax function is used to map the output of contextAwareCNN O to probability distribution V

$V \leftarrow$ logSoftmax(O)

$\hat{y} \leftarrow \arg\max V$

5.3.4 Dynamic Model Selection and Combination

The final stage of *MCUa* model is to dynamically ensemble the most certain models for each image. To this end, we adapted an ensemble-based uncertainty quantification component to allow for a dynamic selection of context-aware models to produce the final prediction for an input image. To measure the uncertainty of the individual context-aware models in *MCUa*, we adopted MC Dropout [29] for each model. MC

dropout is a technique used in deep learning to measure the uncertainty of predictions made by neural network models. The basic idea behind MC dropout is to use dropout, a regularisation technique that randomly drops out connections/neurons in the network during training, as a means of approximating a Bayesian approximation of the model's posterior distribution. In each testing iteration, MC dropout activates a random subset of neurons, causing changes in the output probability distribution or predictions for the input image. By introducing this random neuron activation during testing, MC dropout enables the network to estimate the level of uncertainty in its predictions. We applied MC dropout for each context-aware model, in the test phase, to produce a list of probability predictions for each class of the input image. Then, we calculated the mean and standard deviation for each class. The mean is used to produce the final class label ($\hat{y}$) of the image, while standard deviation is considered as a measure of uncertainty for the context-aware model. Based on such uncertainty measures, a dynamic number of context-aware models will be selected (based on uncertainty threshold value (δ)) for each particular image.

More precisely, each input image will be sensitive to a certain number of context-aware models to form the final model ensemble. A context-aware model will be selected if its uncertainty measure value is less than a pre-defined δ, as described in our experimental study. More importantly, images with zero chosen models during this dynamic selection process can be provided to medical professionals (pathologists) for reviewing and annotating. Once the context-aware models are selected, the mean class predictions is aggregated to produce the final class prediction distribution. Here, we formulate the mean prediction and standard deviation as

$$\mu = \frac{1}{z}\sum_{i=1}^{z}\beta\left(\Phi_i(X); W\right), \tag{5.2}$$

$$\sigma = \frac{1}{z}\sum_{i=1}^{z}\left(\beta\left(\Phi_i(X); W\right) - \mu\right)^2, \tag{5.3}$$

where μ represents the mean prediction, σ defines the uncertainty and z defines the number of MC dropout test passes. The function β represents the context-aware CNN with input X and W denotes the network weights, while Φ_i defines a MC dropout test pass i to the input image X.

Algorithm 3 provides a detailed description of the ensemble process of *MCUa* model. Each model produces a single probability distribution. We applied some MC dropout test passes to generate a list of probability distributions for each model. Then, to get the final class prediction and the measure of uncertainty for each model, we computed the mean and standard deviation of the generated list of probability distributions, respectively. Finally, using a δ value, we include only the most certain models and we aggregate the mean of probability distributions for these models to produce $\hat{y}$.

Algorithm 3: *MCUa* Model

Input: Original image X to be classified

Output: Class label $\hat{y}$

$X_{scale1}, X_{scale2}, ..., X_{scalem} \leftarrow X$ // resize original image to m scales

$\{x_m^{(1)}, x_m^{(2)}, ..., x_m^{(n)}\} \leftarrow$ extractPatches(X_{scalem}, p_w, p_h)

$F_{FeatureExtractor_m} \leftarrow$ FeatureExtractor($\{x_m^{(1)}, x_m^{(2)}, ..., x_m^{(n)}\}$)

for $fm \in F_{FeatureExtractor_m}$ **do**

  $i_1, i_2, ..., i_n \leftarrow$ getPatternIndices(fm)

  $Y \leftarrow fm \parallel k_{i_1} \parallel k_{i_2} \parallel ... \parallel k_{i_n}$

  $T \leftarrow T \parallel Y$

end

$T_{all} \leftarrow \{T_{M1}, T_{M2}, ..., T_{Mn}\}$ // output T_{all} of context-aware stage from n context-aware models $\{M_1, M_2, ..., M_n\}$

for $j \in \mathrm{MC}_{\mathrm{dropout}}$TestPasses **do**

  $\{O_{M1}, O_{M2}, ..., O_{Mn}\} \leftarrow$ contextAwareCNN($\{T_{M1}, T_{M2}, ..., T_{Mn}\}$)

  // probability distribution V from n context-aware models

  $\{V_{M1}, V_{M2}, ..., V_{Mn}\} \leftarrow$ logSoftmax($\{O_{M1}, O_{M2}, ..., O_{Mn}\}$)

  V_{total}.append($\{V_{M1}, V_{M2}, ..., V_{Mn}\}$)

end

// get model-wise mean and uncertainty of probability distributions

$\{\mu_1, \mu_2, ..., \mu_n\} \leftarrow$ mean(V_{total})

$\{\sigma_1, \sigma_2, ..., \sigma_n\} \leftarrow$ standardDeviation(V_{total})

chosenModels $\leftarrow$ {}

for $j \in contextAwareModels$ **do**

  if $\sigma_j < \delta$ **then**

    chosenModels.append(μ_j)

  end

end

// aggregate the mean probability distributions of chosen models

$B \leftarrow$ aggregate(chosenModels)

$\hat{y} \leftarrow \arg\max B$

5.4 Experimental Study

In this section, we present our comprehensive experimental study along with the datasets used for evaluation. Moreover, we introduce an ablation study to show how effective each block in our developed *MCUa* model.

5.4.1 Dataset

In this experimental study, we used BACH dataset which is part of ICIAR 2018 challenge for classification of H&E stained breast cancer histology images. The dataset is composed of two parts (namely Part A and Part B). Part A of the dataset is composed of 400 sections of microscopy images that are equally distributed among four classes (normal, benign, in situ, and invasive). On the other hand, Part B is composed of 10 high-resolution whole slide images, where the annotations are provided for a semantic segmentation task. In this work, we focused on Part A of the dataset to evaluate the performance of the classification models. The dataset was annotated by two medical experts and all microscopy images are relevant to different patients. The total number of patients involved in the production of the dataset was 39. The anonymisation process of the dataset does not allow to retrieve the origin of all images. All the microscopy images have the same size of 2048 $\times$ 1536 pixels at 20X magnification level (where, the pixel resolution of the images is 0.42 μm).

We evaluated the performance of *MCUa* model using 400 training images with stratified five-fold cross-validation. To train and fine-tune patch-wise networks (i.e. pre-trained DCNNs), we used microscopy patches extracted from training images which are augmented using different rotations and reflections. We evaluated the performance of the ensemble of patch-wise networks using the validation set before stacking context-aware networks on the top of patch-wise networks. Likewise, for context-aware models, which are stacked on the top of patch-wise networks, we followed the same training process conducted in patch-wise networks.

5.4.2 Hyperparameter Settings

For multi-scale image features, we managed to try different images scales including the scale of the original image. Based on a comprehensive experimentation as well as the recommendation of the work conducted in [120], we decided to resize the original image (of the size 2048 x 1536) to 448 x 336 (scale 1), and 296 x 224 (scale 2). To extract image patches from the multi-scale resized images, we utilised sliding window technique of size $p_w = p_h = 224$. Also, we set the stride (at scale 1) to 28 and 56 for training data extraction and testing data extraction, respectively. For scale 2, we set the stride to 9 and 18 for training data extraction and testing data extraction, respectively. In this work, for a fair comparison, we followed the same hyperparameter settings as pointed out in [120], where the same backbone networks were used.

The overlapped extracted patches are then fed into the pre-trained DCNN models. We used DenseNet-161 for scale 1 and 2, while ResNet-152 is utilised for scale 1 only. This gives three ensemble pre-trained feature extractors. The choice of these three pre-trained DCNNs with the associated image scales was aligned with the conclusion that has been drawn from the work conducted in [120]. An ablation study was conducted in [120] using several different ImageNet pre-trained networks. The study has included different image scales (2048 x 1536, 1024 x 768, 448 x 336, and 296 x 224) for the BACH dataset and different pre-trained networks (DenseNet-161, ResNet-101, and ResNet-152). They also considered different combinations of the fine-tuned DCNN models (with different image scales) for the ensemble modeling. Our work utilised the optimal combination recommended by their study, which is using DenseNet-161 for scale 1 and 2, while using ResNet-152 for scale 1. In the training process, we applied data augmentation for each patch by applying rotation operation of 90 degrees with/without vertical flipping. This results in eight versions

of a single patch. We set the learning rate to 0.0001 for 5 training epochs with a batch size of 32.

The feature maps extracted from each pre-trained DCNN are then inserted into multi-level context-aware models which present different levels of contextual information. We utilised six multi-level context-aware models for each pre-trained DCNN giving us a total number of 18 context-aware models. Based on initial experimentation, we designed our *MCUa* model in a constructive way by experimenting a group of 3 context-aware models until reaching the total number of context-aware models represented in this work. In our experiments, we considered the amount of GPU memory available and, at the same time, covering different prominent levels of spatial dependencies, different pre-trained DCNN models, and different image scales when choosing the total number of context-aware models.

For context-aware networks, we utilised stride s = 112 for scale 1 and s = 9 for scale 2. The stride values are chosen after comprehensive experimentation to pick up the suitable values which give higher accuracy as well as improving the contextual assumption for *MCUa* model. The settings for a context-aware network are exactly like the pre-trained DCNN settings except that we used 10 training epochs and batch size equals to 8. For data augmentation, we used same transformations applied for pre-trained DCNN models, but using rotation operation of 180 degrees. Moreover, as overfitting is a major problem in this network, dropout was used with 0.7 rate.

As a final stage, for each image, the most certain models have been selected and combined, in a dynamic way. To implement this, we utilised MC-dropout with a total number of 50 test passes (which is sufficient to generate a statistically valid distribution) for each image. Based on the mean and standard deviation of the 50 distributions, we used the mean to produce the final prediction, while standard deviation was used as a measure of uncertainty. The dynamic picking of context-aware models is performed using δ threshold which ranges from 0.001 to 1.75.

5.4.3 Performance Evaluation

We adopt accuracy, precision, recall and F1-score. Precision is intuitively the ability of the classifier not to label as positive a sample that is negative, recall is the ability of the classifier to find all the positive samples and F1-score can be interpreted as a weighted average of the precision and recall. We computed the accuracy, precision, recall and F1-score as shown in equations 4.5, 4.6, 4.7, and 4.8, respectively.

Performance of a Single Context-aware Model

Table 5.1 presents the classification accuracy for our individual context-aware models that have been designed on the top of three pre-trained DCNNs (e.g. DenseNet-161 using two image scales 448 $\times$ 336 (scale 1) and 296 $\times$ 224 (scale 2) and ResNet-152 using image scale 1). The context-aware models are implemented based on different pattern levels and shape indices (P_{2,S_1}, P_{3,S_1}, P_{4,S_1}, P_{4,S_2}, P_{5,S_1}, P_{6,S_1} and P_{8,S_1}). Based on trial and error experiments, we excluded P_{7,S_1} as it gives lower accuracy compared to the other pattern levels. Also, to demonstrate the idea of using different shapes within the same pattern level, we experimented P_{4,S_1} and P_{4,S_2}, where each one of them has a unique set of shape indices. As illustrated by Table 5.1, the highest classification accuracies are obtained by P_{2,S_1}, P_{4,S_2} and P_{5,S_1} with the three pre-trained DCNNs. Moreover, most of the context-aware models for image scale 1 achieved

a classification accuracy which varied between 93% and 94.75%, while the context-aware models for image scale 2 achieved less accuracy ranging between 88.75% and 90.25%.

Static *MCUa* Model

We have presented the accuracy, precision, recall, and AUC of the static ensemble context-aware architecture (i.e. ensemble of the total 18 models) to distinguish each category of images and overall classification accuracy in Table 5.2. As illustrated by the table, invasive carcinoma tissues and benign tissues can be differentiated clearly from other classes. We achieved an average precision of 95.90% $\pm$ 2.40% and an overall classification accuracy of 95.75% $\pm$ 2.44%, which illustrates the viability of our architecture in classifying breast histology images.

Static vs. Dynamic *MCUa* Model

To demonstrate the sensitivity of *MCUa* to the uncertainty quantification component, we studied the performance of the static ensemble of context-aware models and our dynamic ensemble mechanism. For a fair comparison, we utilised two other metrics: (1) Weighted Average Accuracy (WA_{ACC}), which computes accuracy for each fold of the 5 folds weighted by the number of included images in that fold and after that it averages the weighted accuracies of 5 folds over the total number of included images all over the dataset; and (2) abstain percentage (Abs), which calculates the percentage of excluded images in the dataset through different δ values. We formulated WA-ACC and Abs as:

$$WA_{ACC} = \frac{1}{\sum_{i=1}^{r} w_i} \sum_{i=1}^{r} Accuracy_i \times w_i, \tag{5.4}$$

$$Abs = \left(\frac{\sum_{i=1}^{r} \sum_{j=1}^{h} X_{ij}^{''}}{D_{all}} \right) \times 100, \tag{5.5}$$

where $Accuracy_i$ represents classification accuracy i over r folds, w_i is the weight of the included images in fold i, $X_{ij}^{''}$ represents excluded image j over h excluded images in fold i, and D_{all} is the total number of images in BACH dataset.

Table 5.3 illustrates the effectiveness of *MCUa* by improving the classification accuracy obtained by static ensemble mechanism. As demonstrated by Table 5.3, *MCUa* has achieved WA_{ACC} of 98.11% with δ of 0.001 and around 97.70% with δ values of 0.002, 0.003, 0.006 and 0.02.

Figure 5.4 depicts WA_{ACC} curve for included images, Abs, and WA_{ACC} curve for excluded images on BACH dataset, respectively, over δ ranges from 0.001 to 1.75. The WA_{ACC} curve for the included images shows that the best WA_{ACC} is achieved when the δ value is low and the accuracy starts to decrease with increasing δ values until it reaches 0.1. Moreover, as shown by the same figure, the accuracy increases until settling at 95% with δ value of 0.5 to 1.75. On the other hand, Abs curve shows that the percentage of abstained images decreases when we use higher δ values, and starting from 0.25, the number of excluded images dropped to zero. Finally, the WA_{ACC} curve for excluded images shows the performance of *MCUa* model using static ensemble, where the accuracy was around 80% for small δ and then the accuracy starts to drop until reaching 50% at δ of 0.1. The accuracy is zero when the number of excluded images equals to zero. This demonstrates that excluded images

TABLE 5.1: Classification Accuracy for context-aware models based on different pattern levels using stratified five-fold cross-validation on BACH dataset (%).

Pre-trained DCNN (Image Scale)	Context-aware Pattern Levels - Accuracy						
	P_{2,S_1}	P_{3,S_1}	P_{4,S_1}	P_{4,S_2}	P_{5,S_1}	P_{6,S_1}	P_{8,S_1}
DenseNet-161 (Scale 1)	93.75	93.00	93.50	93.25	93.50	93.25	–
DenseNet-161 (Scale 2)	89.00	89.75	–	90.25	89.75	88.75	90.25
ResNet-152 (Scale 1)	94.00	93.25	93.50	94.75	94.75	93.75	–

TABLE 5.2: Performance (mean ± standard deviation) of *MCUa* (static ensemble) on BACH Dataset with stratified five-fold cross-validation (%).

