Milestone One

Confirmation Report

Simon Thomas

42000587

25th September 2019

Advisory Team

Nicholas A. Hamilton

James Lefevre

Glenn Baxter

Additional Committee Members

Quan Nguyen

Mikael Boden

Contents

Acknowledgments

I would like to thank my principle advisor Dr Nicholas A. Hamilton for his guidance over the last year. His criticisms have always been enlightening and often lead me to come up with much more creative solutions. I greatly appreciate his contributions and feedback during the development of this confirmation (milestone 1) report. He only has to read it one more time! Also, any further Richard Feynman books that he has will be well received.

Dr James Lefevre has been an incredible sounding board for my ideas over the year. I greatly appreciate the time he freely (seemingly) gives to work through problems with me and provide helpful suggestions along the way. I am looking forward to this next stage where I will require his advise likely more than I already have.

I would also like to Dr Glenn Baxter for his commitment to the project and I am glad to have him on the team. I have greatly enjoyed my training as a pseudo-pathologist. Under his guidance I expect to be able to further understand the problem. His contribution to the data collection, annotation and interpretability components of the project are invaluable.

Lastly, I would like to thank MyLab Pty Ltd and Southern Sun Pathology for providing access to their archives. This project cannot be done without their interest and support.

1 Introduction

In this report I propose a PhD project that will apply advanced machine learning, specifically, deep learning algorithms, to the histological diagnosis of skin cancer. This follows the trend of applying machine learning to problems across many domains of society. In particular, machine learning applied to problems in medicine has been forecast to revolutionise the healthcare sector. Given the enthusiasm by which it is being embraced, the realistic challenges of overcoming the complexity and difficulty of some problems is often overlooked. Indeed, over-promising and under-delivering has a historical precedent in AI research which should raise scepticism. Have I embarked on one such journey?

Anticipating such a problem, I have placed great importance on contextualising modern artificial intelligence (AI), and discuss at length the impact that this has begun to have on medical research. The point of this is to understand where and how meaningful contribution to the field can be made. In doing so I reveal and emphasise the unique challenges that face medicine. Central to this is recognising that medicine is a high-stakes decision domain. Unlike many other areas where machine learning is applied, the medical AI is obligated to conform to the ethical framework that has underpinned medicine for thousands of years. Therefore, these systems need to be developed in a transparent manner, with full acknowledgement of their applicability and limitations. Furthermore, the systems need to be interpretable, that is, we need to be able to explain how and why a prediction was made. I demonstrate that methods exist that can be used to help satisfy this requirement and suggest ways in which it can be improved which serve as a major component of the thesis.

In reviewing the literature, it is revealed that there is little published work on the histological diagnosis of skin cancer. I also discuss how work on automated diagnosis has so far failed to address the full context of the problem, tending to work on problems which limit interpretability or have little clinical relevance. Indeed, a major limitation of previous work is the limited clinical applicability. For machine learning to be adopted in practice, it needs to solve real problems. To this end, effort is made to thoroughly understand the diagnosis of skin cancer. I introduce the biology of the skin, provide an overview of the pathology of the most common skin cancers, and identify key ways in which deep learning can be usefully applied.

The overarching aim of the project is to produce a machine learning system that is both interpretable and clinically relevant. I detail the progress I have already towards this end, demonstrating the skin cancer is indeed amendable to deep learning algorithms. Building on top of the results I layout a research plan for the next two and half years. In it I describe key experimental objectives which will help achieve the aims of the project. Further I discuss opportunities to share and publish my work, as well as what skills and resources are needed to carry the project forward.

By the end of the report I hope to have demonstrated that this project is of suitable scope and achievable given the timeline. Further, I hope that by having been thoughtful of the context of the problem domain, it serves as an example for the sensible, ethical and practical use of machine learning in medicine.

2 Literature Review

2.1 Introduction

2.1.1 An Artificial Intelligence Revolution?

The field of machine learning has experienced an unprecedented surge in popularity within recent years. This is largely a consequence of deep learning, an algorithmic paradigm that in 2012 began producing a succession of state-of-the-art results on image classification and natural language processing tasks [1,2]. Since then deep learning has become synonymous with artificial intelligence, or AI. Technically, AI is the parent discipline which encompasses a multitude of subdisciplines such as pattern-recognition, computer vision, machine learning, robotics, data mining, logic *etc*. The conflation of the terms is prevalent in the mainstream media, giving AI vast connotations depending on the context. However, in common parlance it refers to algorithms which demonstrate an impressive capacity to perform tasks once thought only possible by humans. These algorithms have had an increasing presence in our lives, which until recently has gone mostly unnoticed. However, there is now wide-spread belief that we are experiencing an "AI revolution" that will transform society [3–6].

Exponential Interest

The exponential interest is self-evident from the world-wide usage statistics from Google Trends[7] for the term "deep learning" (Figure 2.1.1). This proxy captures the interest across many (arguably all) sectors of society, including technologists, scientists, futurists, economists, business people, lawyers and lay people alike. It is easy to dismiss much of the hype surrounding AI from the stand-point that the future is inherently unpredictable. However, there is reason to believe that at least some of the requisite technical discoveries are occurring in a similarly unpresented fashion. The preprint server for mathematics, computer science and physics arXiv.org keeps statistics on paper submissions[8]. Between July 1991 and August 2019, the server had a total of 1,575,188 submissions of which the last five years comprise approximately 36% (574,170). In the field of artificial intelligence and subfields of machine learning and computer vision, the number of submissions has grown exponentially (Figure 2.1.2 - Left). Ten years ago, these categories comprised 10% of all computer science submissions, and within five years grew to 20%, and have since more than doubled to over 40% (Figure 2.1.2 - Right).

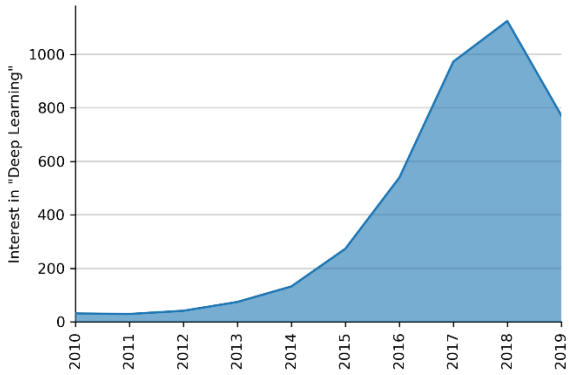

Figure 2.1.1 Google Trends data[7] showcasing the increasing global interest in the topic of deep learning over the last 10 years. The year 2019 includes data up-to and including early August 2019. Interest values are normalised via a Google algorithm and the plot was created using the Python programming language.

These statistics indicate that the research community is both growing and refocusing their efforts on AI-related questions. In fact, the rate of submission is such that it is impossible to keep pace, let alone determine the unique contribution or value of each paper. How many are breakthrough technical papers? How many are minor variations on existing methodologies? How many are applications to new problem domains? In one respect, the scientific community is reliant on the traditional peer-review process to filter and gauge the impact of these works. In another, the sharing of results in large open-access repositories provides the environment for generating, mixing and remixing ideas where the wisdom of the crowd has no-doubt contributed to the rapid progress we are seeing.

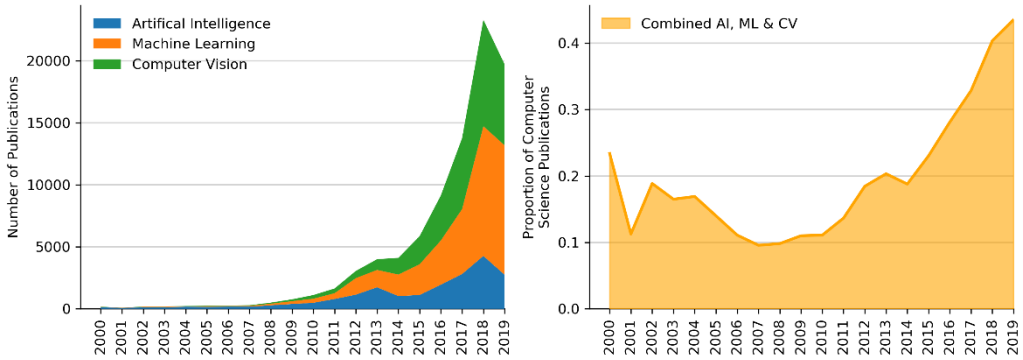

Figure 2.1.2 The increasing number and proportion of artificial intelligence, machine learning and computer vision papers submitted to the arXiv.org preprint server. The data is from arXiv.org and includes data up-to and including early August 2019. The plot was created using the Python programming language.

Inevitable Challenges

In line with increased research outputs, government, industry and non-profits organisations have increasingly directed attention to AI-related projects. The State of AI 2019 report [6] showcases the advances in research, investment, talent procurement, hardware development, industry applications, as well as the political and geopolitical implications of the technology. Importantly, the report highlights numerous challenges that arise from the multi-billion-dollar investment trend. Questions concerning the governance of AI (who has access to the technology), public opinion of automation and job displacement, ethical issues surrounding the transparency of algorithms in high-stakes decisions (*e.g.* predicting recidivism, self-driving cars, health-care *etc.*), militaristic or mass-surveillance use-cases which conflict with current human-rights legislation as well as the multi-level security, institutional and personal vulnerabilities revealed by the recent developments of "deep fakes" [9], all pose serious challenges. Indeed, AI encompasses a disruption that is seemingly at odds with public interests. However, it is a mistake to expect problem-free progress, as is expecting that ethical, social and political problems will disappear as a result of technological abolition. It is an important fact that if all further research into AI is halted, state-of-the-art technology as it already exists can dramatically improve countless aspects of our economic, social, political and personal lives. Consequently, taking these criticisms seriously and addressing them serves to ensure that the benefits we receive outweigh the risks.

Medical AI

One such area in which AI poses an enormous benefit as well as challenge is medicine. The transition to information-based infrastructure has opened up the potential for automation

across the entire health-care sector [10]. However, with the emergence of AI, this automation now extends to include routine tasks once the sole-domain of expert medical practitioners. Specialisations based on visual pattern recognition tasks such as radiology and pathology are the most amenable to automation with current state-of-the-art algorithms. It is this possibility that serves as the general theme of this literature review. The aim is to answer the question, to what extent can current machine learning techniques improve the efficiency and accuracy of medical image analysis? Moreover, what are the limitations of previous works when considering both the technical and practical aspects of the problem? Furthermore, what are the overarching ethical concerns, and how can they be addressed moving forward? There is also a desire to understand the degree to which deep learning has been applied to the area of skin cancer. The above questions will often be answered with this area in mind. Crucially, I want to understand the limitations of previous work so as to devise a unique and significant contribution in applying AI to this field.