Category	Precision	Recall	F1-score	Accuracy
Normal	93.32 ± 5.34	95.00 ± 5	94.07 ± 4.10	97.00 ± 2.09
Benign	96.00 ± 4	95.00 ± 5	95.45 ± 4.15	97.75 ± 2.05
InSitu	95.28 ± 4.68	96 ± 2.24	95.56 ± 1.98	97.75 ± 1.04
Invasive	99.00 ± 1	97 ± 2.74	97.97 ± 2.10	99.00 ± 1
Total	**95.90 ± 2.40**	**95.75 ± 2.44**	**95.77 ± 2.42**	**95.75 ± 2.44**

TABLE 5.3: Accuracy (%) of *MCUa* model with both static and dynamic ensemble on BACH dataset.

Method	δ	Accuracy
MCUa (Static Ensemble)	NA	95.75
MCUa (Dynamic Ensemble)	0.001	**98.11**
	0.002	97.93
	0.003	97.60
	0.006	97.65
	0.01	97.53

are typically harder to classify, and may well require a pathologist to make an expert judgment.

Comparison with Recent Methods

In Table 5.4, we compare the performance of our model with the following state-of-the-art recent methods: (1) a two-stage CNN proposed by Nazeri et al. [73], which consists of patch-wise network for feature extraction and image-wise network for image level classification, (2) a context-aware learning strategy using transferable features, which is based on a pre-trained DCNN and SVM for classification [9], (3)

TABLE 5.4: Performance (mean ± standard deviation) comparison of the *MCUa* model and recent methods on BACH Dataset (%).

Method	Precision	Recall	F1-score	Accuracy
Two-Stage CNN [73]	86.35 ± 2.70	85.50 ± 3.38	85.49 ± 3.25	85.50 ± 3.38
DCNN + SVM [9]	86.88 ± 1.52	85.75 ± 1.90	85.58 ± 1.92	85.75 ± 1.90
Bayesian DenseNet-169 [72]	89.28 ± 4.71	88.50 ± 5.03	88.45 ± 5.05	88.50 ± 5.03
Deep Spatial Fusion CNN [42]	89.93 ± 4.11	89.00 ± 3.89	88.93 ± 4.02	89.00 ± 3.89
Variational Dropout ARA-CNN [79]	90.25 ± 2.87	89.50 ± 3.14	89.48 ± 3.13	89.50 ± 3.14
ScanNet + Feature Aggregation [113]	90.90 ± 3.87	90.50 ± 3.81	90.46 ± 3.86	90.50 ± 3.81
Hybrid DNN [119]	91.79 ± 3.50	91.00 ± 3.46	90.98 ± 3.45	91.00 ± 3.46
EMS-Net [120]	95.23 ± 2.13	95.00 ± 2.17	94.98 ± 2.13	95.00 ± 2.17
3E-Net [90]	95.68 ± 3.15	95.46 ± 3.22	95.45 ± 3.24	95.46 ± 3.22
MCUa (Static Ensemble)	**95.90 ± 2.40**	**95.75 ± 2.44**	**95.77 ± 2.42**	**95.75 ± 2.44**
MCUa (Dynamic Ensemble ($\delta = 0.001$))	**98.25 ± 1.58**	**98.11 ± 1.77**	**98.10 ± 1.78**	**98.11 ± 1.77**

Bayesian DenseNet-169 proposed by Mobiny and Singh [72], which generates uncertainty measure for input images, (4) deep spatial fusion CNN introduced by Huang and Chung [42], which uses patch-wise residual network for feature extraction and deep spatial fusion network that has been designed to capture the spatial relationship among image patches using the spatial feature maps, (5) ARA-CNN introduced by Raczkowski et al. [79], which uses variational dropout during the testing phase to measure the uncertainty of input images, (6) ScanNet with feature aggregation method of [113], which applies feature extraction and concatenation towards the final classification, (7) Hybrid DNN introduced by Yan et al. [119] which uses inception network for feature extraction of image patches along with bi-directional LSTM network which learns contextual information among feature maps generated from inception network, (8) EMS-Net proposed by Zhanbo et al. [120], which applies an ensemble of pre-trained DCNNs, and (9) *3E-Net* Version A [90] which builds an ensemble of image-wise networks with a measure of uncertainty using Shannon entropy to pick the most certain image-wise models for the final image classification. As demonstrated by Table 5.4, our model outperformed other models when both static and dynamic ensemble mechanisms are used. Moreover, Figure 5.6 illustrates ROC curves for *MCUa* (with both dynamic and static ensemble) to confirm the superiority of our presented solution. Consequently, the importance of integrating multi-level contextual information into DCNNs, to alleviating the high visual variability in breast histology images, has been emphasised and experimentally proofed.

Performance of *MCUa* on BreakHis Dataset

To confirm the effectiveness of our solution, we applied *MCUa* model on the Breast Cancer Histopathological Database (BreakHis) [102]. BreakHis has 7909 breast cancer histology images collected from 82 patients, obtained with different magnification levels (40X, 100X, 200X, and 400X). The dataset consists of 2480 benign and 5429 malignant microscopic images with resolution of 700 $\times$ 460 pixels each. We used images with a magnification level of 40X, which included 625 benign and 1370 malignant samples (1995 microscopic samples in total).

In this study, we used the same hyperparameter settings that we used for the BACH dataset. For example, we down-sampled the original input image (700 $\times$ 460) to two image scales (scale 1: 448 $\times$ 336 and scale 2: 296 $\times$ 224). These image scales are fed as input to the pre-trained DCNN models (DenseNet-161 and ResNet-152) for extraction of features from image patches. The extracted features are then inserted into 18 context-aware models to learn the spatial relationships among the image patches. We used the same patch stride and data augmentation settings (that has been applied to the BACH dataset) for both feature extraction and context-aware modeling networks. As BreakHis dataset has 2 classes (Benign and Malignant), we fine-tuned the pre-trained DCNN models by modifying the number of neurons of the last fully connected layer to only two neurons. As shown in Table 5.5, *MCUa* demonstrated to be effective in both static and dynamic techniques. Using 5-fold cross-validation, we achieved a classification accuracy of 99.80% using the static ensemble technique. The model has achieved exceptional classification accuracies of 100%, 99.95%, and 99.90% using dynamic ensemble on δ values of 0.001, 0.003, and 0.03, respectively. Figure 5.5 depicts the WAA and Abs curves for *MCUa* using BreakHis dataset.

Ablation Study

In this work, we describe the ablation study that we conducted to reach the final version of the building components of our *MCUa* model. All the conducted experiments in this ablation study are validated with BACH dataset. As an initial step towards our final version of *MCUa*, we implemented a single DCNN with a target to learn multi-scale and multi-level feature patterns. This is accomplished by using multiple patch scales (224 x 224, 112 x 112, 56 x 56, and 28 x 28) to identify different nuclei sizes in histology images. Then, we utilised all the feature maps extracted from the aforementioned patch scales by applying fusion for the multi-scale, multi-level feature maps for final classification. The single DCNN was built using a sequence of 3 x 3 filters in the convolutional layers, followed by a pooling layer, with the number of channels doubled after each down-sampling. We used 2 x 2 filters in the convolutional layers with stride of 2 for down-sampling the feature maps. Batch normalisation and ReLU activation were used after all convolutional layers. Finally, a fully connected layer followed by softmax layer are used to produce the final image classification. We applied stratified 5-fold cross-validation and achieved classification accuracy of 87.50%.

In another trial, we implemented single DCNNs to extract feature maps from image patches, learn spatial dependencies among image patches arranged in a certain pattern, and generated the final image classification. We used DenseNet-161 with image scales (scale 1: 448 $\times$ 336 and scale 2: 296 $\times$ 224) and ResNet-152 with image scale (scale 1: 448 $\times$ 336) as the single DCNNs in this study. We applied stratified 5-fold cross-validation, and we achieved a classification accuracy of 93.00%

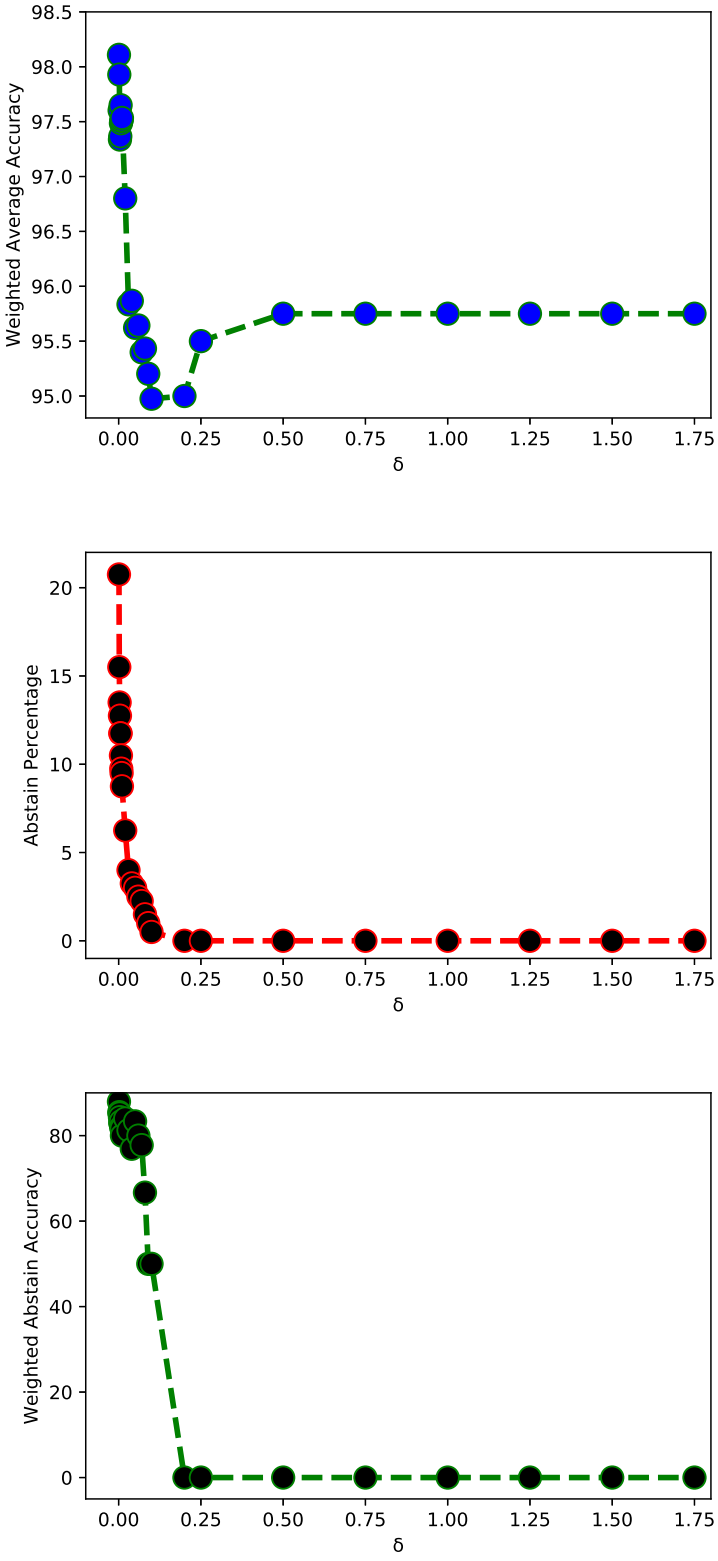


FIGURE 5.4: Weighted average accuracy (WA_{ACC}) for the included images (top), abstain percentage (Abs) (middle) and (WA_{ACC}) (bottom) for the excluded images on BACH dataset.

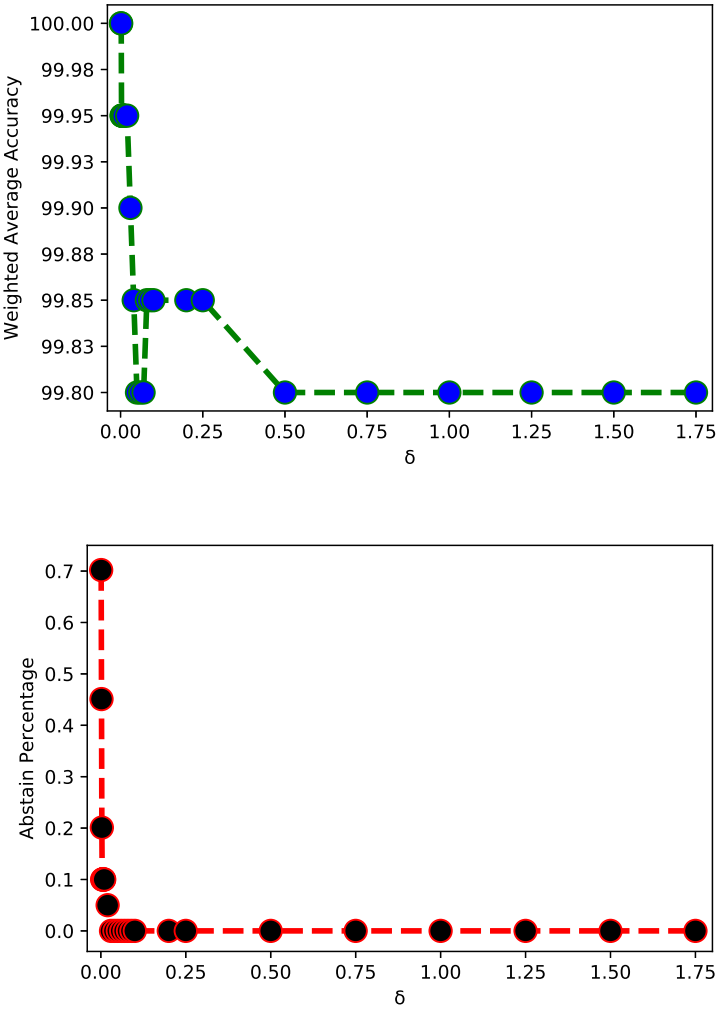


FIGURE 5.5: WA_{ACC} for the included images (top), abstain percentage (*Abs*) (bottom) for the excluded images on BreakHis dataset.

TABLE 5.5: Accuracy (%) of *MCUa* model with both static and dynamic ensemble on BreakHis dataset.

Method	δ	Accuracy
MCUa (Static Ensemble)	NA	99.80
MCUa (Dynamic Ensemble)	0.001	**100**
	0.003	99.95
	0.03	99.90
	0.04	99.85

and 88.50% for DenseNet-161 with scales 1 and 2, respectively, while, ResNet-152 using scale 1 yielded a classification accuracy of 94.50%. Although the aforementioned methods are straightforward and easy to implement, we argue that single

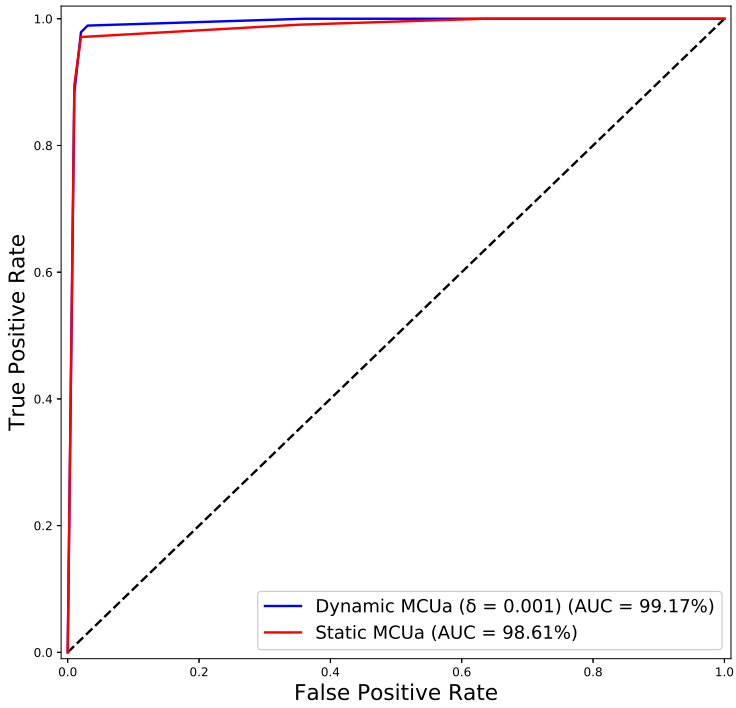


FIGURE 5.6: Receiver Operating Characteristic (ROC) curves for the static and dynamic methods of *MCUa* Model using 5-fold cross-validation on BACH dataset.

DCNNs lack diversity in generating discriminative features which is vital in the usage of ensemble strategy. This helps to generate features from multi-scale and multi-architecture perspectives to help in representing multi-level haematological objects (such as nuclei and glands) within the histology images.

Consequently, we applied an ensemble of three pre-trained single DCNNs with two image scales and achieved a classification accuracy of 95.00%. Furthermore, instead of using the pre-trained DCNNs for classification task, we used them for feature extraction of image patches, then we stacked 18 context-aware models over the three pre-trained DCNNs. The ensemble process of 18 context-ware models yielded a classification accuracy of 95.75% (*MCUa* static ensemble).

In the final stage of *MCUa*, we evaluated the contribution of uncertainty-aware component, which is stacked over 18 context-aware models. This strategy introduces the machine-confidence in the automated prediction of histology images. The full version of *MCUa* (based on the uncertainty-aware component) yielded a classification accuracy of 98.11%. This justifies the effectiveness of using multi-scale input, multi-architecture feature extraction, multi-level context-aware modeling, and uncertainty quantification for the dynamic ensemble mechanism.

5.5 Summary

In this chapter, we introduced a novel dynamic ensemble of context-aware models, we called Multi-level Context and Uncertainty aware (*MCUa*) model, to classify H&E stained breast histology images into four classes including normal tissue, benign lesion, in situ carcinoma and invasive carcinoma. *MCUa* model has been designed in a way to learn the spatial dependencies among the patches in a histology image by integrating multi-level contextual information into the learning framework of deep convolutional neural networks. Capturing spatial relationships among image patches has been accomplished using a pattern of neighbourhood criteria through multiple context-aware models. *MCUa* model has also an uncertainty quantification component that allows for a dynamic ensemble of the context-aware model to not only improve the performance (by improving the learning diversity of the model) but also quantify the difficulties in classifying images. *MCUa* has achieved high accuracy of 95.75% and 98.11% with both static ensemble and dynamic ensemble mechanisms, respectively, on the BACH dataset, and outperformed other related state-of-the-art models.

MCUa model reflects a robust development of a deep learning architecture that can be used as a reliable diagnostic tool. This model is considered a heavy-weight model compared to *3E-Net* as it requires more time during inference and needs more resources. Consequently, *MCUa* can be deployed on a server-based device making it highly efficient in terms of generating accurate predictions and actionable decisions. *MCUa* model distinguishes itself from the *3E-Net* model in several key ways. For instance, it has the capability to accurately detect nuclei objects of diverse dimensions in histopathological images by employing multi-scale inputs and employing a variety of architectural designs for feature extraction. Furthermore, *MCUa* model is capable of learning contextual information at multiple scales and levels. Lastly, it incorporates an uncertainty component, based on MC dropout, which generates list of predictive probability distributions, instead of a single predictive probability distribution as introduced in *3E-Net*.

3E-Net and *MCUa* proved their effectiveness on two different types of deep learning applications: grading and diagnostics, respectively. The selection process of uncertain images which is based on either Shannon Entropy (*3E-Net*) or MC dropout (*MCUa*) are being managed by manual actionability (i.e. manual tuning for threshold value to identify a margin for certain and uncertain images) which is good practice to identify how our models behave when we utilise different values of threshold. This process may lead in excluding lower or higher number of images based on the manual value of threshold set. Therefore, in the next chapter, we introduce an automated actionable method for optimising uncertainty quantification for deep learning architectures (*AUQantO*) that utilises multiple optimisation methods for optimising single and multi-objective functions that aim to find the optimal number of excluded images based on searching for the optimal threshold. This method can be applied to any deep learning architecture that produces probability distribution as sort of output prediction.

Chapter 6

AUQantO: Actionable Uncertainty Quantification Optimisation for Medical Image Classification

In the last chapter, we explained in detail *MCUa* for classification of breast histopathology microscopic images as the second contribution of the thesis. *MCUa* utilised a detailed strategy for building automated diagnosis system. This strategy proved to be highly effective for clinical practice, as it introduces advanced method for uncertainty quantification along with diversity for image scales, feature extraction and context-aware information. The selection process of uncertain images in *3E-Net* and *MCUa* is managed manually by setting a threshold to identify a margin for certain and uncertain images. While this is a good practice to understand the models' behaviour under different threshold values, it may lead to excluding a higher number of images based on the manual value setting. Therefore, in this chapter, we present a model agnostic method for optimising uncertainty quantification for image classification in deep learning models. This method features a fully automatic mechanism to select optimal hyperparameter settings (threshold) for identifying the optimal number of images to be excluded from a particular dataset based on uncertainty quantification methods. Findings reported in this chapter is to be published in [89].

6.1 Overview

Deep learning algorithms have the potential to automate the examination of medical images obtained in clinical practice. Using digitised medical images, convolution neural networks (CNNs) have demonstrated their ability and promise to discriminate among different image classes. As an initial step towards explainability in clinical diagnosis, deep learning models must be exceedingly precise, offering a measure of uncertainty for its predictions. Such uncertainty-aware models can help medical professionals in detecting complicated and corrupted samples for re-annotation or exclusion. In this chapter, we introduce a novel model and data agnostic mechanism, we called Actionable Uncertainty Quantification Optimisation (*AUQantO*), for optimising deep learning architectures' performance for medical image classification. This is by optimising the hyperparameter of entropy-based and Monte-Carlo (MC) dropout uncertainty quantification techniques escorted by single and multi-objective optimisation methods, abstaining classification of images with a high level of uncertainty. *AUQantO* has been validated with four deep learning architectures on two medical image datasets. The results show significant improvements in the

performance of the state-of-the-art deep learning approaches with the ability in generating the optimal number of excluded images based on their level of prediction's confidence.

The chapter is structured as follows. In section 6.2, we give an introductory about the background and the developed work. Section 6.3 discusses, in detail, our *AUQantO* method. Our experimental results and findings are explained in Section 6.4. Section 6.5 summarises our work.

6.2 Introduction

Advances in computer-aided diagnosis (CAD) have a substantial impact on reducing the strain of medical practitioners doing manual investigations and enhancing detection accuracy for various diseases. One of the methods that has been used extensively for automated diagnosis is deep learning. Deep learning approaches have gained gigantic headway and accomplished exceptional outcomes, driving numerous researchers to give automated and fair solutions for a few diverse medical image analysis applications. For the image classification task, Deep Convolutional Neural Networks (DCNNs) are considered as one of the deep learning approaches that have been commonly used to extract prominent image features for several medical image analysis applications [96].

Despite the capability of DCNNs to demonstrate outstanding performance for image classification tasks [42, 73, 120], an initial stage of explainability is required to measure the level of uncertainty in the input samples for medical image analysis applications. Building an uncertainty-aware model can help in identifying any ambiguity that could be therapeutically useful. Uncertainty awareness is also beneficial in terms of actionability to medical samples which could be possibly confusing and challenging to automated diagnosis systems. It additionally permits clinical experts to rate images that ought to be focused on for manual examination.

In this chapter, we present a model agnostic mechanism, coined Actionable Uncertainty Quantification Optimisation (*AUQantO*), to optimise the performance of deep learning architectures for medical image classification. *AUQantO* is guided by uncertainty measurements which assist clinical experts with refining annotations for developing more reliable DCNN models. *AUQantO* employs either an entropy-based mechanism [93] or a Monte-Carlo (MC) dropout [29] technique to measure uncertainty in images, where a new hyperparameter (i.e. a threshold) is introduced and optimised. Our motivation stems from the notion that, despite the abundance of deep learning architectures and their significant potential for reducing the workload strain on medical experts, a small percentage of low quality or indecisive medical images would necessitate the aid of medical experts. The performance of *AUQantO* has been validated using state-of-the-art deep learning architectures (with several optimisation methods) on two medical image datasets.

The contributions of the chapter can be summarised as below:

- introduced an optimised automated actionability component to deep learning architectures, which guides medical experts in identifying contaminated samples for re-annotation or exclusion;
- developed a model and dataset agnostic uncertainty-aware method to improve the usage of deep learning models in clinical practice; and

- conducted a comprehensive experimental study using different deep learning models escorted by two uncertainty measures applied to two publicly available medical imaging datasets.

6.3 *AUQantO* Method

We explain our developed approach (*AUQantO*) for optimising uncertainty quantification in deep learning architectures in this section. As illustrated by Fig. 6.1, an input image is fed into the deep learning architecture for the classification. As a pre-stage to our method, *AUQantO* requires deep learning architectures that can generate probability distributions for their input samples. This requirement helps *AUQantO* to generate an uncertainty score for the image's probability distribution and decides the poor medical samples that need to be manually investigated by medical experts. Consequently, *AUQantO* (as an uncertainty-aware method) has been designed based on Shannon Entropy [93] or MC Dropout [29]. Shannon Entropy is based on the image predictions (or the probability distribution of the output, where each value is associated with a class in the training set) generated by the deep learning architecture. Shannon Entropy is adopted to generate an uncertainty score for how confident the model is in classifying the input image. On the other hand, MC dropout uses dropout layers in the deep learning architecture network for image classification and activates them during the testing phase, resulting in a list of probability distributions whose mean prediction determines the image's final classification while the standard deviation provides a measure of uncertainty. To automatically exclude the poor image samples and keep the confident ones for final classification, *AUQantO* introduces a new hyperparameter, we called threshold (λ). In this work, the optimal threshold value is explored by single and multi-objective optimisation methods. *AUQantO* can quantify uncertainty in medical image samples and automatically tune the threshold hyperparameter against uncertainty values to exclude the highly uncertain images.

6.3.1 Uncertainty Measure

Image predictions generated by the deep learning architectures of *AUQantO* are used for quantifying the uncertainty of the predictions. This is by adopting uncertainty quantification methods, such as Shannon Entropy and MC dropout, to measure the level of uncertainty in classifying an input image. Our choice of Shannon Entropy and MC dropout comes from the fact that we deal with probability distributions and the need to measure the level of randomness in the obtained image's predictions. These uncertainty measures help in the excluding process of poor samples by indicating the ambiguity occurring in either a probability distribution of an input image (Shannon Entropy method) or a generated list of probability distributions from an input image (MC dropout method).

Shannon Entropy

This uncertainty quantification method works by receiving the image prediction generated from a deep learning architecture as an input and generating an uncertainty score. This uncertainty score reflects the level of confusion that comes from the probability values produced for every class label in the training set. A low uncertainty score is obtained when the deep learning architecture is highly confident (in terms of the prediction) with a high probability value associated with a certain

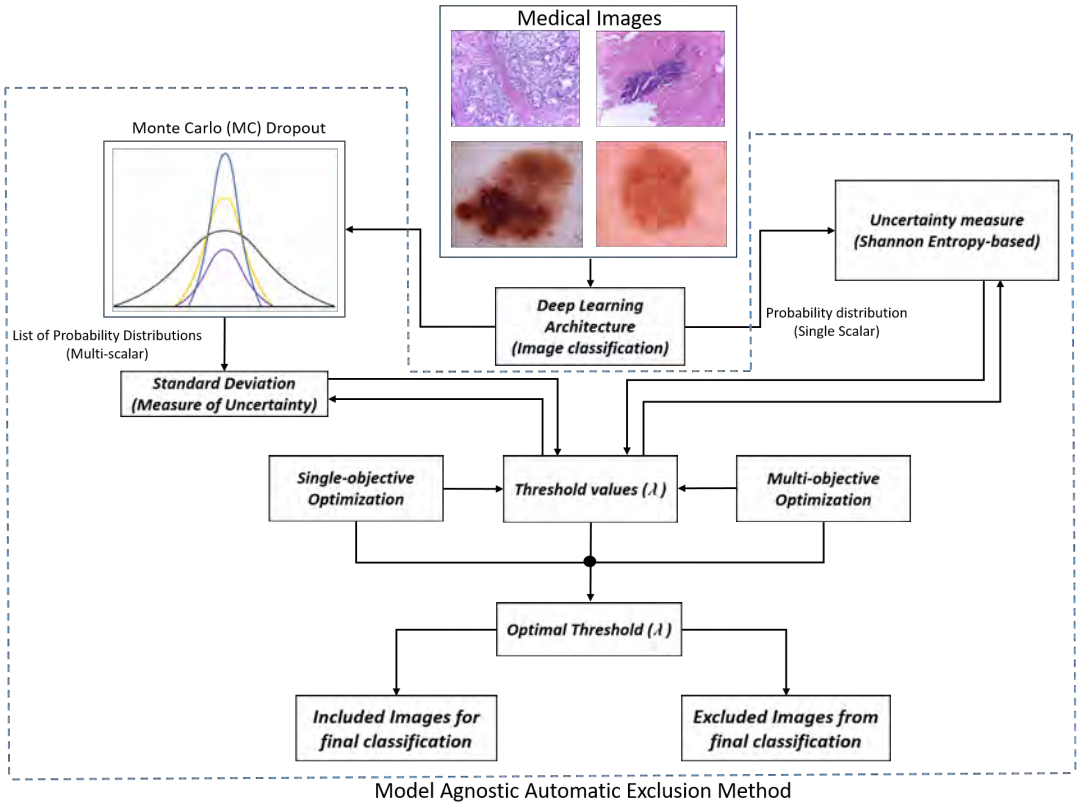


FIGURE 6.1: The workflow of *AUQantO* method. Our method deals with two different uncertainty-awareness techniques. *AUQantO* deals with two types of image predictions. First, *AUQantO* processes a single probability distribution associated with an input image by applying Shannon entropy to generate an uncertainty score. Second, *AUQantO* processes a multi-scalar probability distribution associated with the MC dropout technique. In other words, our method processes a list of probability distributions generated from activating dropout layers throughout the testing process of a deep learning architecture associated with the image classification task. A calculated standard deviation from the list of generated probability distributions for a particular input sample is considered as the measure of uncertainty. Then, for both uncertainty measures, *AUQantO* optimises a hyperparameter or threshold value (λ), of the developed uncertainty quantification component. The importance of λ comes from its effectiveness in selecting whether an image would be included in the final classification (as a strong candidate) or to be excluded as a poor sample which needs further investigation by medical experts.

class label, while a high uncertainty score refers to the conflict that can occur when multiple probability values are quite similar. The formulation of Shannon Entropy is presented in equation 4.4. Detailed description of Shannon Entropy is present in section 4.3.3.