2.1.2 Review Structure

To begin, Section 2.2 introduces the promise that AI holds for medicine more broadly. Here I briefly cover the literature surrounding the hype, opinions of medical experts, and the general outlook for AI in medicine. In attempting to keep criticisms at the fore, in Section 2.3 I explore the ethical challenges facing digital medicine. This mostly covers areas which pertain to machine learning, such as use of patient information, data acquisition and ownership as well as dataset and model biases. Before discussing the details of recent applications of deep learning in medicine, the technology itself will be discussed in Section 2.4. To begin, Section 2.4.1 introduces the fundamentals of deep learning. This foundation is then expanded upon when reviewing recent work in the areas of classification (Section 2.4.2), segmentation (Section 2.4.3) and generative methods (Section 2.4.4). These sections will later be treated as assumed knowledge in Sections 3 and 4. Addressing some of the ethical concerns discussed in Section 2.3, the topic of model interpretability is introduced in Section 2.5. It looks specifically at the methods of dimensionality reduction (Section 2.5.1), attribution (Section 2.5.2) and feature visualisation (Section 2.4.3). Having covered a large component of the field of deep learning, Section 2.6 introduces the burgeoning field of digital pathology and showcases the growing number of deep learning papers involving the histological diagnosis of cancer. Additionally, it identifies skin cancer as an area of enormous opportunity to apply deep learning. To support a project in this new area, Section 2.7 will consider the broad problem of skin cancer diagnosis in a histological setting. It will also look at major components of the dermatopathology work flow, identifying clinically relevant tasks which could be assisted with deep learning. Finally, Section 2.8 summaries and concludes the discussions throughout the whole review.

2.2 The Promise of Artificial Intelligence in Health-Care

The application of the scientific method to human health and disease has seen global average life expectancy increase from 30.9 years in 1900 to 71.5 years in 2019 [11,12]. In this time, the value of health has moved from merely "desirable" to a "right", and consequently societies now assign a large portion of their wealth to the healthcare sector. For example, in 2015 the projected spending on health-related research in Australia for the year 2018-2019 was around 1.06 billion dollars, constituting 57% of government provisioned research funding [13]. This commitment to medical research has led to discoveries which have enabled a greater ability to detect, treat and understand diseases. However, emerging from these successes have been problems of increasing complexity and difficulty and many have promised that AI will be a key technology in solving them.

Controversially, the application of AI to medicine is showcased as a panacea of sorts. Among the hype are claims that AI will give us insight into disease states hitherto unknown to us, or constantly monitor our physiologies to pre-emptively alert us of an otherwise undetectable yet imminent heart-attack [14]. Others extoll the promise of combining whole-body scanners, genome sequencing, metabolomics, and X-ray absorptiometry to automatically detect the smallest of ailments, identify disease risks and extend healthy adult life [15]. Further still, the power of AI will purportedly "usher in an era of quicker, cheaper and more-effective drug discovery", and "lead to a full understanding of human biology" [16]. In evaluating these claims, it is important to consider that AI has been coupled with medicine over 30 years ago in a similarly bold manner [17–25]. This previous paradigm of AI was based on expert-systems; programs which mimicked the decision-making process of human experts. Contemporaneous criticism revealed that the "promises of and expectations for expert systems are seriously overblown" [26], and ultimately problems of unresolved theoretical issues, mechanical limitations, development expenses and naive or superficial results lead to their failure. We must therefore ask, what makes the modern promise of AI different?

Modern Artificial Intelligence

To answer this question, it is helpful to focus on two specific components of AI's modern conception; big data and deep learning. Big data describes the coupling of voluminous amounts of data with data analytics to acquire new knowledge or predictive capability [27–29]. This idea underpins IBM's Watson which was designed to search medical literature, patents, genomics, and chemical and pharmacological data to provide physicians with superhuman knowledge [30]. Having access to an enormous knowledge base addresses the scalability problem faced by previous expert-systems. However, the ability to extract knowledge from unstructured datasets is a challenge. So far, deep learning has proven a powerful method for doing this. For example, using medical image datasets, deep learning techniques have been shown to perform dermatologist-level skin cancer classification [31], diabetic retinopathy detection [32], breast cancer subtyping [33], Gleason index prediction for prostate cancer [34], brain tumour segmentation [35] and localisation of colorectal cancer [36]. Work in natural language processing has been shown to successfully extract information from text data in clinical reports [37–39], and in a recent example could out-perform junior doctors in diagnosing common childhood ailments[40]. In fact, recent reviews[41,42]showcase that deep learning has been successfully demonstrated to perform tasks across many domains of medicine and fundamental biology, from disease and patient categorisation, gene targeting of microRNAs, secondary-structure prediction for proteins, analysing electronic health records (EHRs), gene

expression and splicing, transcription factor prediction, protein-protein interactions, batch-effect correction and variation-calling in DNA and RNA sequencing, problems in neuroscience and many more. Therefore, the modern paradigm of AI has a demonstrated ability across many areas of medicine, supporting the notion that current AI is different from previous eras, not just from what it promises, but what it has already delivered.

Balancing Expectations

Although the potential of current and future applications of AI is established, there is a divide between the promises of technologists and the expectations of how clinicians think AI can be usefully applied. This is a critical obstacle in translating the technology to clinical practice. Blease *et al.* [43] conducted a survey of 729 of general practitioners (GP) from the United Kingdom, collecting responses on the possibility of being replaced by AI in six specific tasks. They categorised the responses into three main areas: (1) limitations of future technology (2) potential benefits of future technology, and (3) social and ethical concerns. They found that GPs generally believe that the doctor patient relationship is based on communication and empathy which they consider to be exclusively human faculties. They expressed consensus on the possibility of AI to improve efficiencies and reduce administrative burdens (although this could also be achieved with well-designed information and administrative systems coupled with basic algorithms). However, they found that some consider AI an inevitability and failing to adopt AI in certain scenarios would result in human harm. Interestingly, they concluded that most of the views expressed considered general practice to be a limited use-case for AI in contrast to most technologists.

Others within the medical community hold the belief that embracing AI carries an uncertainty for the future role of physicians generally[44]. The authors claim that compared to physicians, AI systems have superior ability to identify medical risks which could be used to predict prognosis, readmission, and mortality rates. Moreover, they anticipate that big data analytics could be used widely for checking drug-to-drug interactions, optimising booking systems based on monitored risk-factors (blood sugar, heart-rate, haematocrit, oxygen saturation, infection and inflammation biomarkers *etc.*), which could extend into patient triage in emergency rooms. These claims position the authors to argue that the role of physicians will inevitably change, but "rather than replacing physicians, AI will assist them in making better clinical decisions."[14]. Again, this is in line with the wider belief that AI cannot replace doctors at the bedside[14,44–47], because (current) AI cannot engage in high-level conversation or interaction to gain trust, reassurance or empathy, which are "critical parts of the doctor-patient relationship"[44]. The future role of physicians is therefore seen to be in the interpretation of ambiguous conditions, to integrate medical histories, to conduct physical exams, and to facilitate further discussion.

Subtle Obstacles to Adoption

Another physician, Hamet[45], describes the implementation of AI being in two distinct branches: the virtual and the physical. The virtual branch refers to much of what has previously been discussed. Alternatively, the physical branch refers to technologies such as medical devices and robotic systems that aid surgery [48] or assistive care using "care-bots" [49]. A traditional challenge for successful human-robot interaction is moving through the "uncanny valley"[50], the concept which emphasises the contingencies for robots being perceived as acceptable, feared or rejected. This is arguably an equally important question

for virtual systems. What are the contingencies for medical machine learning systems being perceived as acceptable, feared or rejected? Why should we put our lives in the hands of a disembodied "intelligence"? Is there a similarly awkward "uncanny valley" for software, where machines appear to do everything a doctor does, despite the crucial component of conveying that is actually knows what a doctor is, or what "health" is for that matter? There is scepticism and some distrust for systems which have already been deployed in hospitals[51], suggesting that the technology is not quite ready. For doctors to successfully utilise the technology they certainly cannot fear it. Indeed, for patients to accept the use of AI in their health-assessments, they must not fear it either. It is widely suggested that doctors should be trained in the workings of AI systems to better understand the objective risks of the technology, to be able to assess the quality and relevance of the outputs to their patients. At the very least, patients need to have the assurance that their doctor is comfortable using it. This is matter of medical ethics and will be discussed further in the following section (Section 2.3).

The discussion so far has shown that AI has achieved early success in its application to problems in medicine, making it demonstrably different to previous eras. Clinicians generally conclude that the use of AI shows much promise[14,44–47,52,53] but affirm that further research and demonstration through clinical trials are needed to confirm its feasibility and validity; something that has not yet been done. Furthermore, the technology needs to be clinically useful and be directed to solve real-world problems. This necessitates the inclusion of health-care professionals and medical experts in future developments[54]. This inclusion will also help make clearer the role of physicians in the future.

2.3 Ethics of Artificial Intelligence for Medicine

For thousands of years the medical profession has used the Hippocratic Oath to regulate its conduct. Although modern forms of it vary[55,56], the tradition emphasises that the doctor-patient relationship is unchanging in its commitment to ethical practice. This has implications for the way in which AI is utilised in health-care and in this section, we will explore and address some of the ethical challenges that have been raised in the literature that relate to AI research and clinical applications.

Patient Data

Central to the modern conception of medical ethics is patient privacy. To this end, protected health information (PHI) has been defined to describe patient data that are individually identifiable[57]. Data of this type encompasses the obvious *e.g.* name, date of birth *etc.* to the less obvious *e.g.* image of a unique tattoo. In contrast, data that is deidentified, which means there is no reasonable basis to believe it can be used for identification, is not considered PHI. This is a significant challenge for genomic data[58], however most medical imaging is easily deidentified by simply removing patient metadata. In most jurisdictions, non-PHI data is no-longer considered private and can be used publicly, usually for research purposes. In the case of medical images, the availability of large deidentified datasets have enabled deep learning techniques to match expert humans at cancer classification tasks[31]. As effort is made to create new datasets, or increase the size of existing ones, an area of uncertainty is not-whether the data is deidentified, but rather where the data came from and its original intended use. This is frequently described and arguable mis-characterised as the problem of informed consent: do patients know what their data will be used for?

In a medical setting, informed consent has traditionally referred to consent from a patient to treatment options that have inherent risks[59]. However, there are criticisms around patients having the ability to absorb highly technical concepts or specialised information about their diagnoses to warrant being informed. It is also unclear whether medical photography is even analogous to the concept of informed consent because it is neither a treatment nor a medical intervention. What are the inherent risks to taking a photograph or medical image that is deidentified? Pathologists routinely share photographs of interesting or difficult cases with their colleagues, often on social media platforms[57]. The purposes can range from seeking advice from or educating their colleagues. Has the patient consented to this use? Do they have to? There appears to be a common conflation between informed consent and data ownership. In the case of pathology, a patient consents to treatment options, namely biopsy or excision of tissue. The ownership of the tissue is seen has having been abandoned to the pathology laboratory[60]. Ownership rights to subsequent photographs of products derived from that tissue *i.e.* histology images, are granted to the taker of the images under standard copyright laws[57], which in most cases is the pathologist. They can then distribute the images (if they are de-identified). This interpretation extends to the creation of datasets for machine learning research, and so pre-existing datasets such as archived histology slides can be utilised for these purposes. Despite the absence of clear legislation, university research ethics committees routinely characterise research of this nature as “low or negligible risk”.

Equitable Returns

Given the immense value of data, there are also questions surrounding whether people, and in this case, patients, should be more equitably compensated for their consent. Although this is a

difficult to answer, in previous biomedical lawsuits[61] it has been ruled that, if patients can control how their tissues are used, how products from their tissues are used, or how subsequent profits are shared, then "biomedical research cannot be performed successfully or efficiently"[60]. This pragmatic conclusion makes the implicit assumption that the research or end-product will come to benefit society in the long-run. This is certainly true in general; however, others have argued that to ensure that the public benefits from medical AI, research should be done in an open-source manner[45]. It is argued that if medical AI can provide unprecedented insight into human disease and treatment, then there is an ethical obligation to make it accessible. As discussed in Section 2.1.1, there is a trend in general for AI to be developed in an open-source manner, such as the major deep learning frameworks Tensorflow[62] and Pytorch[63] and the sharing of results via arXiv.org. However, it is not clear that developing non-open-source AI would result in the technology being inaccessible to the public, especially in terms of receiving benefits. Nonetheless, the sharing of data to amass the large datasets that deep learning requires seems to be in the public interest. Indeed, others argue that the web of privacy and legal obstacles is impinging on progress in medical AI[41] and this will ultimately affect patients.

Generalisability

With the intention of developing medical AI that serves the public interest, other questions arise surrounding performance guarantees. Recent controversies surrounding "digital discrimination" have identified failures of technology to incorporate the full diversity of end-users in its design *e.g.* the automatic "Racist Soap Dispenser" whose sensor fails to detect dark-skinned people[64]. In the case of medical AI systems, failure to capture the full diversity within the population, or be aware of which population the system was developed for could harm the patient. Consequently, when papers say, "dermatologist level classification"[31], the performance metric cannot be assessed independently of the data on which the model was trained and tested. Therefore, it must be clear what assumptions have been made about the dataset and how it applies to the specific medical problem. For instance, it is well known that some indigenous populations are more susceptible to heart-disease and diabetes, and skin cancer primarily affects Caucasians but not exclusively; will systems be robust to these differences? Similarly, natural language processing has so far been demonstrated for electronic health records in English and Mandarin, two widely used languages. Is there enough data available for other languages to achieve equal performance? Or what about nuanced semantic differences in word use in the same language, such as between different English-speaking countries? Will a system trained on data in the United States be appropriate to use in hospitals in South Africa or New Zealand? These are serious challenges which need to be considered before translating AI research to clinical practice. We may come to learn that generalised AI systems are not practical to train or use, and consequently narrow AI systems *i.e.* task specific, should be used in full knowledge of their biases and limitations. As widely suggested, a good starting point is to develop medical AI applications in the company of domain experts who can determine how well AI systems generalise to real-world scenarios.

A Right to Explanation

The final criticism that raises several challenges to AI in clinical practice is their "black-box" nature. Anticipating the future use of algorithmic decision making, in 2016, the European

Union introduced regulations which mandate a "right to explanation"[65]. To satisfy this, we must be able to explain why or how a system came to give a prediction or diagnosis. Furthermore, if deep learning algorithms will provide new perspectives on disease or treatments, we need to also know what that perspective is, not just that the algorithm is using it[41]. Importantly, the explanation must be meaningful to both the clinician and the patient. This is critical for the successful translation of deep learning research to clinical practice because on top of performance, they must also convince the medical community of their safety. The problem of interpretability is an area of ongoing research, but as will be explored further in Sections 2.5, it is not insurmountable.

2.4 An Overview of Deep Learning Methods

Having established the impact that deep learning can have on medicine, this section will look at the technology itself. In Section 2.4.1, the fundamental concepts will be introduced in the context of supervised learning and then unsupervised learning. Neural network architectures specific for image analysis will then be introduced and how these methods learn to detect abstract patterns is discussed. Image classification (Section 2.4.2) will then be explored in more detail, reviewing and comparing the state-of-the-art architectures. This will also be done for image segmentation (Section 2.4.3). Finally, Section 2.4.4 provides a brief introduction to generative methods and as well as a short discussion on why they are a valuable area of research.

2.4.1 The Basics of Deep Learning

2.4.1.1 Supervised Deep Learning

Deep Learning is characterised by the use of "deep" artificial neural networks in contrast to their "shallow" counter-parts which were originally conceived of in 1957 by Frank Rosenblatt[66]. Named the "perceptron", the original artificial neural network attempted to mimic the basic principles in our understanding of biological neurons; multiple inputs (axons) are connected to a neuron (soma), and the sum of incoming inhibitory and excitatory stimuli are sufficient to reach an activation threshold, causing the neuron to generate an output (action potential). This concept can be represented graphically with neurons as nodes and the existence and strengths of connections as edges (Figure 2.6.1). The computation can be represented using linear algebra where the inputs are a vector X of size n and their relative contributions (weights) to the sum are a vector W of size n. The dot product $W \cdot X$ produces a scalar value z which is then fed to a non-linear activation function, such as the sigmoid function. In the simplest case, a network can give a continuous scalar output $\hat{y}$ between 0 or 1 and a threshold (usually 0.5) can be used to assign a class, either 0 or 1. In this way, a network approximates a function that maps the input X discretely to the output $y \in \{0, 1\}$, where $\hat{y} \approx y$ *e.g.* $0.978 \approx 1$. In addition to the weights, a bias term b can also be added.

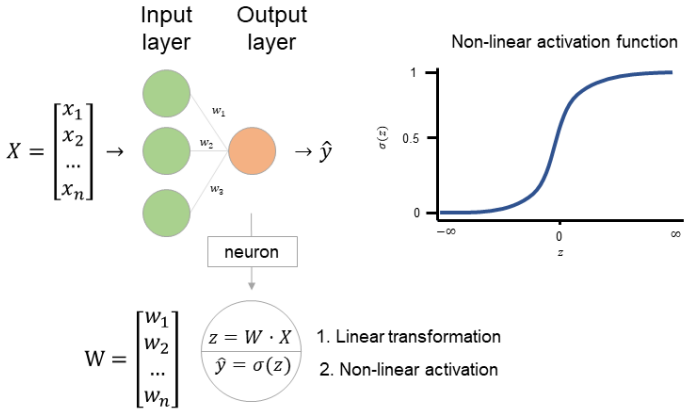

Figure 2.4.1 A single layer neural network. The vector X holds values in the input layer which are then linearly transformed by the weight vector W. This operation computes z which is then given to a non-linear activation function $\sigma(z)$ outputting $\hat{y}$.

Initially, the optimal weights of W are not known and so are set randomly. From this sub-optimal starting point, an output is computed and the error in the prediction is calculated according to a loss function, $L(Y, \hat{Y})$. In most cases, the loss is calculated across multiple predictions, and so $\hat{Y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n]$, is a vector. Importantly, the loss function must be

differentiable, such as the mean squared-error (L_2) loss (Eq. 2.4.1), or binary cross-entropy loss (Eq. 2.4.2). This enables a continuous derivative of the loss with respect to each of the weights to be calculated. The weights can then be adjusted incrementally using the derivative according to a specified step size (or learning rate), improving the prediction. Iteratively optimising the weights via this process is called gradient descent and is the algorithm that underlies the learning component of deep learning. Specifically, this method of learning is called supervised learning, because a mapping from X to y is already known and needed to compute the loss.

$$L(Y, \hat{Y}) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

Equation 2.4.1: Mean Squared-Error Loss

$$L(Y, \hat{Y}) = -\frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \log(\hat{Y}_i) + (1 - Y_i) \cdot \log(1 - \hat{Y}_i)$$

Equation 2.4.2: Binary Cross-Entropy Loss

By design, we can arbitrarily increase the number of neurons in a layer, creating multiple neurons in a multi-layered network (Figure 2.4.2). In this way, the output of one layer serves as the input to the next. The number of weights for each layer is therefore equal to the product of the number of neurons in the two connected layers. For example, for the single-output 5-layered network in Figure 2.4.2, the number of weights in the third layer is $5 \times 3 = 15$, because of the 5 input neurons and 3 output neurons. If we include bias terms for all operations, the network has a total of 44 learnable parameters. In a multilayered network the backpropagation algorithm calculates the derivatives of the loss with respect to all the parameters, enabling the whole network to be trained via gradient descent. One of the challenges of using this algorithm on larger networks is that the magnitude of the gradient diminishes as training progresses. This has come to be called the vanishing gradient problem. Since 2012, alternative activation functions such as rectified linear-units (relu)[67], batch-normalisation[68], weight initialisation techniques[69] and modifications to gradient descent such as the improved Adam[70] algorithm have helped make training faster and more stable. The training of large, multi-layer networks resulting in millions of learnable weights is therefore what is meant by the term deep learning.

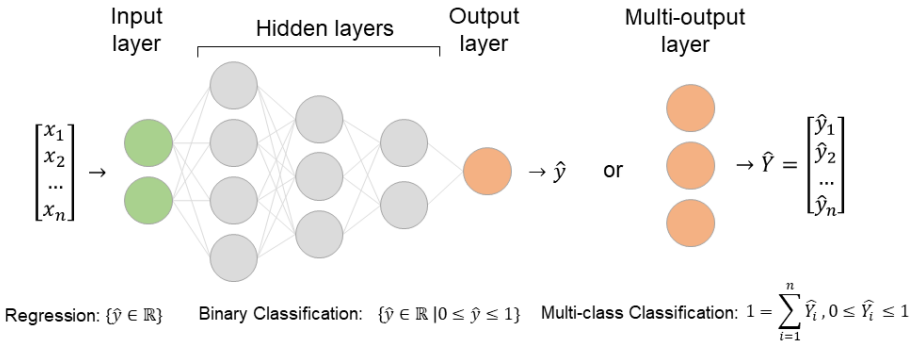

Figure 2.4.2 A multi-layered artificial neural network. The learnable parameters of the network are contained within the hidden layers. The hidden layers consecutively transform the input data into a representation that makes them linearly separable for classification, or weighted features which can perform regression (continuous value prediction). As shown, a network can also be designed to output multiple values.

The power of deep neural networks comes from the way they process information via consecutive layers of non-linear transformations, creating new representations of the original input data. In fact, each representation is defined in relationship to the previous layer, resulting in a nested-hierarchy of features with increasing complexity and abstraction. Having been optimised, these features are thought to be useful when making predictions in classification and regression tasks. However, the exact meaning of these features remains elusive to us and so layers between the input and output of the network are called *hidden layers*. This component is central to the "black-box" nature of neural networks. Attempts to understand these learned representations are the basis of Section 2.5.

2.4.1.2 Unsupervised Deep Learning

Unsupervised machine learning methods attempt to answer the question, is there meaningful structure in the data? Data such as gene expression or patient metadata are typically high-dimensional, therefore patterns may not be immediately, if at all, obvious. In this case, it is common to analyse the raw data using principal component analysis (PCA) or clustering techniques *e.g.* k-means, hierarchical *etc*. In applying these techniques to raw data, important relationships between variables in answering a particular question *e.g.* expression levels of genes x_1 and x_2 on predicting cancer risks, may not be as clear (Figure 2.4.3 - Left). However, if the space is non-linearly transformed, the new representation may improve our chances of seeing the important pattern (Figure 2.4.2 - Right). As mentioned above, deep learning enables high-level representations to be learned as a consequence of non-linear transformations of the data, and consequently using these representations with traditional clustering methods has been shown to be useful[71]. In contrast to supervised learning, these representations can be learned without the need for labels *i.e. a priori* knowledge of the input-output mappings, and instead can be treated as an information theory problem.