Bayesian Approximation using MC Dropout

Unlike Shannon entropy quantification method, MC dropout works by applying dropout layers to the deep learning architectures and activating these layers during

the testing phase. MC dropout method features a key hyperparameter that regulates the number of times the network is tested (test passes). During each testing pass, MC dropout randomly activates a subset of neurons, leading to alterations in the output probability distribution (or predictions) associated with the input image. This random activation of neurons during testing provides a means of measuring uncertainty in the predictions made by the network. As a consequence, a list of probability distributions for each input image is generated where their average is used to produce the final class prediction and their standard deviation reflects the level of uncertainty. The formulation of mean prediction and standard deviation are presented in equations 5.2 and 5.3, respectively. Detailed description of MC dropout is present in section 5.3.4.

The uncertainty score generated by any of the two above-mentioned uncertainty quantification methods is then verified by a hyperparameter (e.g. λ). The optimal threshold value (which aids in excluding the optimal number of poor samples) is calculated by the minimisation of our objective function(s).

6.3.2 Objective Function

Single-objective Function

A new hyperparameter (λ) has been introduced by our objective function, to be tuned based on the generated uncertainty scores. This is by checking if the input image has an uncertainty score greater than the λ then the image will be excluded from the final classification process, otherwise, the image will be classified by the model. More precisely, to calculate the optimal threshold value (λ), we introduce a single-objective function to be minimised. Our objective function has two terms that have been designed to encode the confidence of probability distributions for both included and excluded images. We used cross-entropy for the probability distributions against the ground truth labels. For example, we customised the cross-entropy equation by multiplying the probability distribution of a given image by the one hot-encoding labelling for the same image. Consequently, we formulated H_{exc} and H_{inc} to present a summation of cross-entropy values for excluded and included images, respectively. H_{exc} and H_{inc} can be represented as:

$$H_{exc} = \sum_{i=1}^{n} \sum_{j=1}^{c} p_{ij} \times q_{ij} \tag{6.1}$$

$$H_{inc} = \sum_{i=1}^{m} \sum_{j=1}^{c} p_{ij} \times q_{ij} \tag{6.2}$$

where p_{ij} represents the probability value j over c class probability values, while q_{ij} represents the one hot-encoding value j over c class categories of image i over either n excluded images or m included images.

The average cross-entropy for both excluded and included images is then calculated by dividing H_{exc} and H_{inc} by n excluded images and m included images, respectively. Using single-objective optimisation methods, the main target is to reach an optimal λ which minimises the summation of the two terms of the objective function. For example, a possible scenario to minimise the cross-entropy of excluded images H_{exc} is to have cases where images are misclassified with high confidence. This means that the evaluation of cross-entropy equation (assuming we have a classification problem of two classes) will have a very small probability value p (tends to zero) for the correct class multiplied by $q = 1$ to represent the one hot-encoding of

correct class. While, for incorrect class, a very high probability value p is multiplied by $q = 0$. A similar scenario could happen for the maximisation of the cross-entropy of included images H_{inc} by having images that are correctly classified with high confidence and by subtracting this term from a value of one, we convert it into a term to be minimised instead. Both scenarios for included and excluded images lead to a very small value for the output of the objective function and hence we can reach high level of trustworthiness for included images that are classified by a deep learning architecture and exclude images that are truly uncertain with high confidence for further annotation and investigation by medical experts. The single-objective function can be defined as:

$$F(\lambda) = argmin_{(SE|\sigma<\lambda||SE|\sigma\geq\lambda)} \left(\left(\frac{H_{exc}}{n} \right) + \left(1 - \frac{H_{inc}}{m} \right) \right) \tag{6.3}$$

where $F(\lambda)$ is the output of the single-objective function and λ is the optimal threshold value. λ is verified by Shannon Entropy SE or MC dropout's standard deviation σ to differentiate between included and excluded image groups and measure the average cross-entropy.

Multi-objective Function

As can be noticed from the above-mentioned single-objective function, that we have two terms to work on both included and excluded images. The two terms can be presented in two separate objective functions that can be optimised simultaneously to reach the optimal threshold which achieves the selection of (1) highly certain images to be included in the final classification and (2) highly uncertain images to be excluded from classification and to be returned to medical experts for manual exploration. In that sense, we introduce a multi-objective function with the target of maximising the rate of included images and minimising the rate of excluded images based on their uncertainty and confidence of deep learning architecture's predictions.

Our multi-objective function can be defined as:

$$\min \{F_{exc}(\lambda), F_{inc}(\lambda)\} \tag{6.4}$$

where:

$$F_{exc}(\lambda) = H_{exc}/n$$
$$F_{inc}(\lambda) = 1 - (H_{inc}/m)$$

subject to:

$$n \geq 1$$
$$\lambda \leq SE|\sigma \quad for\, F_{exc}(\lambda)$$
$$m \geq 1$$
$$\lambda > SE|\sigma \quad for\, F_{inc}(\lambda)$$
$$\lambda_{\min} \leq \lambda \leq \lambda_{\max},$$

where $F_{exc}(\lambda)$ and $F_{inc}(\lambda)$ represent the objective functions which are based on the average cross-entropy of excluded and included images, respectively. The number of excluded images n and included images m in the multi-objective function are subject to number of images not less than value of one. Moreover, λ value is subject to a pre-defined range ($\lambda_{\min}$ to $\lambda_{\max}$) while using a multi-objective optimisation method.

6.3.3 Optimisation methods

Our optimisation problem is based on a non-convex objective function. Non-convex optimisation refers to the optimisation problems where the objective function to be minimized or maximized is not a convex function. A convex function is a mathematical function that is always above its tangent, meaning that its slope does not change abruptly, resulting in a single global minimum. In contrast, non-convex functions have multiple local minima, meaning that they can have multiple regions where the function value is the lowest. In the context of the given explanation, the exclusion rate of images can affect the accuracy of the deep learning architectures in a non-linear and unpredictable manner. This means that increasing or decreasing the exclusion rate may not always lead to higher accuracy. As a result, the objective function that represents the relationship between the exclusion rate and accuracy is non-convex and has multiple local minima. In this case, the search for the optimal threshold hyperparameter will be conducted randomly within a search space, which means that different solutions will be explored randomly in the hope of finding the global optimum (stochastic objective function). However, due to the non-convex nature of the objective function, there is no guarantee that the optimum found will be the global optimum, but rather a local optimum. All the optimisation methods adopted in this work are known to be effective when applied to non-convex optimisation problems [105].

Bayesian Optimisation using Gaussian Processes (GP)

For function spaces with domains that are continuous, discontinuous, mixed, or even hierarchical, Bayesian optimisation using GP [81] provides a rich and versatile collection of non-parametric statistical models. The aim is to use GP to approximate the objective function. More precisely, the values of the functions are developed to follow a multivariate Gaussian distribution. GP is considered as stochastic process defined on a continuous domain $\mathcal{X} \subset R^n$. A function F is considered as a GP if for a finite tuple of hyperparameter values $\lambda_{tuple} = (\lambda_1, \cdots, \lambda_k) \subset \mathcal{X}^k$ the random vector $\mathbf{Y} = [F(\lambda_1), \cdots, F(\lambda_k)]^T$ is a multivariate Gaussian random variable. A GP is characterised by two functions: mean function $\mu(\lambda)$ and a covariance function $k(\lambda, \lambda^{'})$. Then, we can present the multivariate probability density as:

$$P(\mathbf{Y}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{Y}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mu)\right], \tag{6.5}$$

where $\mu = [\mu(\lambda_1), \cdots, \mu(\lambda_k)]^T, \mathbf{\Sigma} = [\Sigma_{ij}] = [k(\lambda_i, \lambda_j)]$.

Constrained Optimisation by Linear Approximation (COBYLA)

COBYLA is a derivative-free simplex method introduced by [78]. In general, the simplex technique aims to minimise the objective function by employing simplices, with simplex referring to the convex hull of $s+1$ points in s-dimensional space and s indicating the number of variables. The objective function is evaluated at the vertices of an initial simplex, and the simplex is then changed so that the objective function obtains generally smaller values at the vertices of the new simplex than at the vertices of the prior simplex. Let the initial value of λ be required and let's assume we have function values $F(\lambda_i), i = 0, 1, \cdots, s$, where $F(\lambda_0) \leq F(\lambda_1) \leq \cdots \leq F(\lambda_s)$. An iteration of original simplex method evaluates the next function $F(\lambda)$, where the next point is $\hat{\lambda}$

$$\hat{\lambda} = (2/s) \sum_{i=0}^{s-1} \lambda_i - \lambda_s \tag{6.6}$$

If $F(\hat{\lambda}) < F(\lambda_{s-1})$ has been reached, then $\hat{\lambda}$ takes place λ_s as a vertex of the simplex. Otherwise, a contraction is recommended by retaining λ_0, but λ_i being replaced by $\frac{1}{2}(\lambda_0 + \lambda_i)$ for $i = 1, 2, \ldots, s$, and then s more function values have to be evaluated for the upcoming iteration. In both conditions, the new set of function values has to be arranged into an ascending order style.

Dual Annealing

Dual Annealing is a stochastic global optimisation method. It is a generalised simulated annealing algorithm [116], which is an extension of simulated annealing. In addition, it is associated with a local search algorithm, which runs automatically at the end of the simulated annealing process. This optimisation is designed for objective functions that have a nonlinear response surface. The algorithm mimics the slow cooling of metals, which is characterised by a progressive decrease in atomic movements, which lowers the density of lattice distortion until the smallest energy state is reached. Similarly, the simulated annealing algorithm generates a new feasible solution to the optimisation problem by modifying the existing state according to a specified criterion at each virtual annealing temperature. The newly reached state is approved if it meets the Metropolis criterion, and the procedure is repeated until convergence is reached. Let's consider the objective function $F(\lambda_i)$ with a set of $\lambda = \lambda_1, \cdots, \lambda n$. Each new candidate point λ_{i+1} was approved throughout the annealing process with a temperature-dependent probability P_t determined by

$$P_t = \begin{cases} 1 & \text{if } F(\lambda_{i+1}) \leq F(\lambda_i), \\ e^{\left(\frac{F(\lambda_i) - F(\lambda_{i+1})}{k \times t}\right)} & \text{if } F(\lambda_{i+1}) \geq F(\lambda_i), \end{cases} \tag{6.7}$$

where t is the current temperature, k is the Boltzmann constant, and $F(\lambda_i)$ and $F(\lambda_{i+1})$ are the function values for the current and new point to check.

Non-dominated Sorting Genetic Algorithm (NSGA-II)

NSGA-II was first introduced in [23] as a multi-objective optimisation method. NSGA-II is an evolutionary algorithm that can overcome the drawbacks of traditional direct and gradient-based methods when dealing with non-linearities and complicated interactions. NSGA-II utilises an elitist principle (a population's elites are allowed to pass down to the following generation). It uses an explicit mechanism for maintaining diversity (crowding distance) and focuses on non-dominant solutions.

NSGA-II algorithm starts by applying a non-dominated sorting in the pair of parent and offspring populations and classifying them by fronts (sorted based on ascending level of non-domination). Then, front-ranking is used to generate a new population. Crowding distance sorting is then applied when one front is partially taken where the sorting is relevant to the density of solutions around each solution. Then the less dense ones are chosen. Finally, crowded tournament selection, crossover, and mutation are performed to create offspring population from the generated new population.

6.4 Experimental Study

We validated *AUQantO* with 16 different case studies, where the case studies are associated with four different deep learning architectures on two medical datasets using both Shannon-entropy and MC dropout uncertainty quantification methods. A 5 x 4 nested cross-validation has been used to evaluate the performance of the methods in all the case studies.

6.4.1 Datasets

In this work, we used two medical image datasets. Fig. 6.2 depicts samples of the used datasets along with different class categories.

Breast Cancer Dataset

BreAst Cancer Histology images (BACH) dataset [7] of hematoxylin & eosin stained breast cancer histology images divided into two parts (A and B). Images of Part B were provided for pixel-wise classification tasks. Consequently, in this work, we used images of part A of the dataset which is composed of 400 microscopy images of size 2048 x 1536 pixels and 20X magnification level. The 400 images are divided into four groups (normal, benign, in situ, and invasive).

Skin Cancer Dataset

Skin cancer dataset [84] is introduced by the International Skin Imaging Collaboration (ISIC). Over 2,000 individuals contributed 33,126 dermoscopic images of benign and malignant skin lesions. For computational and memory efficiency, we utilised a smaller version of the dataset [1] which comprises of 3297 image samples (with 224 x 224 pixels) distributed between the two classes of skin lesions as 1800 images for benign and 1497 for malignant.

6.4.2 Deep Learning Architectures

In this section, we explain in detail the deep learning architectures used in the experimental study. The choice of the deep learning models was based on two different classes of architectures: First, single deep learning architectures, where the input image is transformed into small patches and fed into a feature extraction network (patch-wise network). Then, the extracted feature maps are then fed into an image-wise networks for the final classification. The second class of architectures is the ensemble architecture of deep learning models. In general, ensemble architectures have a number of deep learning models to learn image features using different learning perspectives and hence introduce diversity in the final ensemble of image prediction.

Two-stage CNN

Two-stage CNN [73] is a single deep learning architecture that works by taking an input image then dividing the image into smaller patches of equal size. Then, the image patches are inserted into a custom patch-wise network which acts as a feature extractor. The network made up of a set of 3 x 3 filters in the convolutional layers,

[1] https://www.kaggle.com/fanconic/skin-cancer-malignant-vs-benign

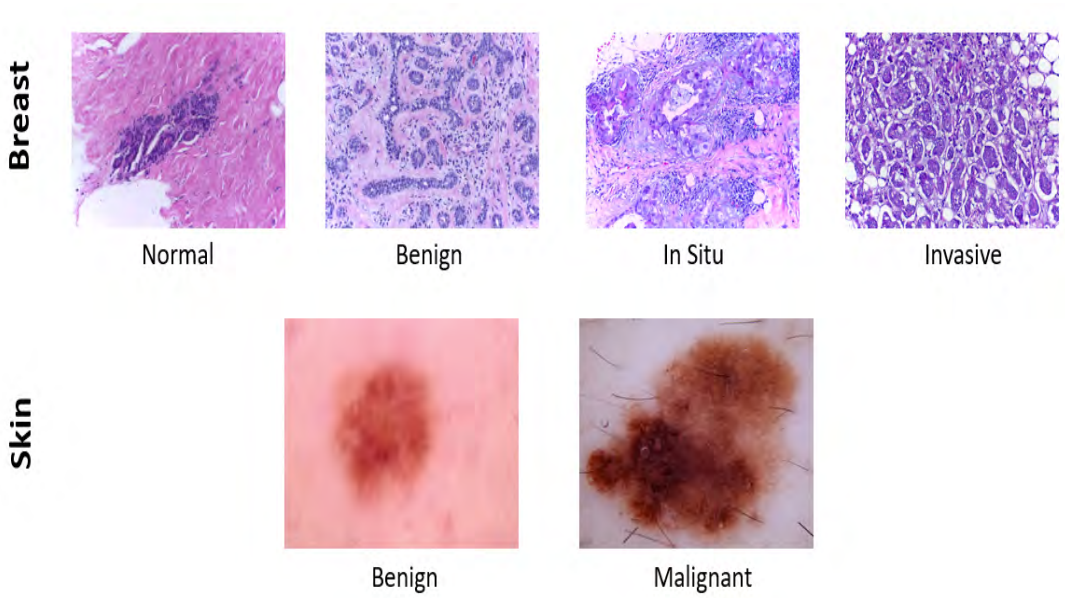


FIGURE 6.2: Overview of the used datasets. Breast dataset has 4 classes while skin cancer dataset has 2 classes. Breast dataset is classified by three tumour classes and one class to represent normal samples. For skin cancer, the samples are classified as benign and malignant cases.

succeeded by a pooling layer. After each down-sampling, the number of channels is doubled. The feature maps are down-sampled using 2 x 2 filters in the convolutional layers with a stride of 2. After each convolutional layer, ReLU activation and batch normalisation were utilised. The feature maps extracted from the patch-wise network are used by an image-wise network that produces the final image classification. The image-wise network follows a similar architecture to the patch-wise network. The network has a series of 3 x 3 convolutional layers accompanied by a pooling layer with a stride of 2 for down-sampling. ReLU activation and batch normalisation have been used after each layer. A 1 x 1 convolutional layer is then used for the spatial average of activation maps followed by 3 fully connected layers with a softmax classifier.