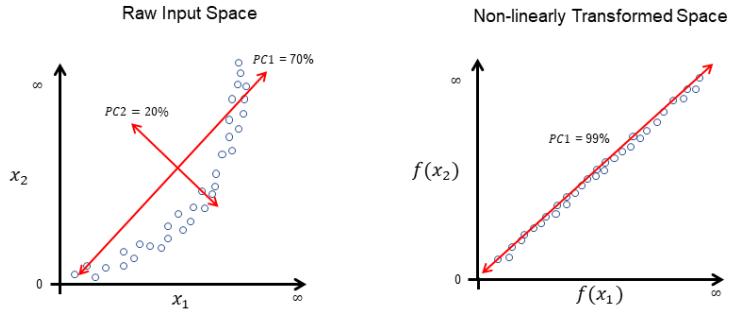

Figure 2.4.3 Looking for relationships using PCA on the raw data may not reveal the true significance of patterns in the data (Left). Transforming the raw data using non-linear unsupervised learning may instead result in a representation where the pattern is clear. Note: the variation captured by each principle component is estimated for demonstration purposes.

In information theory, we typically want to encode (or compress) a message for transmission and then decode (decompress) the message faithfully upon receiving it. The problem is often formulated as, what is the minimum number of bits necessary to transmit and reconstruct the message? In the context of unsupervised learning, we are similarly asking, how much information can be discarded so that the meaning of the original data can be reproduced faithfully? In this case, we are using a deep neural network to learn what information is important and what is not through the process of encoding the input, and then decoding it. The key concept is that we want to learn a mapping from X to X that involves encoding

important information in a code, z. The architecture to do this is called an autoencoder (Figure 2.4.4).

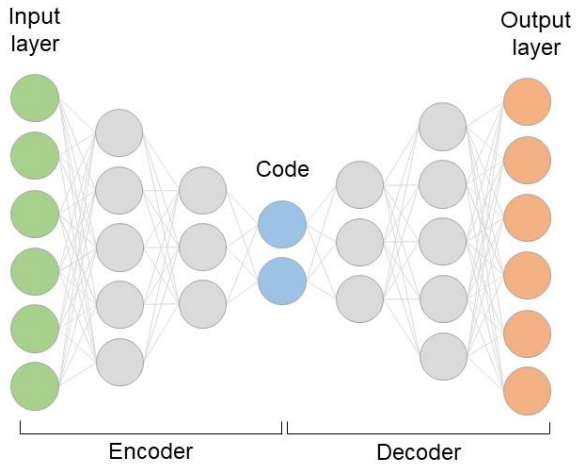

Figure 2.4.4 Autoencoder architecture typically used for unsupervised learning tasks in deep learning. The input layer is transformed to a lower-dimensional representation and then it is transformed back to the original. The loss is a reconstruction loss usually measured using L_2. The size of the code vector, z, is arbitrarily chosen and the overall architecture is designed empirically.

The training process can be performed end-to-end with a single loss function, or reconstruction error, usually L_2 (Eq. 2.4.1). The dimensionality of z is chosen arbitrarily and is problem dependent. For example, $n \in \{2, 3\}$ allows z to be immediately visualised, however $n \gg 3$ is typically used in combination with PCA or other dimensionality reduction techniques (Section 2.5.1). Importantly, the code is assumed to contain salient information, yet the actual meaning remains hidden to us and may include spurious correlations. Thus, generalisability is inferred by training the network on a large dataset and testing on an unseen validation set. If the reconstruction loss is sufficiently low on unseen data, then it is inferred that the code contains salient information. However, since the code contains information to generate outputs by itself, it is also possible to explore this space by generating new data using just the decoder. Interpolating values been two known input codes may reveal the extent to which important information has been captured alongside any spurious correlations. Indeed, more advanced unsupervised learning methods such as Variational Autoencoders and Generative Adversarial Networks (Section 2.4.4) build upon the basic idea of learning a code, or latent space, which summaries the content of images realistically.

2.4.1.3 Image Analysis with Deep Learning

In the context of images, it is usually quite clear to us that there is structure in the data. Indeed, there are obvious spatial correlations between neighbouring pixels, such as edges, horizontal stripes, or the symmetry in a face. Moreover, there are colour correlations, such as the sky or grass. The question then is not whether such patterns exists, but how do we get a computer to detect those patterns? To this end, the network architecture is modified to incorporate the expected structure in the data. Spatial correlations are captured using filters (or kernels) which can be thought of as a block of neurons. The filter, a $n \times n$ weight matrix, is moved across the whole image. At each step the weighted sum of the $n \times n$ pixel-region and the overlayed $n \times n$ filter is computed, generating a convolved matrix (Figure 2.4.5 – a) to which a non-linear activation function can then be applied (Figure 2.4.5 – b). In the example below, the weights are known *a priori* to create a vertical edge detector. In the context of deep learning, the weights are set randomly, and useful feature detectors are

learned via gradient descent according to some training objective *e.g.* minimising a loss function. A pooling layer can also be used to summarise the activations within a given area e.g. 2×2. The summary captures either the maximum or mean values of the activations in the underlaying pixels (Figure 2.4.5 – c) and there are no learnable parameters for this step. Figure 2.4.4 demonstrates these operations in two dimensions, however, three dimensions is common for colour images. This means that the filter detects correlations depth-wise as well.

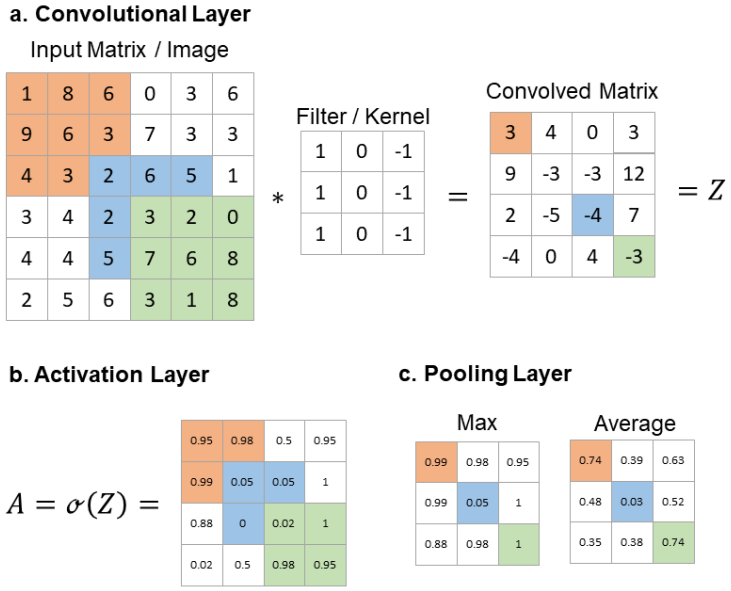

Figure 2.4.5 Examples of the basic operations in a Convolutional Neural Network. a) Convolutional layers are optimised to detect patterns. b) Activation functions allow for non-linear transformations. c) Pooling layers are used to summarise regions of an image.

These operations can be chained together in multiple layers, creating what is referred to as a convolutional neural network (CNN) (Figure 2.4.6). There is enormous variation in the architecture of CNNs, however, they usually consist of repeating "blocks" of convolution and pooling layers. As a result of these layers, the output decreases in height and width, however, the number of filters (or features) progressively increases. The consequence of this is that the original image is summarised as a code or feature vector which can be used for classification, regression or object-detection tasks. Alternatively, convolutions can be performed in the opposite direction, combining up-sampling layers (bilinear interpolation) with convolutions in chains to perform segmentation or reconstruction tasks. This ability means CNNs can also be designed for unsupervised learning for images *i.e.* as autoencoders.

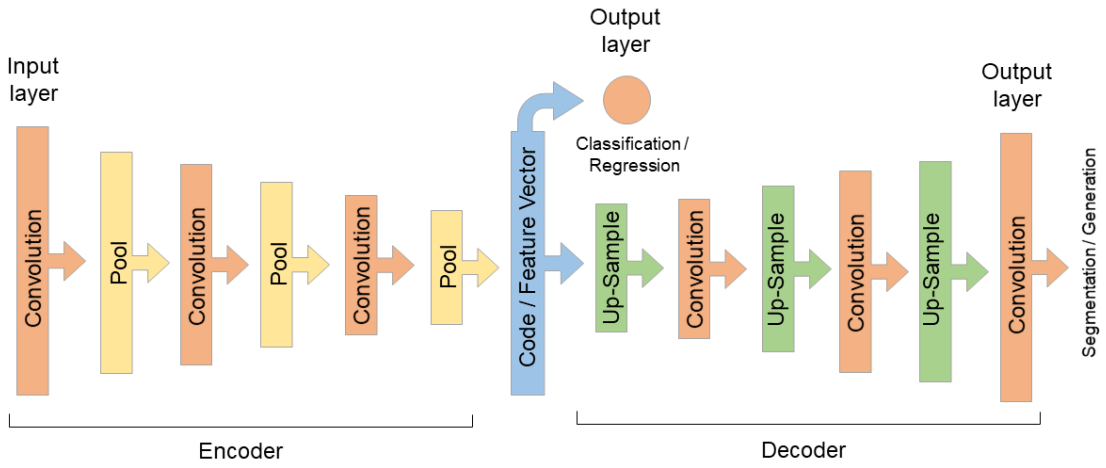

Figure 2.4.6 General architectural design for convolutional neural networks. Consecutive blocks of convolution and pooling layers detect and summarise features which can be used for classification and regression tasks. The number of filers tends to increase with depth as there are many more high-level features than there are atomic features e.g. horizontal edges or dots. A decoder network can also be added to reconstruct the input *i.e.* autoencoder, or perform segmentation and generation tasks.

Similar to fully-connected networks, each layer (in the encoding phase) is a new representation of the original input data and forms a nested-hierarchy of features with increasing complexity and abstraction. In practice this means that early layers detect edges and colours, middle layers detect textures and later layers detect objects or concepts such as "insect-ness" (Figure 2.4.7). The feature vector used for classification therefore contains high-level concepts which can be used for classification, regression or segmentation tasks.

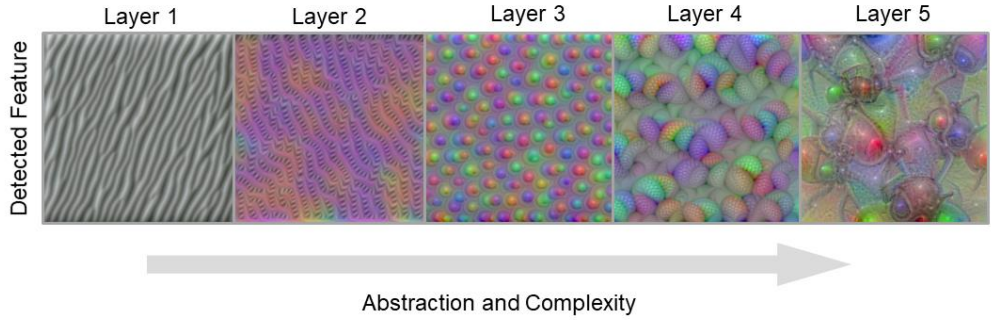

Figure 2.4.7 The nest-hierarchy of features enables convolutional neural networks to detect high-level features in images (and also audio). Early layers detect atomic features such as edges and simple patterns. Middle layers combine these features to make more complex features from which the later layers can detect highly abstract and complex features. The above examples are filter visualisations (Section 2.5.3) derived from the pretrained network VGG16 on ImageNet.