Deep Spatial Fusion CNN (DSF-CNN)

Deep Spatial Fusion CNN (DSF-CNN) introduced by [42] as a single deep learning architecture that can capture spatial dependencies among image patches, where it takes an image as input then divides the image into patches to be inserted into a patch-wise residual network. The extracted feature maps are inserted into a deep spatial fusion network to learn the spatial relationship between image patches. The spatial fusion network consists of a sequence of fully connected layers, where a dropout layer has been used after each fully connected layer.

Hybrid LSTM

Hybrid Long short-term memory (LSTM) [119] follows the same architecture as the two previously described single architectures. Image patches extracted from the input image are inserted to the inception patch-wise network for extracting rich multi-level features then the extracted features are inserted to bi-directional LSTM which

learns spatial information between feature maps extracted from the inception network.

EMS-Net

For ensemble deep learning architectures, we utilised an Ensemble of Multi-Scale Network (EMS-Net) [120]. The architecture takes an input image, then resizes the image into two image scales. The ensemble architecture consists of three pre-trained deep learning models. The first two models use DenseNet-161 on the two images scales, while the third model uses ResNet-152 on one of the two image scales. Then, using the ensemble process, image predictions extracted from the three models are combined to produce the final image classification.

6.4.3 Experimental setup

For evaluating the single architectures (Two-stage CNN, DSF-CNN, and Hybrid LSTM) on skin dataset, we used the resized images of 224 $\times$224 pixels. During the training of a patch-wise network of single architectures, we extracted overlapped image patches of size 112 $\times$ 112 pixels from input images using a patch stride of 56. For the image-wise network of single architectures, we extracted non-overlapped image patches using a patch-stride of 112. We used data augmentation to rotate the training patches 90 degrees while flipping them horizontally and vertically. Adam optimiser [48] has been used to reduce the loss function of the networks. Patch-wise and image-wise networks are trained using a learning rate of 0.0001 and a batch-size of 32.

In the BACH dataset, we used the original image size (2048 $\times$ 1536 pixels) as an input to the single architectures (Two-stage CNN, DSF-CNN, and Hybrid LSTM). We extracted overlapped image patches of size (512 $\times$ 512 pixels) using a patch stride of 256 to train the patch-wise network. Non-overlapped image patches are used for the image-wise network of the single architectures (using patch stride 512).

Lastly, we employed an ensemble architecture (EMS-Net) to the two datasets (BACH and skin). We utilised the exact hyperparameter settings for the BACH dataset as described in [120]. This is by utilising two image scale levels (scale 1: 448 $\times$ 336, scale 2: 296 $\times$ 224) for the three pre-trained DCNN models. We extracted patches of size 224 $\times$ 224 pixels and we fine-tuned the pre-trained DCNN models based on the BACH dataset. We changed the number of neurons in the last fully connected layer of the pre-trained models to 4 (where BACH has 4 classes). Moreover, during the training, we used patch-strides of 28 and 9 for scales 1 and 2, respectively, while during testing, we used patch strides of 56 and 18 for scales 1 and 2, respectively. Finally, we followed the same augmentation settings similar to the single architectures and we used Adam optimiser with learning rate of 0.0001 and batch-size of 32.

We applied a similar strategy to the one used for the EMS-Net on the BACH dataset to skin dataset. We utilised two image scales (scale 1: 224 $\times$ 224, scale 2: 112 $\times$ 112) for the three pre-trained DCNN models. we extracted patches of size 112 $\times$ 112 and 56 $\times$ 56 for scales 1 and 2, respectively. We modified the number of neurons in the last fully connected layer of the pre-trained models to 2 (where skin dataset has 2 classes). We used patch strides of 56 for scale 1 and 28 for scale 2 during the training and testing phases. Finally, the remaining settings in terms of data augmentation and Adam optimiser are the same as EMS-Net on BACH dataset. As can be seen from the settings, we employed to evaluate the *AUQantO* method

using different dataset image settings (e.g. image scales) and different deep learning architectures including pre-trained DCNN models.

We employed the four optimisation methods explained earlier (Bayesian optimisation using GP, COBYLA, dual annealing, and NSGA-II) to all case studies. In the single-objective optimisation methods (Bayesian optimisation using GP, COBYLA, and dual annealing), we set the λ range from 1×10^{-9} to 2 while the evaluation step is set to 50. In COBYLA, the initial search point is set to 0.01. Finally, in the multi-objective optimisation method (NSGA-II), we set the number of variables to 1 as we optimise only one hyperparameter (e.g. λ), number of objectives to 2, number of generations to 50, population size to 1, and we utilised the same λ range as in the single-objective optimisation.

To perform the uncertainty measure using Bayesian approximation with MC dropout to the deep learning architectures, we employed 50 test runs (which has been proved to be adequate to establish a valid distribution) for each image.

6.4.4 Results and Analysis

In this work, we introduce three different metrics to measure the effectiveness and robustness of *AUQantO*. First, we introduce Weighted Average Accuracy (WAA), which measures average classification accuracy weighted by the included images in each test fold. Second, Accuracy Difference (AD) measures the difference between the accuracy of included images and the accuracy of excluded images. Third, The Abstain Percentage (AP) calculates the proportion of excluded images in each dataset compared to the total number of images. The three metrics are formulated as follows:

$$WAA = \frac{1}{\sum_{i=1}^{r} L_i} \sum_{i=1}^{r} Acc_i \times L_i \tag{6.8}$$

$$AD = WAA_{inc} - WAA_{exc} \tag{6.9}$$

$$AP = \left(\frac{\sum_{i=1}^{r} V_i}{D} \right) \times 100 \tag{6.10}$$

where L_i is the number of images (whether they are included or excluded images) in fold i. Acc_i is the classification accuracy in fold i over a total number of r folds. Acc equals to $(TP + TN)/(TP + TN + FP + FN)$, where TP and TN represent the correct predictions by our model, while FP and FN represent the incorrect predictions. WAA_{inc} and WAA_{exc} are the weighted average accuracy for included and excluded images, respectively. V_i is the number of excluded images in fold i, while D is the total number of images in each dataset.

Performance of Deep Learning Architectures

Table 6.1 shows the average test accuracy for all case studies before applying *AUQantO* method to exclude images. After evaluating the deep learning models on each dataset, it can be noticed that EMS-Net has higher test accuracy on BACH (94.00%) and skin (91.30%) datasets among all deep learning architectures. This is because of the usage of an ensemble architecture that applies diversity in learning variety of image features. While, in terms of single deep learning architectures, we can see that DSF-CNN and Hybrid-LSTM have high average test accuracy comparable to EMS-Net. DSF-CNN achieved average test accuracy of 91.25% and 90.14% on BACH and

skin datasets, respectively, while Hybrid-LSTM achieved average test accuracy of 90.25% and 90.02% on BACH and skin datasets, respectively.

TABLE 6.1: Average test accuracy (without image exclusion - *AUQantO* method) for all case studies.

Architecture	Dataset	
	BACH	Skin
Two-stage CNN	88.25%	83.90%
DSF-CNN	91.25%	90.14%
Hybrid-LSTM	90.25%	90.02%
EMS-Net	94.00%	91.30%

Uncertainty measure - Shannon Entropy

Table 6.2 demonstrates the performance of *AUQantO* (in terms of the weighted average test accuracy of included images) using Shannon Entropy with four optimisation methods (Bayesian optimisation using GP, COBYLA, Dual Annealing, and NSGA-II). *AUQantO* shows a significant improvement in all case studies using the four optimisation methods. Also, this improvement showed the capability of *AUQantO* in automatically excluding poor samples that are misclassified or even the uncertain images that are correctly classified. Moreover, NSGA-II demonstrated the highest average test accuracy among other optimisation methods for all case studies except for two case studies (EMS-Net and Hybrid-LSTM on BACH dataset) which have the highest improvement reported by dual annealing (94.78%) and COBYLA (92.65%), respectively. NSGA-II achieved highest test accuracy of 90.29% for Two-stage and 97.61% for DSF-CNN on BACH dataset, while NSGA-II on skin dataset reported 93.46% for Two-stage, 99.06% for DSF-CNN, and around 96% for both Hybrid-LSTM and EMS-Net. In terms of single architectures, DSF-CNN has shown higher performance for all optimisation methods on BACH and skin compared to Two-stage CNN and Hybrid-LSTM.

In Table 6.3, we present the performance of *AUQantO* on the excluded poor samples along with the abstain percentage (which presents the number of excluded images to the total number of images). As shown in Table 6.3, the excluded image accuracy (in all case studies) varies between 20% to 74% which indicates how effective our method is on excluding poor samples. Moreover, the Hybrid-LSTM excluded the least number of images with the lowest abstain percentage for all optimisation methods among all architectures except in one case (e.g. Hybrid-LSTM on skin dataset using NSGA-II) where EMS-Net on skin dataset showed the lowest abstain percentage (17.65%). Also, the evaluation of Hybrid-LSTM on BACH dataset showed very low excluded images accuracy of 25%, 42.11%, 20%, and 25% for GP, COBYLA, dual annealing, and NSGA-II, respectively. This proves that *AUQantO* is effective and successful in minimising the exclusion rate by excluding the most challenging and poor samples that need manual investigation by medical experts.

To further demonstrate the effectiveness of our method in excluding poor image samples, Fig. 6.3 shows (1) Accuracy Improvement which presents the level of improvement (in terms of accuracy) reported after using *AUQantO* by excluding

TABLE 6.2: Average test accuracy of included images using *AUQantO* method (Uncertainty measure: Shannon Entropy) for all case studies.

Architecture (Dataset)	Optimisation Method			
	GP	COBYLA	Dual Annealing	NSGA-II
Two-Stage (BACH)	89.34%	88.99%	89.16%	90.29%
DSF-CNN (BACH)	94.57%	93.35%	93.58%	97.61%
Hybrid-LSTM (BACH)	91.58%	92.65%	91.14%	92.27%
EMS-Net (BACH)	94.75%	94.44%	94.78%	94.68%
Two-Stage (Skin)	91.01%	90.79%	91.74%	93.46%
DSF-CNN (Skin)	95.58%	93.88%	94.12%	99.06%
Hybrid-LSTM (Skin)	93.75%	93.51%	93.36%	96.67%
EMS-Net (Skin)	96.25%	96.64%	95.60%	96.69%

the uncertain samples and (2) Accuracy Difference (AD) which represents the difference between average test accuracy of included and excluded images. Figs. 6.3 (a) and (b) show the accuracy improvement in all deep learning models on BACH and skin datasets using the four optimisation methods, confirming the effectiveness of *AUQantO*. Also, the NSGA-II optimisation method showed the highest level of accuracy improvement for all deep learning architectures on skin dataset (Fig. 6.3 (b)), where an accuracy improvement of approximately 10% has been achieved by Two-stage and DSF-CNN, almost 7.5% by Hybrid-LSTM, and 5% by EMS-Net. Figs. 6.3 (c) and (d) show the accuracy difference between included and excluded images in all deep learning models on BACH and skin datasets using the four optimisation methods. Hybrid-LSTM showed the highest accuracy difference with all optimisation methods on the BACH dataset (Fig. 6.3 (c)) by achieving accuracy difference of around 70% for GP, dual annealing, and NSGA-II and 50% for COBYLA. In the skin dataset, the obtained accuracy differences among all deep learning architectures and optimisation methods look the same with accuracy difference varies between 22% and 38% (Fig. 6.3 (d)).

Uncertainty measure - MC Dropout

Here, we describe the experimental study conducted to the case studies using MC dropout as an uncertainty quantification measure. Table 6.4 presents the average test accuracy of the images included to the final classification. NSGA-II showed the highest average test accuracy among other optimisation methods for the following 3 case studies: Two-stage on BACH dataset (89.72%), Two-stage on skin dataset (89.54%), and DSF-CNN on skin dataset (96.39%). The other 5 case studies have comparable records (varies between approximately 92% and 97%) among single-objective optimisation methods. In terms of single architectures, DSF-CNN showed high accuracy with all optimisation methods (97.92%, 94.97%, 96%, and 96.76% for GP, COBYLA, dual annealing, and NSGA-II, respectively) on BACH dataset. Also, DSF-CNN achieved higher accuracy for all optimisation methods except GP on skin dataset, where accuracy measures of 93.23%, 93.93%, 96.39% have been reported

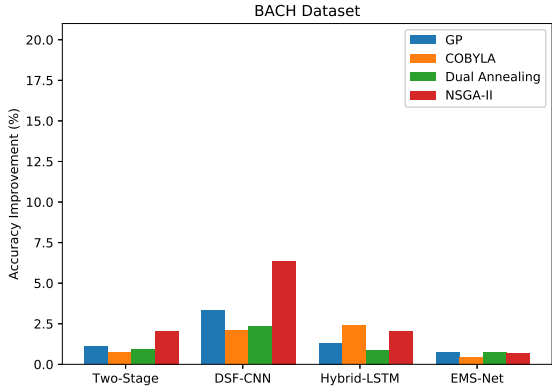
TABLE 6.3: Average test accuracy of excluded images and (Abstain percentage of dataset images) using *AUQantO* method (Uncertainty measure: Shannon Entropy) for all case studies.