2.4.2 Image Classification

Image classification is the most fundamental machine learning task in computer vision. The aim is to classify an image as belonging to one of C predefined classes and so is in the domain of supervised learning. In a multi-class problem, the final layer of a network contains a vector s of size C and is treated as an estimate of the log-likelihood for a particular class (logits). The SoftMax activation function, $f(s)$ transforms the logits, $\mathbb{R} \in (-\infty, \infty)$, to estimated likelihoods for each class, $\mathbb{R} \in [0,1]$ (Eq. 2.4.3), which are often interpreted as probabilities. These probabilities are then used to compute the loss, typically categorical cross-entropy loss (Eq. 2.4.4). To do this, the ground-truth is one-hot encoded as a vector to match the network output *e.g.* $\hat{y} = [0.25 \quad 0.65 \quad 0.1]$ and $y = [0 \quad 1 \quad 0]$ and thus represent

probability distributions. Training proceeds to minimise the difference in the two distributions (the Kullback-Leibler Divergence or KL Divergence).

$$\hat{y}_i = f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}}$$

Equation 2.4.3: SoftMax function for a single output vector, $\hat{y}$

$$L(y, \hat{y}) = -\sum_{i=1}^{C} y_i \cdot \log(\hat{y}_i)$$

Equation 2.4.4: Categorical Cross-Entropy Loss for a single output vector, $\hat{y}$. In practice the average loss is calculated across several (n) predictions, e.g. Y and $\hat{Y}$.

Transfer Learning

Deep learning networks typically have millions of parameters. This means that one possible solution to many problems is to just learn the answers *i.e.* to remember the correct output for a given input. In this extreme case, when testing the network on unseen data, it fails completely. To avoid this, it is necessary to use large datasets that have many more bits of information than there are parameters, encouraging a network to recognise the true pattern. However, in practice enough data may not be available, and so a network often *overfits* on the training data and performs poorly on the test data. The problem of overfitting can be addressed with regularisation techniques. However, a major way to avoid overfitting in the absence of large amounts of data is through the practice of transfer learning.

Transfer learning involves using a pretrained network that performs well on the ImageNet dataset, consisting of approximately 1.2 million images which belong to 1,000 classes in everyday life e.g. cats, dogs, boats, birds, buildings, food *etc* [72,73]. The advantage of transfer learning is that information that is useful in this problem domain *e.g.* distinguishing cats and boats, is equally useful in others *e.g.* distinguishing different cancer types. An intuition for why this is so can be seen from Figure 2.4.7 in the previous section. All images contain atomic features such as those found in early layers. It is really only the final layers where abstract concepts are learned. Therefore, transfer learning usually consists of taking a pretrained network, attaching a new classifier layer *e.g.* a four-class cancer classifier, and then training the network on new data but only updating weights on the last few layers. This approach has now become standard practice.

Classification Benchmarks

Success with transfer learning relies on the performance of previously trained networks on the ImageNet dataset. As a benchmark, human level performance on ImageNet, measured as top-5 accuracy (the true class is in the top 5 predictions) is estimated to be 94.9%[73]. The ImageNet challenge has for years aim to surpass this bench mark, and in recent years have surpassed it. To understand how deep learning compares to humans, the most common models used in transfer learning and their respective performance accuracy is shown in Table 2.4.1.

The majority of work that utilises transfer learning cites the use of the VGG16 or 19 networks, the ResNet50 or 101 networks or InceptionV3. Access to these models is provided

through the Keras[74], PyTorch[63] and Tensorflow[62] programming libraries. The reason for their popularity may be that they were all developed between 2014 and 2015 have since established a foothold in the research community. Despite their wide spread use, these five networks sit slightly below the human performance (Table 2.4.1). Moreover, MobileNetV2 outperforms VGG16 and 19 with 38x less parameters which should encourage more experimentation. However, the table reveals a trend that improved performance correlates with increased parameters.

Despite the widespread use of models that arguable match human performance, models that exceed it have since been produced and are available. However, the trend of increased performance is met with a possibly prohibited cost of parameters. As of August 2019, state-of-the-art models have *hundreds of millions* of trainable parameters (Table 2.4.2). Models of this size are impractical for most research and business applications and this likely contribute to why most transfer learning research utilises the five aforementioned networks.

Table 2.4.1: Comparison of ImageNet performance on typical models used in transfer learning. * models that exceed human-level performance.

Model	Top 1 Accuracy	Top 5 Accuracy	# Parameters	Year	Ref.
AlexNet	63.3%	84.6%	60M	2012	[2]
MobileNet	70.6%	89.5%	4.24M	2017	[75]
VGG16	74.4%	91.9%	138M	2014	[76]
VGG19	74.5%	92.0%	144M	2014	[76]
MobileNetV2	74.7%	-	3.47M	2018	[77]
DenseNet-121	76.39%	93.34%	>27M	2016	[78]
ResNet50	77.15%	93.29%	25.6M	2015	[79]
ResNeXt50	77.8%	-	25M	2017	[80]
ResNet101	78.25%	93.95%	44.5M	2015	[79]
ResNet152	78.57%	94.29%	60.2M	2015	[79]
InceptionV3	78.8%	94.4%	23.8M	2015	[81]
Xception	79%	94.5%	22.8M	2017	[82]
Pre-ResNet200*	79.9%	95.2%	64.7M	2016	[83]
InceptionV4*	80.1%	95.1%	55.8M	2017	[84]
ResNeXt101*	80.9%	95.6%	<44M	2017	[80]
NASNet*	82.7%	96.2%	89M	2017	[85]

Table 2.4.2: Comparison of the top 5 performing networks on ImageNet. * models with extra training data.

Model	Top 1 Accuracy	Top 5 Accuracy	# Parameters	Year	Ref.
FixResNeXt-101 32x48d*	86.4%	98.0%	829M	2019	[86]
ResNeXt-101 32x48d*	85.4%	97.6%	829M	2018	[87]
ResNeXt-101 32x32d*	85.1%	97.5%	466M	2018	[87]
EfficientNet-B7	84.4%	97.1%	66M	2019	[88]
GPIPE	84.3%	97%	557M	2018	[89]

Classification in the Future

It should be acknowledged that impressive progress has been made, especially since the difference in Top 1 Accuracy and Top 5 Accuracy between VGG16 and FixResNeXt-101 32x48d is 12% and 6.1% respectively. However, it appears that the significant increase in parameters without matched performance means that the difference these models contribute in practice is negligible. Of course, EfficientNet-B7 is an exception and could possibly be

utilised more in the future. Yet critically, it is still approximately three times larger than InceptionV3.

Another point of criticism is that these large networks are overparameterised for new domains. There is opportunity to explore the question of whether fine-tuned models have redundant filters. It might be the case that most filters have small weights, and so are useful collectively, but by themselves aren't detecting features meaningful to humans. This directly impacts our efforts to interpret networks. In this case, attempts to prune networks to force individual filters to detect salient features may prove useful. Furthermore, given the amount of previous work, early models are still the best understood, albeit not completely. As will be discussed further in Section 2.5, the majority of interpretability methods have been developed for VGG16 and 19, ResNet50 and early Inception models. Therefore, despite improvements in the state-of-the-art, it is still expected that these models will feature in future work.

2.4.3 Image Segmentation

Instead of giving the image a single classification, segmentation has the goal of classifying each pixel in the image. For binary segmentation, problems are designed so as to detect background and foreground content, or cancer and non-cancer regions. In a rich multi-class problem, it is referred to as semantic segmentation, because each pixel is assigned to a meaningful class within the context of the image *e.g.* self-driving cars assign every pixel in their camera to a meaningful class in the context of a driving environment. Segmentation has traditionally been performed using hand-engineered features and thresholding techniques, and in many cases, the later method is still useful and efficient. However, many segmentation problems require the high-level feature detection provided by deep learning which makes feature engineering obsolete.

Performing semantic segmentation via deep learning requires large amounts of labelled data. This often comes at a prohibitive time cost in most problem domains. Similar to ImageNet, benchmark datasets such as PASCAL VOC[90] and Cityscapes[91] facilitate the development and comparison of segmentation networks. These networks can also leverage the features learned from ImageNet by adapting pretrained networks to fully convolutional networks (FCNs) (Figure 2.4.8). This simply converts fully connected layers to convolutional layers with a 1×1 filter and using the same weights can perform classification on each sub-region of the image (point-wise convolution). The first demonstration of this proved to be hugely successful for segmentation and most works are derived from the foundation created by Long *et al.* [92] in 2015. They converted and fine-tuned AlexNet, VGG16 and InceptionV1 on the PASCAL VOC dataset. As seen in Figure 2.4.8, in general a network processes an image and outputs a segmentation mask. As a supervised learning tasks, this requires ground-truth labels to be in the form of a segmentation mask. The mask allocates every pixel in the image to one of the predefined classes.

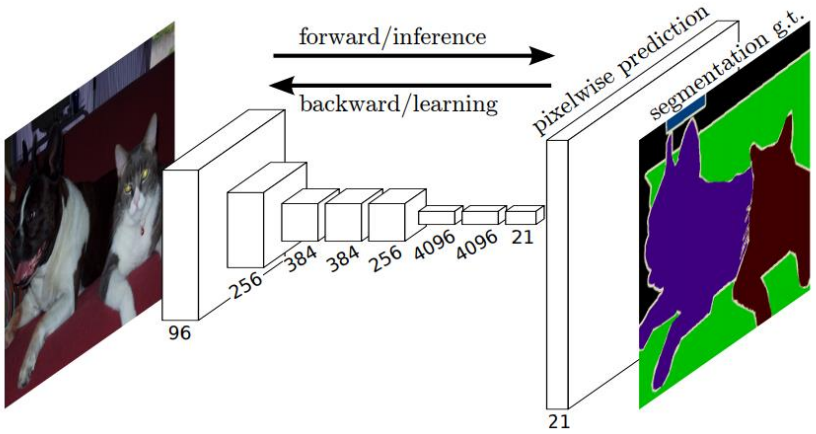

Figure 2.4.8: An example of how FCN performs semantic segmentation. The ground-truth segmentation mask assigns every pixel to a class. The figure is from Long et al., reproduced under Creative Commons licence.