Architecture (Dataset)	Optimisation Method			
	GP	COBYLA	Dual Annealing	NSGA-II
Two-Stage (BACH)	71.85% (13.25%)	72.49% (10.50%)	72.46% (11.25%)	74.24% (27.50%)
DSF-CNN (BACH)	68.00% (12.50%)	71.79% (9.75%)	71.43% (10.50%)	73.83% (26.75%)
Hybrid-LSTM (BACH)	25.00% (2.00%)	42.11% (4.75%)	20.00% (1.25%)	25.00% (3.00%)
EMS-Net (BACH)	57.90% (4.75%)	63.64% (5.50%)	58.82% (4.25%)	66.67% (6.00%)
Two-Stage (Skin)	56.10% (20.38%)	56.30% (19.99%)	57.37% (22.84%)	63.46% (31.88%)
DSF-CNN (Skin)	68.90% (20.38%)	60.54% (11.22%)	57.42% (10.83%)	77.16% (40.79%)
Hybrid-LSTM (Skin)	57.27% (10.22%)	57.76% (9.77%)	56.95% (9.16%)	68.26% (23.42%)
EMS-Net (Skin)	65.27% (15.89%)	66.12% (17.65%)	64.77% (13.86%)	66.14% (17.65%)

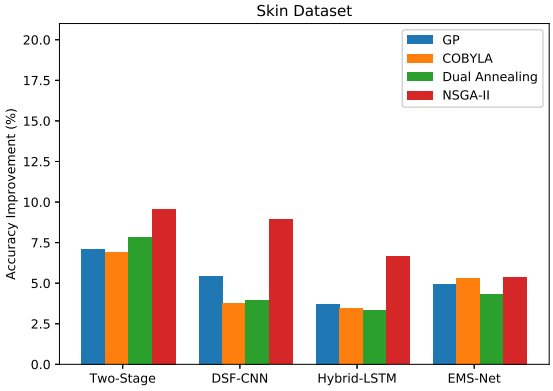
by COBYLA, dual annealing, and NSGA-II, respectively. For GP on skin dataset, Hybrid-LSTM showed an accuracy of 95.44%.

Table 6.5 demonstrates the performance of our method on excluded images and the associated abstain percentage of datasets for all case studies using MC dropout. Hybrid-LSTM showed the lowest excluded image accuracy for all optimisation methods on BACH dataset with excluded images accuracy of 27%, 43.22%, 22%, and 26% for GP, COBYLA, dual annealing, and NSGA-II, respectively. While for skin dataset, Hybrid-LSTM showed the lowest excluded image accuracy for all optimisation methods except GP, where excluded images accuracy measures of 56.22%, 57.40%, and 61.09% have been achieved by COBYLA, dual annealing, and NSGA-II, respectively. For GP on skin dataset, DSF-CNN showed lowest excluded image accuracy of 63.80%. Moreover, generally, the least abstain percentage was obtained by Hybrid-LSTM on BACH and skin datasets.

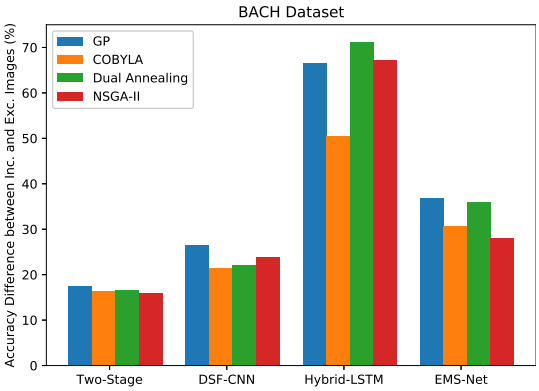
Fig. 6.4 confirms the effectiveness of our method with MC dropout, where the accuracy has been improved in all case studies (Figs. 6.4 (a) and (b)). For BACH dataset, DSF-CNN showed the highest accuracy improvement among the other deep learning architectures by achieving accuracy improvement of approximately 5% for all optimisation methods. The other deep learning architectures (Two-stage, Hybrid-LSTM, and EMS-Net) reported improvement of around 1% for all optimisation methods. The accuracy improvement for skin dataset is approximately 5% for all deep



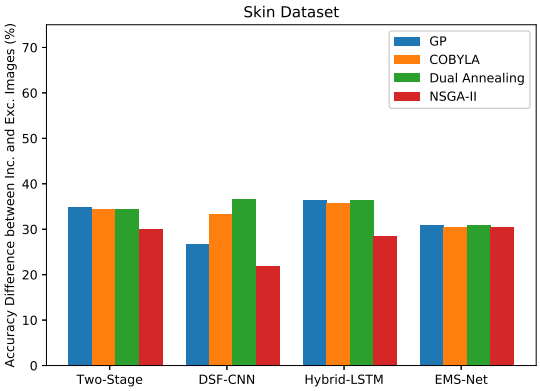
(A) Accuracy Improvement on BACH Dataset



(B) Accuracy Improvement on Skin Dataset



(C) Accuracy Difference on BACH Dataset



(D) Accuracy Difference on Skin Dataset

FIGURE 6.3: Accuracy Improvement using *AUQantO* method and accuracy difference (AD) between included and excluded images for all deep learning architectures on the two medical datasets using Shannon Entropy as uncertainty measure.

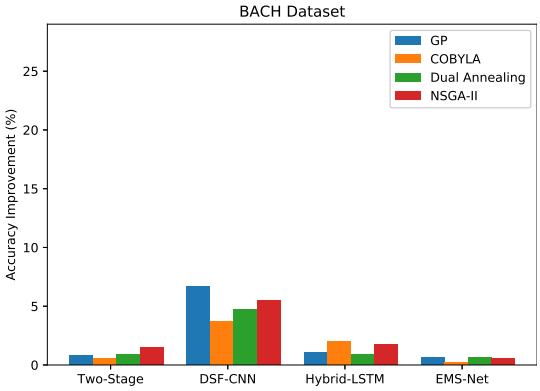
learning architectures and all optimisation methods. Figs. 6.4 (c) and (d) for the accuracy difference (AD) between included and excluded images). Moreover, Hybrid-LSTM, generally, reported highest accuracy difference on BACH dataset with accuracy difference varies between almost 50% and 70% for all optimisation methods. Two-stage and DSF-CNN show accuracy difference of around 20%, while EMS-Net has achieved accuracy difference of around 30% for all optimisation methods. For skin dataset, the accuracy difference has comparable records for all deep learning architectures using all optimisation methods, where an accuracy difference of around 30% has been achieved.

TABLE 6.4: Average test accuracy of included images using *AUQantO* method (Uncertainty measure: MC Dropout - 50 test passes) for all case studies.

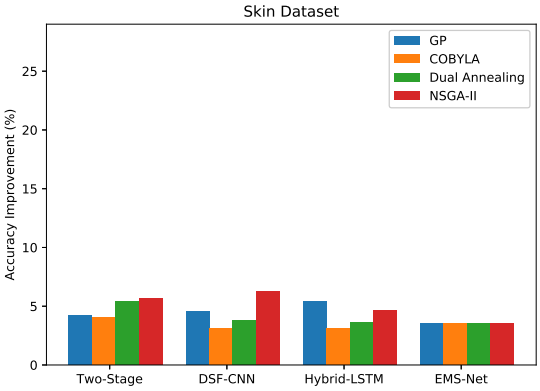
Architecture (Dataset)	Optimisation Method			
	GP	COBYLA	Dual Annealing	NSGA-II
Two-Stage (BACH)	89.10%	88.81%	89.12%	89.72%
DSF-CNN (BACH)	97.92%	94.97%	96.00%	96.76%
Hybrid-LSTM (BACH)	91.32%	92.23%	91.16%	92.03%
EMS-Net (BACH)	94.63%	94.24%	94.68%	94.59%
Two-Stage (Skin)	88.07%	87.92%	89.33%	89.54%
DSF-CNN (Skin)	94.73%	93.23%	93.93%	96.39%
Hybrid-LSTM (Skin)	95.44%	93.14%	93.63%	94.62%
EMS-Net (Skin)	94.82%	94.82%	94.82%	94.82%

TABLE 6.5: Average test accuracy of excluded images and (Abstain percentage of dataset images) using *AUQantO* method (Uncertainty measure: MC Dropout - 50 Test Passes) for all case studies.

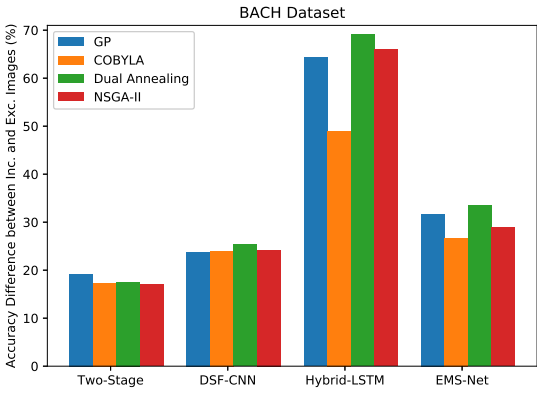
Architecture (Dataset)	Optimisation Method			
	GP	COBYLA	Dual Annealing	NSGA-II
Two-Stage (BACH)	69.86% (11.25%)	71.45% (9.25%)	71.57% (10.50%)	72.67% (22.50%)
DSF-CNN (BACH)	74.10% (28.00%)	70.97% (15.50%)	70.67% (18.75%)	72.52% (22.75%)
Hybrid-LSTM (BACH)	27.00% (2.50%)	43.22% (5.00%)	22.00% (1.75%)	26.00% (3.25%)
EMS-Net (BACH)	62.90% (4.50%)	67.50% (5.00%)	61.24% (4.50%)	65.67% (5.20%)
Two-Stage (Skin)	61.40% (13.83%)	62.36% (13.86%)	61.97% (18.11%)	61.23% (18.23%)
DSF-CNN (Skin)	63.80% (14.83%)	59.46% (9.13%)	60.95% (11.50%)	73.03% (26.66%)
Hybrid-LSTM (Skin)	64.30% (17.50%)	56.22% (8.52%)	57.40% (10.04%)	61.09% (13.80%)
EMS-Net (Skin)	66.14% (11.65%)	66.14% (11.65%)	66.14% (11.65%)	66.14% (11.65%)



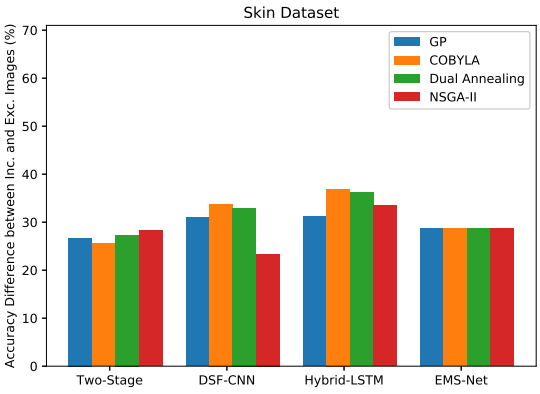
(A) Accuracy Improvement on BACH Dataset



(B) Accuracy Improvement on Skin Dataset



(C) Accuracy Difference on BACH Dataset



(D) Accuracy Difference on Skin Dataset

FIGURE 6.4: Accuracy Improvement using *AUQantO* method and accuracy difference (AD) between included and excluded images for all deep learning architectures on the two medical datasets using MC dropout as uncertainty measure.

6.5 Summary

In this chapter, we introduced a model and a data agnostic method, we called Actionable Uncertainty Quantification Optimisation (*AUQantO*) for optimising uncertainty quantification in deep learning architectures. *AUQantO* can measure uncertainty level in medical images and exclude poor samples based on a hyperparameter (e.g. threshold) that is optimised using single and multi-objective optimisation methods. We validated and evaluated the performance our method on 16 different case studies using two commonly used uncertainty measures. Experimental results showed a favorable performance in the exclusion of highly uncertain images, confirming its automated actionability with different deep learning architectures.

The developed method in this chapter adds a very important characteristic by providing an automated actionability for deep learning models specialised for classification task. We experimentally showed that this method aids in excluding the highly uncertain images and optimises the number of excluded images to be checked by medical professionals for manual investigation. Showing how to benefit clinical practice has been achieved in the three contributions listed in this thesis by providing automated systems that are able to not only show high performance in classification, but also to prove the ability of taking actions for samples that are uncertain as an important step in explainability. The main aim here is to reduce the workload for manual investigation of medical samples and to assist professionals in taking decisions.

Chapter 7

Conclusion and Future Work

7.1 Overview

This thesis studies the impact of applying context-aware learning techniques and uncertainty quantification to improve medical image analysis and introduce benefits for clinical practice. The aim and objectives presented in this thesis focus mainly on the development of novel deep learning models for medical image analysis, specifically for grading and diagnosis applications. The developed methods: (1) utilise context-aware learning techniques to improve performance and (2) offer trustworthiness by applying uncertainty quantification to identify the level of uncertainty introduced by the predictions of deep learning models. In addition, the objectives include the introduction of an actionable method to optimise uncertainty quantification for deep learning models. This method aims to accelerate deep learning models using an actionable strategy to decide on the quality of a particular input image based on the uncertainty measure reported. This measure is supported by a threshold hyperparameter which has been optimised using optimisation methods for single- and multi-objective functions.

Chapter 1 presents an introduction to computer vision, supervised classification tasks, and deep learning along with the main challenges in medical image analysis. Then, motivation has been highlighted to address the research gap available in the area and aim and objectives have been identified for the work conducted in this thesis. Chapter 2 introduces the necessary background and theoretical explanation of the important concepts and methods required to build the contributions stated in the thesis. Chapter 3 reviews the literature work conducted in the field of histopathological image analysis including single and ensemble image classification methods, context-aware methods, and uncertainty quantification methods conducted for medical imaging.

Chapter 4 discusses the first contribution, which addresses the first three objectives in Section 1.4. The chapter presents the *3E-Net* model for classifying grades of invasive breast carcinoma microscopic images. *3E-Net* has the feature of using image-wise models, which learn contextual information in an elastic ensemble fashion and apply Shannon Entropy for uncertainty measure. Chapter 5 discusses the second contribution which addresses also the first three objectives in an advanced and high-level approach. The chapter presents the *MCUa* model for the classification of breast cancer microscopic images which uses multi-scale images, multi-architectures for feature extraction, dynamic ensemble strategy, and uncertainty quantification using MC dropout. Finally, chapter 6 introduces an automated actionable method for optimising uncertainty quantification to address the fourth objective in Section 1.4. The method presents a strategy for finding the optimal number of excluded images from a particular dataset based on uncertainty quantification

measures. In the next section, a reflection on how the objectives of this work have been accomplished is presented.

Overall, we conclude the thesis by introducing two sections. Initially, we discuss how we developed our three contributions presented in the thesis to achieve our research objectives and fulfil our aim. This also includes a discussion of possible prospects and views of the contributions in the thesis. Lastly, possible future directions have been proposed.

7.2 Research Summary

Medical image analysis using deep learning techniques has been a growing field in research for several years. Deep learning techniques have been shown to be effective in improving the performance of automated systems for medical image analysis [65]. Due to the large size of medical images, deep learning techniques encounter problems in processing high-resolution images in one pass. Therefore, dividing an image into small patches is a common approach for dealing with high-resolution images. One of the main challenges is that most patch-based deep learning models focus on local information of image patches without considering the contextual information among different feature maps extracted from these patches. Contextual information is paramount for enriching deep learning models by more context of the spatial dependencies of patch locations and hence improving the vision of the whole image rather than focusing on local regions (patches) in the image.