During training, a network is optimised via gradient descent, often using binary cross-entropy or categorical cross-entropy loss. There is often a class imbalance in segmentation problems, particularly for background pixels. To compensate, the contribution of each class can be weighted, β_i so that each class contribution is balanced (Eq. 2.4.5). Focal loss attempts to down-weight the contribution of easy examples to focus the CNN on hard cases[93]. This includes two hyperparameters α and γ which typically need to be manually tuned (Eq. 2.4.6). Another common loss function is Dice loss (Eq. 2.4.7) which measures overlap of positive and negative predictions. These losses can be combined with other regularisation terms to improve performance *e.g.* total variation[94].

$$WCE = L(y, \hat{y}) = -\sum_{i=1}^{C} \beta_i y_i \cdot \log(\hat{y}_i)$$

Equation 2.4.5: Weighted Cross-Entropy for a multi-class problem. The average loss would normally be computed over n examples.

$$FL = L(y, \hat{y}) = -\sum_{i=1}^{C} \alpha(1 - \hat{y}_i)^{\gamma} \cdot \log(\hat{y}_i)$$

Equation 2.4.6: Focal Loss. The average loss would normally be computed over n examples.

$$DL = L(y, \hat{y}) = \sum_{i=1}^{C} 1 - \frac{2y_i\,\hat{y}_i + 1}{y_i + \hat{y}_i + 1}$$

Equation 2.4.7: Dice Loss. The average loss would normally be computed over n examples.

The introduction of deconvolution (or transpose convolution) layers by Long *et al.* lead to the development of U-Net[95], an autoencoder style network (Figure 2.4.9) that performs segmentation at the original input image size. It further contributed the use of long-range skip-connections, which reintroduce low-level information to improve the quality of the segmentation. This network architecture has inspired other work for both 2D[35,96,97] and 3D[98] segmentation tasks.

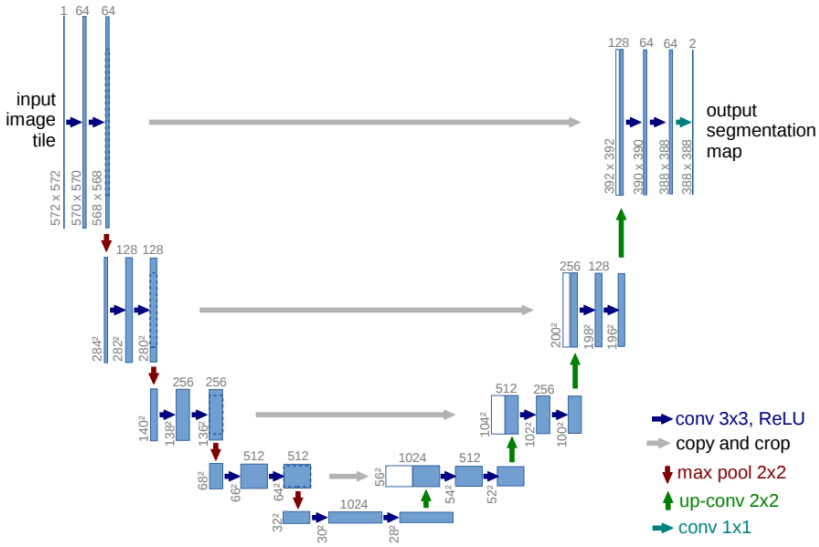

Figure 2.4.9: U-Net architecture used in semantic segmentation tasks. The encoder phase extractions information in a progressively high-level manner. The decoder phase then up-samples the high-level representations back to the original image size. At each up-sampling step, low-level information from the encoder phase is added using skip connections (copy and crop). This enables fine details to be used for segmentation on-top-of the high-level features. The figure is from Ranneberger *et al.*[95] reproduced under Creative Commons licence.

Measuring Segmentation Performance

The performance of segmentation tasks can be measured in several ways. Let n_{ij} be the number of pixels of class i predicted to belong to class j, where there are n_{cl} different classes. Let $t_i = \sum_j n_{ij}$ be the total number of pixels of class i. With that we compute pixel accuracy (Eq. 2.4.5), mean accuracy (Eq. 2.4.6) and mean intersection over union (IoU) (Eq. 2.4.6).

$$\frac{\sum_i n_{ii}}{\sum_i t_i}$$

Equation 2.4.8: Pixel Accuracy

$$\frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i}$$

Equation 2.4.9: Mean Accuracy

$$\frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{(t_i + \sum_j n_{ji} - n_{ii})}$$

Equation 2.4.10: Mean IoU

What metric is reported varies across publications, but for comparison on challenge datasets the IoU is used. In 2015, Long *et al.* reported an IoU of 67.2% on the PASCAL VOC challenge. The current state-of-the-art of an IoU of 89% was reported in 2018 by Chen *et al.*[99].

Advanced Segmentation Architecture

The model from Chen *et al.*[99] is called DeepLabv3+ and is an update from previous versions[100,101]. The key contributions of the DeepLab method are modifications to standard convolution processing. The first are dilated (or atrous) convolutions which enlarge the field

of view of filters to incorporate larger context without increasing the number of parameters[100](Figure 2.4.10.a). The second is Atrous Spatial Pyramid Pooling (ASPP), which performs several dilated convolutions at several different rates and merges the outputs using a pointwise convolution (Figures 2.4.10.b, Figures 2.4.11.b). These operations are performed on the feature maps generated by VGG16 and ResNet-101. The unique contribution of DeepLabv3+ is utilising depthwise-separable-dilated convolutions (Figure 2.4.11) in combination with ASPP on a modified Xception model. They also utilised a U-Net-like decoder module to improve the boundaries of the segmentation. This state-of-the-art model therefore is the culmination of all previous successful methods. In addition to the mean IoU, qualitatively the results are very impressive. Importantly, the large gains in performance were also due to extra pre-training on the COCO[102] and JFT[103] datasets as well as merging several multi-scale inputs (1x, 0.75x, 0.5x). However, the demonstrated success of dilated convolutions to detect features at different scales will likely be a useful method for both segmentation and classification problems in the future.

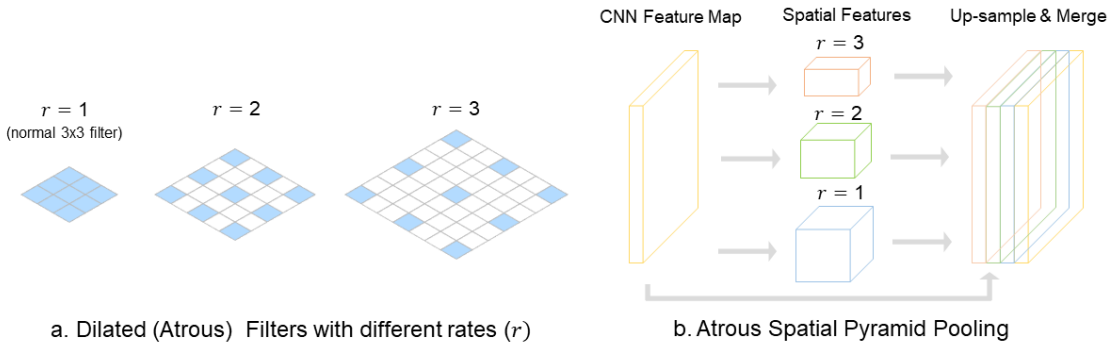

Figure 2.4.10: a) Dilated convolutions involve modifying the size of the filter by a rate. This increases the receptive field of the filter without increasing the number of weights. b) The receptive field can be tuned to detect different features and so combining features from different scales improves performance.

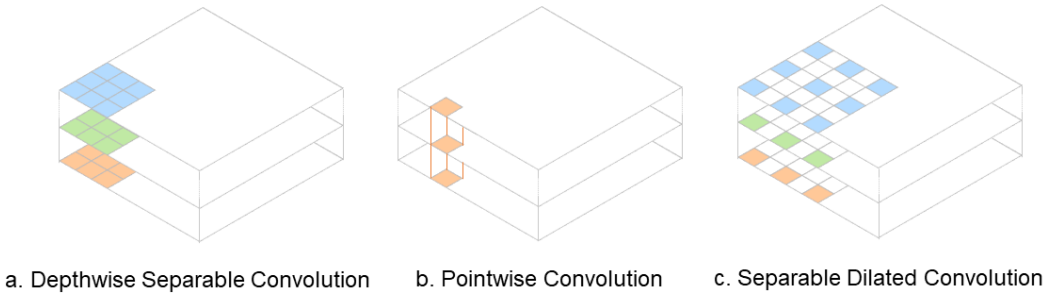

Figure 2.4.11: Depthwise separable convolutions (a) convolve channels separately and then merge them using pointwise convolutions (b). DeepLabv3+ introduces the concept of separable-dilated convolutions (c) which is used in the ASPP as well as the modified Xception backbone.

2.4.4 Generative Methods

Generative methods provide a unique way to interpret the learned representations of deep learning models because they have been trained to generate realistic outputs. Although the type of output can be anything from speech, to music, to cartoon drawings, a major area of focus is generating realistic natural images. In Section 2.4.1 the autoencoder was introduced and in the form of a CNN can input an image, learn to encode it and then decode it back to a realistic image. This demonstrates that the latent space of an autoencoder (the code) can

produce realistic images. However, the challenge is that there is no way to predict the distribution of values in the latent space. This is not a problem when generating known images from their code, but when generating new images, it becomes obvious the latent space is sparsely populated with realistic codes because the resulting generated images are not realistic images. The question then is, can we learn a latent space where every possible code has some meaning and so a realistic image can be generated with it? There are two popular methods which work towards this end.

2.4.4.1 Variational Autoencoders

Variational Autoencoders (VAE) were first introduced in 2013[104] and aim to make the latent space more predictable by forcing the latent variables to be normally distributed. Instead of the encoder transforming the input to the latent space, it instead transforms the data to create the mean μ and standard deviation σ^2 parameters of a normal distribution $\mathcal{N}(\mu, \sigma^2)$. The latent space is then created by sampling from this distribution. To avoid the network just learning the μ and σ^2 of a standard autoencoder, a standard normal distribution is defined as $\mathcal{N}(0, 1)$ and the Kullback-Leibler divergence, the difference between the learned and standard distributions, is included as an additional term in the loss. Therefore, the loss consists of the Kullback-Leibler divergence and the reconstruction loss.

Importantly, taking a random sample for the latent space means there is no gradient to enable learning. Instead, VAEs take advantage of the fact that any normal distribution can be expressed in terms of a standard normal distribution (Eq. 2.4.11). This is called the *reparametrisation trick.*

$$N(\mu, \sigma^2) \sim u + \sigma^2 \cdot \epsilon, \epsilon \leftarrow \mathcal{N}(0, 1)$$

Equation 2.4.11: General form of normal distributions as an expression of a defined normal standard normal distribution.

Therefore, a sample ϵ can be made from the normal distribution and transformed into the latent space sample z (Eq. 2.4.12). Gradients can then be computed with respect to u and σ^2.

$$z = u + \sigma^2 \cdot \epsilon, \epsilon \leftarrow \mathcal{N}(0, 1)$$

Equation 2.4.12: Transform a sample from a normal distribution to the latent space.

The architecture of a VAE is much the same as a standard autoencoder, however, the latent space is sampled from the encoded distribution mean and standard deviation (Figure 2.4.12). After training, the latent space can then be explored, usually revealing a structured latent space where variation between different inputs is meaningfully represented (Figure 2.4.13).