Another important factor that deep learning models lack is the existence of trustworthiness and a measure of predictions' reliability. This factor is essential for clinical practice, as it reflects how confident a model is in its predictions and whether to count on the predictions of that model on certain input samples. It is crucial to escort automated grading/diagnosis models with uncertainty quantification techniques that introduce a level of explanation on how accurate the prediction generated from DNNs is. The uncertainty measure brings up the notion of introducing actionability to deep learning models. This notion indicates that, on the basis of the uncertainty measures generated from predictions of medical samples, an action can be generated by the automated system. This action is either to produce an automated prediction for a particular sample or to exclude the sample due to high uncertainty. To the best of our knowledge, this approach has not been applied to any deep learning model for grading/diagnosis applications.

Another factor that improves the performance of medical image analysis is the use of an ensemble learning strategy which aims in designing an automated system with many deep learning models. This approach can be utilised to develop multiple models that learn different levels of contextual information (different learning perspectives) that boost performance when we combine predictions generated from all models. As an enhancement for the ensemble learning strategy, a dynamic ensemble strategy based on uncertainty quantification is required. This dynamic ensemble strategy works on selecting only the confident models to contribute to the final image prediction in an automated system that includes many deep learning models. In other words, an image that is inserted into an ensemble architecture with n deep learning models may have a dynamic number of models (less than n) to contribute to the final prediction rather than having all n models contributing to the final prediction.

We achieved our research objectives stated in section 1.4 by introducing three research contributions. We designed, developed, and implemented our first contribution in this thesis which is presented in Chapter 4. We developed "***3E-Net*: Entropy Elastic Ensemble Model for Classifying Grades of Invasive Breast Carcinoma Images**" to address the first three objectives in section 1.4 which are (1) introducing context-aware learning, (2) using uncertainty quantification method, and (3) introducing generalisation and robustness by developing an elastic ensemble learning strategy which combines the only confident models each time we have a new input sample and provide exclusion for uncertain input samples. *3E-Net* is mainly designed and used for grading task where it classifies different grades of invasive breast carcinoma samples. *3E-Net* takes a histopathological image and then divides it into small patches that are inserted into a single DCNN which acts as a feature extractor to extract salient features. These features are then passed to multiple image-wise CNNs, that learn contextual information and generate image-wise predictions (probability distributions) for the input image. Image-wise predictions are then inserted into a new uncertainty quantification stage based on Shannon Entropy to generate uncertainty scores for the input image. These uncertainty scores are then compared against a threshold value to identify image-wise CNNs with low uncertainty scores that contribute to the final image prediction. The predefined threshold and the uncertainty scores generated for a particular image indicate the number of image-wise CNNs used in the ensemble, and here comes the elasticity of our ensemble approach.

Developments motivated us to introduce the second contribution in this thesis are:

- For histopathology images, having multiple image scales enriches an automated diagnosis/grading system with a vision of different nuclei sizes and distributions.
- The availability of multiple feature extractors can lead to a variety of extracted features, which can increase performance.
- The usage of flexible uncertainty quantification method can generate multi-scalar predictions which can introduce high level of confidence and reliability for an automated diagnosis/grading system.

Therefore, we designed, developed, and implemented "***MCUa*: Multi-level Context and Uncertainty aware Model for Classification of Breast Cancer Images**" as our second contribution in Chapter 5 to achieve the first three objectives in section 1.4 which are specifically about (1) introducing different levels of context-aware learning, (2) using multi-scalar uncertainty quantification method, and (3) introduce generalisation and robustness by developing a dynamic ensemble learning strategy which combines the confident models based on MC dropout and provide exclusion mechanism for excluding poor image samples. *MCUa* has been designed and used for diagnosis task that differentiates between breast cancer samples: normal, benign, in situ carcinoma, or invasive carcinoma class categories. *MCUa* applies pre-processing to an input histopathology image by capturing multiple image scales. The multiple image scales are divided into small patches, which are then passed to two different DCNN (DenseNet and ResNet) for feature extraction. Then, the extracted features from the previous stage are inserted into a more developed multi-level context-aware stage which has different levels of context-aware networks. This context-aware stage introduces multi-level feature maps' combinations and builds spatial dependencies information among image patches. Finally,

MC dropout has been used in context-aware networks to introduce an uncertainty-aware component to measure the uncertainty of image predictions. MC dropout works by randomly deactivating different neurons in the neural network during the testing phase. Therefore, a particular image can have many predictions based on the variation of the deactivated neurons during the testing phase. A particular image can have a list of image predictions based on network variations. The mean of the image prediction list represents an aggregated probability distribution for all class categories that indicates the final class label. While the standard deviation of the image predictions list indicates an accurate uncertainty measure. Based on the MC-dropout uncertainty measure, a dynamic ensemble of multi-level context-aware models is introduced which works on selecting the most confident models in the ensemble architecture for generating the final image prediction.

MCUa model has been compared with other deep learning models from the literature in Table 5.4 in Chapter 5. The results prove that *MCUa* has surpassed all the models and showed that the development conducted in terms of multi-scale images, multi-architecture for feature extraction, multi-level context-aware modeling, and uncertainty quantification method using multi-scalar probability distribution enhanced the performance.

3E-Net and *MCUa* provide novel ensemble strategies, which are based on a hyperparameter threshold to identify whether an image is automatically classified or excluded from automated classification to be manually investigated due to high uncertainty. Both architectures succeeded in excluding uncertain images and classifying certain images only. Some observations motivated us to introduce the third contribution in this thesis:

- The threshold hyperparameter used in *3E-Net* and *MCUa* is manually tuned. This may lead to setting the threshold value to low/high value which indicates high/low levels of image exclusion, respectively.
- The actionability of including/excluding images for/from the classification is a manual process based on the threshold.

Therefore, based on the above observations, we designed, developed, and implemented a model agnostic method named "***AUQantO*: Actionable Uncertainty Quantification Optimisation for Medical Image Classification**" presented in Chapter 6. This contribution achieves the fourth objective in section 1.4 which are about developing an automated actionable technique for optimising uncertainty quantification in deep learning architectures. *AUQantO* is a model and dataset agnostic method which can be utilised for any deep learning model that generates probability distribution for input predictions. The method works on probability distributions generated from deep learning models. The method uses Shannon Entropy for single-scalar probability distributions and MC dropout for multi-scalar ones. The generated uncertainty scores from both Shannon Entropy and MC dropout uncertainty quantification techniques are then compared against a hyperparameter threshold. This threshold is optimised using single- and multi-objective functions to achieve the minimised number of excluded images of a particular dataset. We used four optimisation methods and two datasets to prove how effective our method is in excluding the optimal number of highly uncertain images.

Based on all these objectives, we fulfilled our aim which was about developing automated classification systems that have characteristics of context-aware learning,

trustworthiness using uncertainty quantification, generalisation and robustness using a novel dynamic ensemble strategy, and automated actionability by optimising uncertainty quantification.

7.3 Future Directions

The methods described in this thesis pave the way for further advances that should be theoretically investigated and empirically evaluated in the future. To name a few possibilities for future work, consider the following.

- The developed models can be extended to cope with the semantic segmentation problem of whole-slide images and study the effect of multi-level contextual information on the robustness of the segmentation.
- Attention mechanism in transformers can be utilised to provide context-aware information for different regions of a histopathology image, resulting in improved accuracy and effectiveness in diagnoses [38].
- An explainability component using the post-hoc method can be added to the developed models to understand the decision and internal working mechanism of the developed model. Also, for the aid of action taking through visual explanation on the classified image.
- Our solution can be applied to different histopathological tissues, such as prostate and colorectal cancer.
- *3E-Net* and *MCUa* models can be extended by using the Bayesian-based dynamic ensemble method and comparing the performance with current settings.
- *AUQantO* can be extended by including the trial of other optimisation methods.
- The adoption of machine teaching as the next level for developing useful automated systems for clinical practice.

In conclusion, this thesis advances the field of medical image analysis by developing sophisticated deep learning architectures for applications of grading and diagnosis, accompanied by uncertainty quantification and actionability via the dynamic ensemble technique.

Bibliography

[1] Moloud Abdar et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges". In: *Information Fusion* (2021).

[2] Nouman Ahmad, Sohail Asghar, and Saira Andleeb Gillani. "Transfer learning-assisted multi-resolution breast cancer histopathological images classification". In: *The Visual Computer* (May 2021). ISSN: 1432-2315.

[3] Alper Aksac et al. "BreCaHAD: a dataset for breast cancer histopathological annotation and diagnosis". In: *BMC Research Notes* 12 (Dec. 2019).

[4] Md Zahangir Alom et al. "Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network". In: *Journal of Digital Imaging* 32.4 (Aug. 2019), pp. 605–617. ISSN: 1618-727X.

[5] U. B. Angadi, Anil Rai, and G. Uma. "MBFerns: classification and extraction of actionable knowledge using Multi-Branch Ferns-based Naive Bayesian classifier". In: *Soft Computing* 25.13 (2021), pp. 8357–8369. ISSN: 1433-7479.

[6] Teresa Araújo et al. "Classification of breast cancer histology images using Convolutional Neural Networks". In: *PLOS ONE* 12 (June 2017), e0177544.

[7] Guilherme Aresta et al. "BACH: Grand challenge on breast cancer histology images". In: *Medical Image Analysis* 56 (2019), pp. 122–139. ISSN: 1361-8415.

[8] Eirini Arvaniti et al. "Automated Gleason grading of prostate cancer tissue microarrays via deep learning". In: *Scientific Reports* 8 (Aug. 2018).

[9] Ruqayya Awan et al. "Context-Aware Learning Using Transferable Features for Classification of Breast Cancer Histology Images". In: *Image Analysis and Recognition*. Ed. by Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny. Cham: Springer International Publishing, 2018, pp. 788–795. ISBN: 978-3-319-93000-8.

[10] Ruqayya Awan et al. "Glandular Morphometrics for Objective Grading of Colorectal Adenocarcinoma Histology Images". In: *Scientific Reports* 7 (Dec. 2017).

[11] Jocelyn Barker et al. "Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles". In: *Medical Image Analysis* 30 (2016), pp. 60–71. ISSN: 1361-8415.

[12] Sunitha Basodi et al. "Gradient amplification: An efficient way to train deep neural networks". In: *Big Data Mining and Analytics* 3.3 (2020), pp. 196–207.

[13] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information science and statistics. Springer New York, 2013. ISBN: 9788132209065.

[14] J. Brownlee. *Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python*. Machine Learning Mastery, 2019.

[15] Gabriele Campanella et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". In: *Nature Medicine* 25.8 (Aug. 2019), pp. 1301–1309. ISSN: 1546-170X.

[16] Yusuf Celik et al. "Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images". In: *Pattern Recognition Letters* 133 (2020), pp. 232–239. ISSN: 0167-8655.

[17] Haoyuan Chen et al. "IL-MCAM: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach". In: *Computers in Biology and Medicine* 143 (2022), p. 105265. ISSN: 0010-4825.

[18] Pingjun Chen et al. "Automatic whole slide pathology image diagnosis framework via unit stochastic selection and attention fusion". In: *Neurocomputing* 453 (2021), pp. 312–325. ISSN: 0925-2312.

[19] Pingjun Chen et al. "Interactive thyroid whole slide image diagnostic system using deep representation". In: *Computer Methods and Programs in Biomedicine* 195 (2020), p. 105630. ISSN: 0169-2607.

[20] Sai Saketh Chennamsetty, Mohammed Safwan, and Varghese Alex. "Classification of Breast Cancer Histology Image using Ensemble of Pre-trained Neural Networks". In: *Image Analysis and Recognition*. Ed. by Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny. Cham: Springer International Publishing, 2018, pp. 804–811. ISBN: 978-3-319-93000-8.

[21] François Chollet. *Deep Learning with Python*. Manning, Nov. 2017.

[22] François Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 1800–1807.

[23] K. Deb et al. "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE Transactions on Evolutionary Computation* 6.2 (2002), pp. 182–197.

[24] J. Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[25] Li Deng. "A tutorial survey of architectures, algorithms, and applications for deep learning". In: *APSIPA Transactions on Signal and Information Processing* 3 (2014), e2.

[26] Kosmas Dimitropoulos et al. "Grading of invasive breast carcinoma through Grassmannian VLAD encoding". In: *PLOS ONE* 12.9 (Sept. 2017), pp. 1–18.

[27] Babak Ehteshami Bejnordi et al. "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images". In: *Journal of Medical Imaging* 4 (May 2017).

[28] Muhammad Fraz et al. "FABnet: feature attention-based network for simultaneous segmentation of microvessels and nerves in routine histology images of oral cancer". In: *Neural Computing and Applications* 32 (July 2020).

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

[30] Sourodip Ghosh et al. "Colorectal Histology Tumor Detection Using Ensemble Deep Neural Network". In: *Engineering Applications of Artificial Intelligence* 100 (2021), p. 104202. ISSN: 0952-1976.

[31] Rafael C Gonzalez and Richard E Woods. *Digital Image Processing*. Pearson Prentice Hall, 2007.

[32] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[33] Simon Graham et al. "MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images". In: *Medical Image Analysis* 52 (2019), pp. 199–211. ISSN: 1361-8415.

[34] Yao Guo et al. "Breast Cancer Histology Image Classification Based on Deep Neural Networks". In: *Image Analysis and Recognition*. Ed. by Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny. Cham: Springer International Publishing, 2018, pp. 827–836. ISBN: 978-3-319-93000-8.

[35] Vibha Gupta and Arnav Bhavsar. "Sequential Modeling of Deep Features for Breast Cancer Histopathological Image Classification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2018.

[36] Zabit Hameed et al. "Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models". In: *Sensors* 20.16 (2020). ISSN: 1424-8220.

[37] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[38] Zhu He et al. "Deconv-transformer (DecT): A histopathological image classification model for breast cancer based on color deconvolution and transformer architecture". In: *Information Sciences* 608 (2022), pp. 1093–1112. ISSN: 0020-0255.

[39] Irum Hirra et al. "Breast Cancer Classification From Histopathological Images Using Patch-Based Deep Learning Modeling". In: *IEEE Access* 9 (2021), pp. 24273–24287.

[40] Andrew G. Howard et al. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". In: *CoRR* abs/1704.04861 (2017).

[41] G. Huang et al. "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2261–2269.

[42] Yongxiang Huang and Albert Chi-Shing Chung. "Improving High Resolution Histology Image Classification with Deep Spatial Fusion Network". In: *Computational Pathology and Ophthalmic Medical Image Analysis*. Ed. by Danail Stoyanov et al. Cham: Springer International Publishing, 2018, pp. 19–26. ISBN: 978-3-030-00949-6.

[43] Z. Huang et al. "Medical Image Classification Using a Light-Weighted Hybrid Neural Network Based on PCANet and DenseNet". In: *IEEE Access* 8 (2020), pp. 24697–24712.

[44] Asmaa Ibrahim et al. "Artificial intelligence in digital breast pathology: Techniques and applications". In: *The Breast* 49 (2020), pp. 267–273.

[45] Fahdi Kanavati et al. "A deep learning model for the classification of indeterminate lung carcinoma in biopsy whole slide images". In: *Scientific Reports* 11.1 (Apr. 2021), p. 8110. ISSN: 2045-2322.

[46] Sara Hosseinzadeh Kassani et al. "Classification of Histopathological Biopsy Images Using Ensemble of Deep Learning Networks". In: *Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*. CASCON '19. IBM Corp., 2019, 92–99.