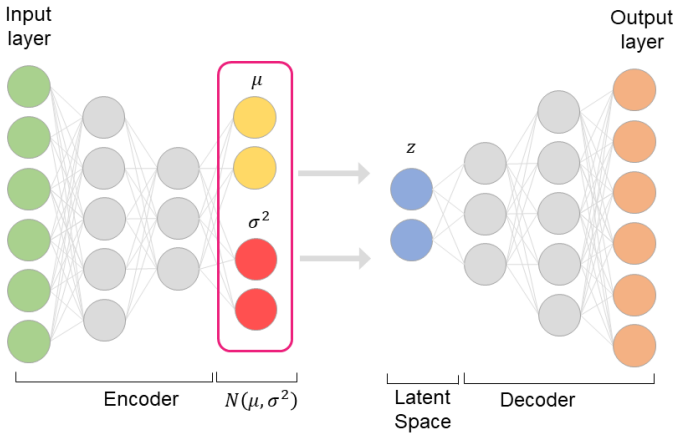

Figure 2.4.12: A variational autoencoder gives structure to the latent space by forcing it to be a normal distribution. The encoder network transforms the input to a mean and standard deviation. This is then used to sample z which is then used to reconstruct the input.

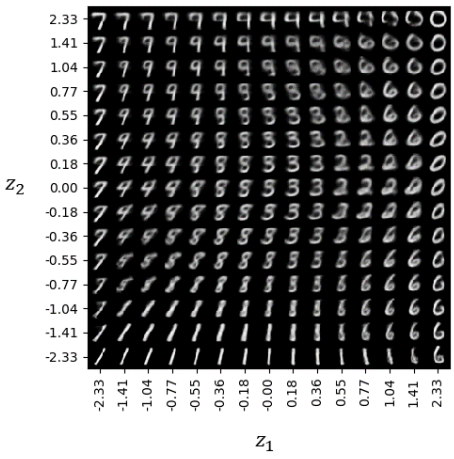

Figure 2.4.13: Exploring the latent space of a variational autoencoder reveals that codes which produced realistic images encompass the full space. It also reveals meaningful variation within and relationships between the different classes. This is an example of a VAE trained on the MNIST dataset.

2.4.4.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) attempt to force the latent space to have structure in a similar manner to VAEs by forcing the latent variables to belong a normal distribution. However, there is no encoder part to the network, the latent space is randomly sampled directly, and a generator network transforms z into a realistic image (Figure 2.4.14).

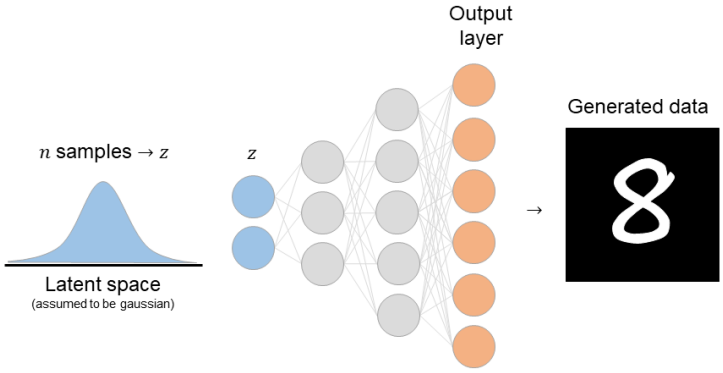

Figure 2.4.13: Generator network that transforms the latent space vector z into a realistic image.

Instead of having an explicit loss function, the generator network is trained using a *learned loss function* in the form of a discriminator network. Uniquely, there are two training phases for GANs. In the first phase, the discriminator learns to classify images as being either real or fake, with a mixed training set of real reference images and a set of fake generated images. In the second phase, the generator network uses the gradients of the discriminator to learn how to create more realistic images. GANs training therefore consists of a dynamic training regime where the two networks attempt to outperform each other (Figure 2.4.14).

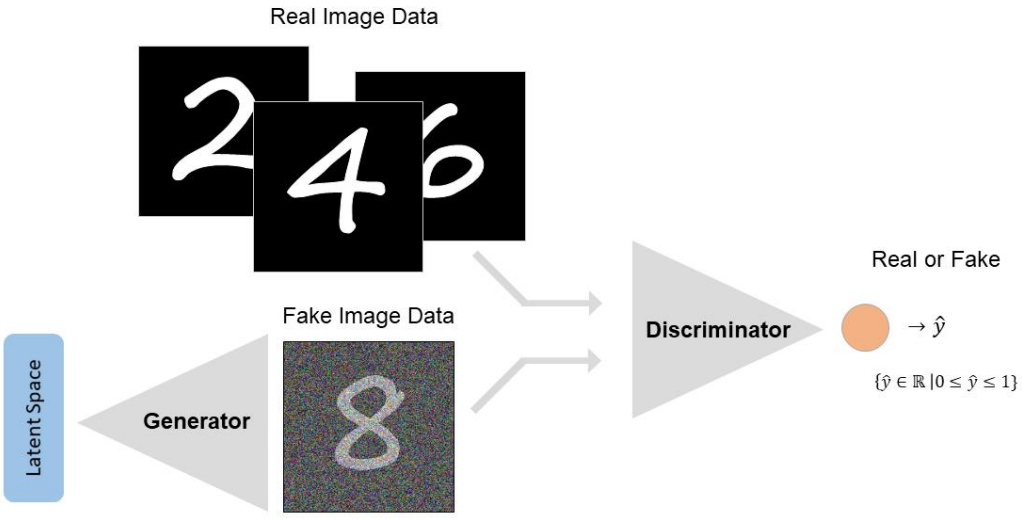

Figure 2.4.14: Training regime of a generative adversarial network.

The aim is to balance training so that as the generator improves at creating realistic images, the discriminator improves at detecting fakes. The original GAN paper[105] formalised the loss as a mini-max game between two adversaries (Eq. 2.4.13). Because it is a dynamic loss *i.e.* there is no known global minimum, the training can become unstable. Where one network starts to dominate another, it is referred to as model collapse. To avoid this, the research in the last several years has developed techniques to improve this, such as using tanh activation functions[106], using filter and step sizes that are factors of the image size[107], modifying the loss[108], iteratively growing the network[109,110] or modifying the assumptions about the latent space[111].

$$\min_{G} \max_{D} \mathbb{E}_{x \sim q_{data}(x)}[\log(D(x))] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))]$$

Equation 2.4.12: Formalisation of the original mini-max game that characterises the loss in generative adversarial network training.

GANs have now been shown to produce highly realistic looking images in almost any imaging domain. Furthermore, they also produce a highly interpretable latent space. For example, in 2019 the state-of-the-art in human face generation created an architecture called StyleGAN[111] and produced highly realistic human faces of people who don't exist (Figure 2.4.15 – a). Similar to VAEs, the latent space can be explored, demonstrating that abstract features such as hair colour, smile and glasses have been learned independently[111]. Another recent work created an architecture called BigGAN[112] demonstrated their model can represent huge amounts of variation and successfully reconstruct images from ImageNet (Figure 2.4.15 - b).

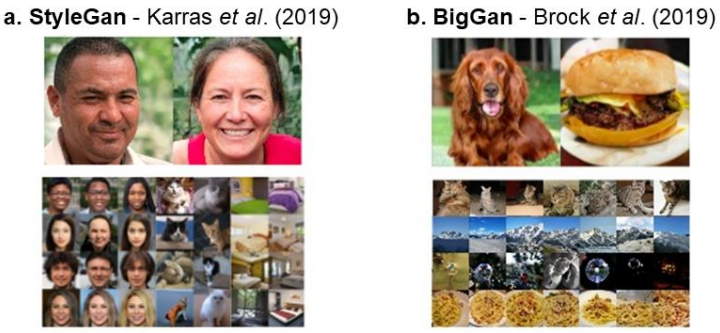

Figure 2.6.15 State-of-the-art GAN models that produce highly realistic images across diverse domains. The figure is adapted from results produced by Karras *et al.*[111]. and Brock *et al.*[112] under Creative Commons Licence.

2.5 An Overview of Interpretability Methods

Coinciding with the development of deep learning has been a burgeoning interest in understanding how these systems work. To that end, attempts at interpreting deep learning models are wide and varied, and have recently come to be complimentary. Indeed, they all attempt to translate machine representations to representations understandable by humans. As discussed in Section 2.3, this is a crucial component in the development of medical AI. Therefore, this section will explore the degree to which current interpretability methods can provide insight into how models represent the world and how they make decisions.

2.5.1 Dimensionality Reduction

Raw data often contains too much information and is too high-dimensional to make sense of. As discussed in the previous section, the hidden layers of neural networks transform data into lower-dimensional representations, often referred to as the latent space. Networks utilise the lower-dimensional latent space to successfully perform classifications *etc*. However, a problem that faces network interpretability is that the latent space is still too high dimensional for humans to make sense of. For example, the size of the feature vectors learned by VGG16 and ResNet-50 to perform classification are 4096 and 2048 respectively. How should we interpret these vectors?

Each feature to contains high-level concepts which are sufficient to distinguish a class. Therefore, the vector can be thought of as evidence for the presence or absence of a particular feature in an image. From this perspective, we can represent the information in an image as a point in this high dimensional space e.g. 4096 dimensions where each dimension is a feature. We can then use dimensionality reduction algorithms such as PCA, t-SNE[113] or UMAP[114] to transform the 2048D space into a 2D or 3D space, which can then be easily visualised (Figure 2.7.1). Qualitative correlations can then be assessed by comparing images within each cluster or neighbourhood that may emerge. As an interpretability method it enables us to check that a network is representing the world according to our own expectations *e.g.* cats and dogs cluster together. If it isn't, perhaps it discovered unknown patterns and so can be used for hypothesis generation. Dimensionality reduction techniques can also be used for labelling data that has been analysed using unsupervised learning. Clusters can be visually inspected and quickly allocated to a known class.

2.5.1.1 Principle Component Analysis

PCA transforms a set of possibly correlated variables into a set of linearly uncorrelated variables called principle components. These components are defined in such a way that the successive components explain less and less of the variance observed. Each image is thus given a weighting for each component, and the first two or three can be visualised. The components themselves often have

little meaning in complex datasets and may not correspond to human-relatable directions of variance *e.g.* a "dog-ness" vector. However, given that neural networks can extract high-level information such as "pointy ears" or "floppy ears", some components may represent orthogonality between such features for some problems *e.g.* main difference between cats and dogs. PCA is a hugely popular technique across many disciplines and is a valuable tool given that it is inexpensive computationally.

2.5.1.2 t-Distributed Stochastic Neighbour Embedding (t-SNE)

t-SNE is a method that assigns high-dimensional points to points in two or three dimensions. Through a stochastic optimisation process, the algorithm attempts to maximise the probability that similar points in higher dimensions are modelled by nearby points in lower dimensions. Likewise, dissimilar objects in higher dimensions should have a high probability of being modelled by distant points in lower dimensions. This technique has wide spread use in the machine learning literature as it often provides compelling visualisations. However, interpreting the results is difficult and developing experience with the behaviour of the algorithm is recommended to avoid common pitfalls[115]. For example, the algorithm is non-linear and applies transformations to different regions of the data. Therefore, the size of clusters has no meaning as it expands dense clusters and shrinks sparse ones. Furthermore, the distances between clusters might not mean anything. The algorithm also requires fine-tuning of hyperparameters such as perplexity (attention to local and global aspects of the data), learning rate, and steps. There is also no set number of steps that yields a stable result and most datasets will need to be visually inspected for quality.