[47] Pegah Khosravi et al. "Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images". In: *bioRxiv* (2017).

[48] Diederik Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations* (Dec. 2014).

[49] Matthias Kohl et al. "Assessment of Breast Cancer Histology Using Densely Connected Convolutional Networks". In: *Image Analysis and Recognition*. Ed. by Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny. Cham: Springer International Publishing, 2018, pp. 903–913. ISBN: 978-3-319-93000-8.

[50] Kansei Komaki, Nobuya Sano, and Akira Tangoku. "Problems in histological grading of malignancy and its clinical significance in patients with operable Breast Cancer". In: *Breast Cancer* 13.3 (July 2006), pp. 249–253. ISSN: 1880-4233.

[51] Ismaël Koné and Lahsen Boulmane. "Hierarchical ResNeXt Models for Breast Cancer Histology Image Classification". In: *Image Analysis and Recognition*. Ed. by Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny. Cham: Springer International Publishing, 2018, pp. 796–803. ISBN: 978-3-319-93000-8.

[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Curran Associates Inc., 2012, 1097–1105.

[53] Chen Li et al. "Cervical Histopathology Image Classification Using Multilayer Hidden Conditional Random Fields and Weakly Supervised Learning". In: *IEEE Access* 7 (2019), pp. 90378–90397.

[54] Jiahui Li et al. "Hybrid Supervision Learning for Pathology Whole Slide Image Classification". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by Marleen de Bruijne et al. Cham: Springer International Publishing, 2021, pp. 309–318. ISBN: 978-3-030-87237-3.

[55] Jiayun Li et al. "A multi-resolution model for histopathology image classification and localization with multiple instance learning". In: *Computers in Biology and Medicine* 131 (2021), p. 104253. ISSN: 0010-4825.

[56] Lingqiao Li et al. "Multi-task deep learning for fine-grained classification and grading in breast cancer histopathological images". In: *Multimedia Tools and Applications* 79 (June 2020).

[57] Xia Li et al. "Classification of breast cancer histopathological images using interleaved DenseNet with SENet (IDSNet)". In: *PLOS ONE* 15 (May 2020), e0232127.

[58] Xingyu Li et al. "Discriminative Pattern Mining for Breast Cancer Histopathology Image Classification via Fully Convolutional Autoencoder". In: *IEEE Access* 7 (2019), pp. 36433–36445.

[59] Yixin Li et al. "A hierarchical conditional random field-based attention mechanism approach for gastric histopathology image classification". In: *Applied Intelligence* (Jan. 2022). ISSN: 1573-7497.

[60] Yuexiang Li et al. "Reverse active learning based atrous DenseNet for pathological image classification". In: *BMC Bioinformatics* 20.1 (Aug. 2019), p. 445. ISSN: 1471-2105.

[61] Yuqian Li, Junmin Wu, and Qisong Wu. "Classification of Breast Cancer Histology Images Using Multi-Size and Discriminative Patches Based on Deep Learning". In: *IEEE Access* 7 (2019), pp. 21400–21408.

[62] Gongbo Liang et al. "Improved Trainable Calibration Method for Neural Networks on Medical Imaging Classification". In: *British Machine Vision Conference (BMVC)*. 2020.

[63] Ming Liang and Xiaolin Hu. "Recurrent convolutional neural network for object recognition". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3367–3375.

[64] Qin Liang et al. "Effect of Ki-67 Expression Levels and Histological Grade on Breast Cancer Early Relapse in Patients with Different Immunohistochemical-based Subtypes". In: *Scientific Reports* 10 (May 2020), p. 7648.

[65] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88. ISSN: 1361-8415.

[66] Yun Liu et al. "Detecting Cancer Metastases on Gigapixel Pathology Images". In: *CoRR* abs/1703.02442 (2017). arXiv: 1703.02442.

[67] Marc Macenko et al. "A Method for Normalizing Histology Slides for Quantitative Analysis." In: vol. 9. June 2009, pp. 1107–1110.

[68] Rui Man, Ping Yang, and Bowen Xu. "Classification of Breast Cancer Histopathological Images Using Discriminative Patches Screened by Generative Adversarial Networks". In: *IEEE Access* 8 (2020), pp. 155362–155377.

[69] Bahram Marami et al. "Ensemble Network for Region Identification in Breast Histopathology Slides". In: *Image Analysis and Recognition*. Ed. by Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny. Cham: Springer International Publishing, 2018, pp. 861–868. ISBN: 978-3-319-93000-8.

[70] Mamoun T. Mardini and Zbigniew W. Raś. "Extraction of actionable knowledge to reduce hospital readmissions through patients personalization". In: *Information Sciences* 485 (2019), pp. 1–17. ISSN: 0020-0255.

[71] Sachin Mehta et al. "Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2018.

[72] Aryan Mobiny and Aditi Singh. "Risk-Aware Machine Learning Classifier for Skin Lesion Diagnosis". In: *Journal of Clinical Medicine* 8 (Aug. 2019), p. 1241.

[73] Kamyar Nazeri, Azad Aminpour, and Mehran Ebrahimi. "Two-Stage Convolutional Neural Network for Breast Cancer Histology Image Classification". In: *Image Analysis and Recognition*. Springer International Publishing, 2018, pp. 717–726. ISBN: 978-3-319-93000-8.

[74] Long Nguyen et al. "Biomedical image classification based on a feature concatenation and ensemble of deep CNNs". In: *Journal of Ambient Intelligence and Humanized Computing* (Mar. 2019), pp. 1–13.

[75] Emilio Soria Olivas et al. *Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques - 2 Volumes*. Hershey, PA: Information Science Reference - Imprint of: IGI Publishing, 2009. ISBN: 1605667668.

[76] Emanuela Paladini et al. "Two Ensemble-CNN Approaches for Colorectal Cancer Tissue Type Classification". In: *Journal of Imaging* 7.3 (2021). ISSN: 2313-433X.

[77] Pushpak Pati et al. "HACT-Net: A Hierarchical Cell-to-Tissue Graph Neural Network for Histopathological Image Classification". In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*. Ed. by Carole H. Sudre et al. Cham: Springer International Publishing, 2020, pp. 208–219. ISBN: 978-3-030-60365-6.

[78] M. J. D. Powell. "A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation". In: *Advances in Optimization and Numerical Analysis*. Ed. by Susana Gomez and Jean-Pierre Hennart. Dordrecht: Springer Netherlands, 1994, pp. 51–67. ISBN: 978-94-015-8330-5.

[79] Lukasz Raczkowski et al. "ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning". In: *Scientific Reports* 9 (Oct. 2019).

[80] Hooman Rashidi et al. "Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods". In: *Academic Pathology* 6 (Sept. 2019), p. 237428951987308.

[81] Carl Edward Rasmussen. "Gaussian Processes in Machine Learning". In: *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*. Ed. by Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 63–71. ISBN: 978-3-540-28650-9.

[82] Abtin Riasatian et al. "Fine-Tuning and Training of DenseNet for Histopathology Image Representation Using TCGA Diagnostic Slides". In: *Medical Image Analysis* (2021), p. 102032. ISSN: 1361-8415.

[83] P Robbins et al. "Histological grading of breast carcinomas: A study of interobserver agreement". In: *Human Pathology* 26.8 (1995), pp. 873–879. ISSN: 0046-8177.

[84] Veronica Rotemberg et al. "A patient-centric dataset of images and metadata for identifying melanomas using clinical context". In: *Scientific Data* 8.1 (Jan. 2021), p. 34. ISSN: 2052-4463.

[85] Kaushiki Roy et al. "Patch-based system for Classification of Breast Histology images using deep learning". In: *Computerized Medical Imaging and Graphics* 71 (2019), pp. 90–103. ISSN: 0895-6111.

[86] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3 (Dec. 2015), pp. 211–252. ISSN: 1573-1405.

[87] Stuart J. (Stuart Jonathan) Russell. *Artificial intelligence : a modern approach*. Includes bibliographical references (pages 1063-1093) and index. Third edition. Upper Saddle River, N.J. : Prentice Hall, 2010.

[88] Caglar Senaras et al. "Optimized generation of high-resolution phantom images using cGAN: Application to quantification of Ki67 breast cancer images". In: *PLOS ONE* 13.5 (May 2018), pp. 1–12.

[89] Zakaria Senousy, Mohamed Medhat Gaber, and Mohammed Abdelsamea. "AUQantO: Actionable Uncertainty Quantification Optimization in Deep Learning Architectures for Medical Image Classification". In: *UNDER REVIEW* (2022).

[90] Zakaria Senousy et al. "3E-Net: Entropy-Based Elastic Ensemble of Deep Convolutional Neural Networks for Grading of Invasive Breast Carcinoma Histopathological Microscopic Images". In: *Entropy* 23.5 (2021). ISSN: 1099-4300.

[91] Zakaria Senousy et al. "MCUa: Multi-level Context and Uncertainty aware Dynamic Deep Ensemble for Breast Cancer Histology Image Classification". In: *IEEE Transactions on Biomedical Engineering* (2021), pp. 1–1.

[92] M. Shaban et al. "Context-Aware Convolutional Neural Network for Grading of Colorectal Cancer Histology Images". In: *IEEE Transactions on Medical Imaging* 39.7 (2020), pp. 2395–2405.

[93] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[94] Zhuchen Shao et al. "TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 2136–2147.

[95] Yash Sharma et al. "Cluster-to-Conquer: A Framework for End-to-End Multi-Instance Learning for Whole Slide Image Classification". In: *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*. Ed. by Mattias Heinrich et al. Vol. 143. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 682–698.

[96] Dinggang Shen, Guorong Wu, and Heung-Il Suk. "Deep Learning in Medical Image Analysis". In: *Annual Review of Biomedical Engineering* 19.1 (2017). PMID: 28301734, pp. 221–248.

[97] Yiqing Shen and Jing Ke. "A Deformable CRF Model for Histopathology Whole-Slide Image Classification". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by Anne L. Martel et al. Cham: Springer International Publishing, 2020, pp. 500–508. ISBN: 978-3-030-59722-1.

[98] Xiaoshuang Shi et al. "Pairwise based deep ranking hashing for histopathology image classification and retrieval". In: *Pattern Recognition* 81 (2018), pp. 14–22. ISSN: 0031-3203.

[99] Rebecca L. Siegel et al. "Cancer Statistics, 2021". In: *CA: A Cancer Journal for Clinicians* 71.1 (2021), pp. 7–33. DOI: https://doi.org/10.3322/caac.21654.

[100] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations*. 2015.

[101] Sudhir Sornapudi et al. "Deep Learning Nuclei Detection in Digitized Histology Images by Superpixels". In: 9 (Mar. 2018), p. 5.

[102] Fabio A. Spanhol et al. "A Dataset for Breast Cancer Histopathological Image Classification". In: *IEEE Transactions on Biomedical Engineering* 63.7 (2016), pp. 1455–1462.

[103] Bernard Stewart and Christopher Wild. *World Cancer Report 2014*. The International Agency for Research on Cancer, 2014.

[104] Chunli Sun et al. "Deep Learning-Based Classification of Liver Cancer Histopathology Images Using Only Global Labels". In: *IEEE Journal of Biomedical and Health Informatics* 24.6 (2020), pp. 1643–1651.

[105] Shiliang Sun et al. "A Survey of Optimization Methods From a Machine Learning Perspective". In: *IEEE Transactions on Cybernetics* 50.8 (2020), pp. 3668–3681.

[106] Christian Szegedy et al. "Going deeper with convolutions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9.

[107] Mingxing Tan and Quoc Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 6105–6114.

[108] Yeeleng S. Vang, Zhen Chen, and Xiaohui Xie. "Deep Learning Framework for Multi-class Breast Cancer Histology Image Classification". In: *Image Analysis and Recognition*. Ed. by Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny. Cham: Springer International Publishing, 2018, pp. 914–922. ISBN: 978-3-319-93000-8.

[109] Sulaiman Vesal et al. "Classification of Breast Cancer Histology Images Using Transfer Learning". In: *Image Analysis and Recognition*. Ed. by Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny. Cham: Springer International Publishing, 2018, pp. 812–819. ISBN: 978-3-319-93000-8.

[110] Jelte Peter Vink et al. "Efficient nucleus detector in histopathology images". In: *Journal of microscopy* 249 (Dec. 2012).

[111] Jingwen Wang et al. "Weakly Supervised Prostate Tma Classification Via Graph Convolutional Networks". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020, pp. 239–243.

[112] Pin Wang et al. "Classification of histopathological whole slide images based on multiple weighted semi-supervised domain adaptation". In: *Biomedical Signal Processing and Control* 73 (2022), p. 103400. ISSN: 1746-8094.

[113] Xi Wang et al. "Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis". In: *IEEE Transactions on Cybernetics* 50.9 (2020), pp. 3950–3962.

[114] Tian Xia et al. "Patch-level Tumor Classification in Digital Histopathology Images with Domain Adapted Deep Learning". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018, pp. 644–647.

[115] Tiange Xiang et al. "DSNet: A Dual-Stream Framework for Weakly-Supervised Gigapixel Pathology Image Analysis". In: *IEEE Transactions on Medical Imaging* (2022), pp. 1–1.

[116] Y Xiang et al. "Generalized simulated annealing algorithm and its application to the Thomson model". In: *Physics Letters A* 233.3 (1997), pp. 216–220. ISSN: 0375-9601.

[117] Dan Xue et al. "An Application of Transfer Learning and Ensemble Learning Techniques for Cervical Histopathology Image Classification". In: *IEEE Access* 8 (2020), pp. 104603–104618.

[118] Yuan Xue et al. "Selective synthetic augmentation with HistoGAN for improved histopathology image classification". In: *Medical Image Analysis* 67 (2021), p. 101816. ISSN: 1361-8415.

[119] Rui Yan et al. "Breast cancer histopathological image classification using a hybrid deep neural network". In: *Methods* 173 (2020), pp. 52–60. ISSN: 1046-2023.

[120] Zhanbo Yang et al. "EMS-Net: Ensemble of Multiscale Convolutional Neural Networks for Classification of Breast Cancer Histology Images". In: *Neurocomputing* 366 (July 2019).

[121] Hongdou Yao et al. "Parallel Structure Deep Neural Network Using CNN and RNN with an Attention Mechanism for Breast Cancer Histology Image Classification". In: *Cancers* 11.12 (2019). ISSN: 2072-6694.

[122] Jason Yosinski et al. "How transferable are features in deep neural networks?" In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014.

[123] Changjiang Zhou et al. "Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning". In: *Computerized Medical Imaging and Graphics* 88 (2021), p. 101861. ISSN: 0895-6111.

[124] Qing Zhou et al. "Grading of hepatocellular carcinoma using 3D SE-DenseNet in dynamic enhanced MR images". In: *Computers in Biology and Medicine* 107 (Feb. 2019).

[125] Y. Zhou et al. "CGC-Net: Cell Graph Convolutional Network for Grading of Colorectal Cancer Histology Images". In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2019, pp. 388–398.

[126] Chuang Zhu et al. "Breast cancer histopathology image classification through assembling multiple compact CNNs". In: *BMC Medical Informatics and Decision Making* 19.1 (Oct. 2019), p. 198. ISSN: 1472-6947.