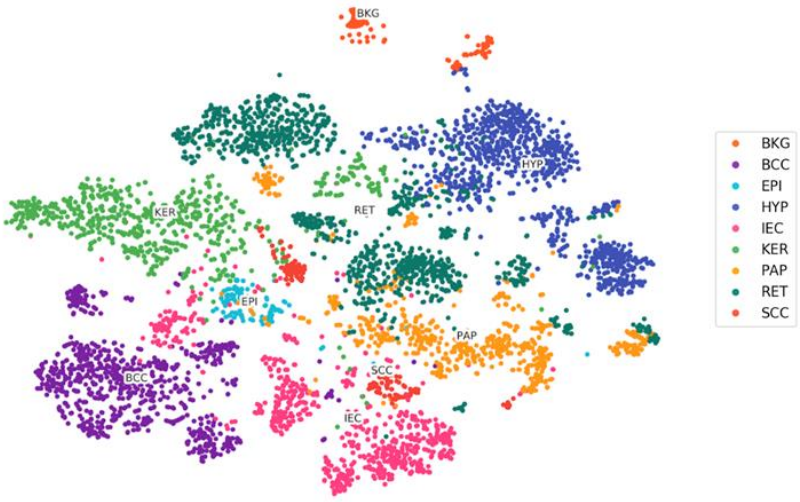

Figure 2.5.1: Example of dimensionality reduction on the ResNet-50 feature vector. This shows the results of the t-SNE algorithm applied to features of nine tissue classes in the context of skin cancer. The graph should be interpreted with full knowledge of the limitations of the algorithm.

2.5.1.3 Uniform Manifold Approximation and Projection (UMAP)

UMAP is a recent (2018) dimensionality reduction algorithm that has had increasing presence in the literature. It has solid theoretical backing as a manifold approximation technique that combines fuzzy-logic, topological data analysis and optimisation theory. It produces similar looking visualisations as t-SNE, however, it arguably has some advantages over it. Firstly, it is faster than t-SNE and purportedly captures global structure better. It is also a general-purpose dimensionality reduction technique whereas t-SNE is designed for two or three dimensions. Wider use of this algorithm will enable its merits and limitations to be better understood.

2.5.2 Attribution

Optimisation via gradient descent modifies the weights of a network relative to their contribution to the loss. A non-obvious feature of this process is that it is also possible to find the derivative of the loss with respect to the input image. Utilising a pretrained network, the contribution of each pixel in the image to an objective function *e.g.* the probability of being "Golden Retriever", can be determined and visualised (Figure 2.5.2). The visualisations are called saliency maps and are a popular spatial attribution method[116–120]. They consist of simple heatmaps highlighting areas in the image that most caused a particular classification. However, individual pixels are not a useful unit of attribution, because by themselves are far removed from high-level concepts like the output class. Instead of looking for pixel-level attributions, we can look at high-level features detected by the network, and then project them back to the original pixels.

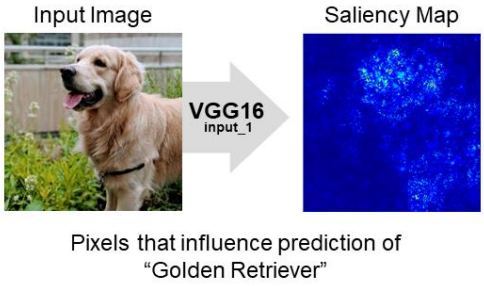

Figure 2.5.2: A saliency map generated via back-propagation, using gradients to determine which pixels contribute to the classification of "Golden Retriever". The general outline of the dog is clear however individual pixels do not carry meaningful information for us to interpret.

2.5.2.1 Grad-CAM

The most common attribution method is called Grad-CAM[118,121]. The method consists of taking an output feature map from a CNN for a given image and weighing every channel in the feature map by the gradient of the class with respect to the channel. The mean weighted activation across all channels is calculated for each spatial region. The resulting heatmap can then be used to interpret the importance of high-level features detected in the image to a specific class activation. For example, Figure 2.5.3 shows several Grad-CAM saliency maps for an image predicted as "Golden Retriever", and what image regions contribute to various other classifications. Critically, for some networks and for some layers, saliency map methods have been shown not to be robust against class randomisation or weight randomisation[122]. Indeed, layers preceding the last spatial feature map tend to give meaningless results *i.e.* all pixels contribute to all classifications. Recent work has also shown this method to be vulnerable to adversarial attacks[123]. Generally, these methods need to be interpreted with care and should include sanity checks *e.g.* in Figure 2.5.3 a random class of "bakery" shows that the image doesn't contain any meaningful information for that prediction. Importantly, this technique traditionally limits the user's ability to control what concepts of interest are present in the map. In the case of having two cat pictures, one map may attribute ears while the other may attribute eyes and nose. How can we compare how ears, nose and eyes contribute generally to the prediction of cat?

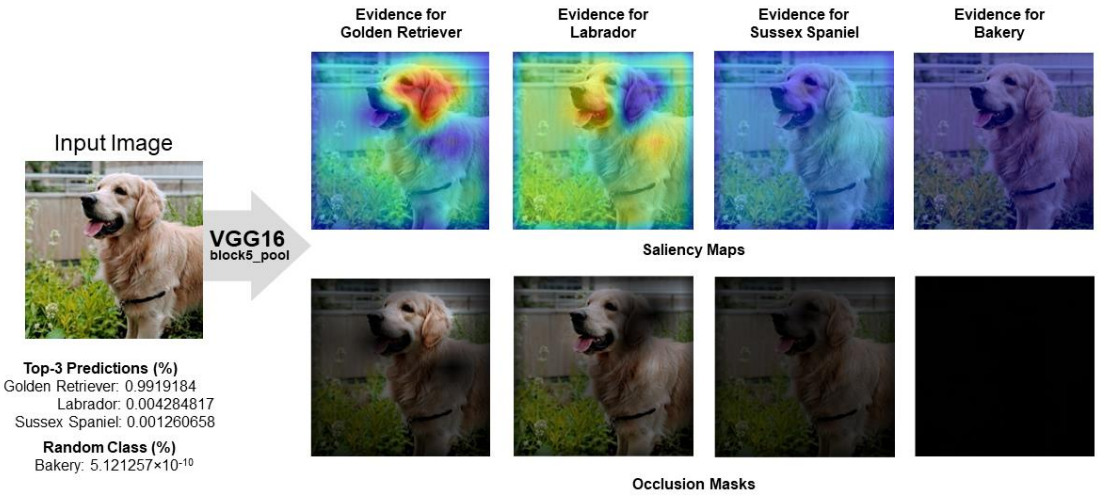

Figure 2.5.3: An example of Grad-CAM applied to an image of a dog. The Top 3 predictions can be used to determine which parts of the image contribute to each class. The saliency map shows that ears are important in classifying "Gold Retriever", perhaps more so than anything else. The saliency map for "Labrador" shows that a large part of the image influences that class, particular the dog face and body. A random class of "Bakery" is used as a sanity check, showing that no information in the image is useful for that classification.

2.5.2.2 TCAV

An alternative method is called Testing with Concept Activation Vectors (TCAV)[124]. The method generates activation features from layer l for a set of positive concepts of interest *e.g.* images that contain "ears", as well as from a negative set *e.g.* "not ears". Given the activation space, a linear separation is performed, and then an orthogonal vector is found which represents a continuum of "ear" to "not ear", *i.e.* a CAV. For a given image x of class k, the sensitivity of the prediction to changes in the feature map (at layer l) in the direction of the CAV can be gauged using derivatives. This computes a score $S_{c_{k,l}}(x)$, either positively or negatively affecting a given prediction. A model's conceptual sensitivity across an entire class, *i.e.* TCAV score, can be measured as the proportion of positive scores for all images of class k. A TCAV of 0 means the concept negatively affects the class prediction, a TCAV of 1 means the concept positive affects the class prediction, and a TCAV of 0.5 shows no affect. To guard against spurious results, this process is repeated 500 times using a random negative set, to ensure the CAVs are actually significant for a class prediction. This is measured using a two-tailed t-test. Different concepts may be captured at different layers of a network, and so generating CAVs for each layer provides further insight into the way a network is modelling the world. In the case of medical diagnosis, TCAV scores can help gauge whether deep learning systems are using the same evidence as doctors in generating a diagnosis[124] (Figure 2.5.4). Potentially this could be used to tune the importance that networks give to each CAV to be better aligned with background knowledge. However, a limitation of this approach is that images need to be labelled according to the concepts you want to check are being represented. This can be time consuming and difficult because it is not clear how to visually capture abstract concepts such as "symmetrical". It is also not clear whether a network is using an unbeknown conception of symmetry that just happens to be different to the CAV. The testing of the robustness of the CAVS also adds to the time required to interrogate the method.

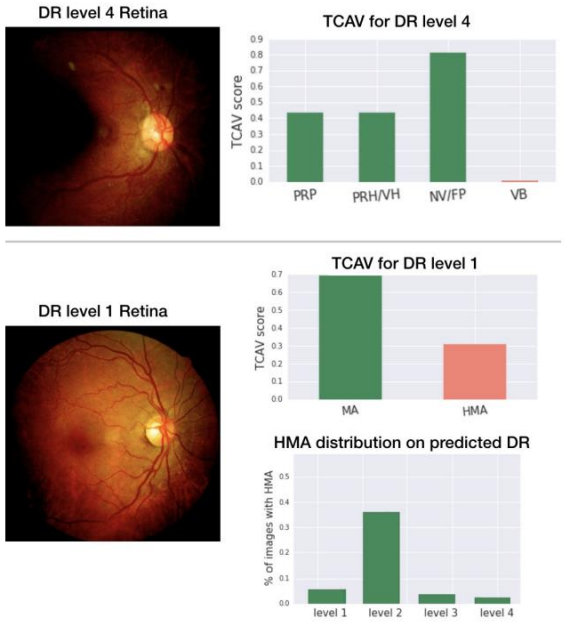

Figure 2.5.4: Example of diagnostically relevant CAVs and their contributions in grading diabetic retinopathy. The figure is from Kim *et al.*[124] reproduced under Creative Commons licence.

2.5.3 Feature Visualisation

Feature Visualisation attempts to understand a network by synthesising input images to elicit a desired output response. As a generative method, this approach allows a network to directly show us what it has learned. This is a powerful way to understand a network because it separates causative features from things which merely correlate with them. For a given layer of a network, we can select a filter and then use gradient descent to optimise a random input image to maximally activate the filter. Applying this method to layers of increasing depth reveals that, as expected, networks learn to detect abstract concepts *e.g.* fur, pelican head, dog ear and snake body (Figure 2.5.5). However, a common artefact of these visualisations is the rainbow colour spectrum of which the patterns are comprised[125–128]. In effort to generate more natural looking images, regularisation techniques *e.g.* image parameterisation tricks, gradient blurring and image blurring and jittering, have been employed. Many hypotheses have been conjectured as to why regularisation is required, however recent work suggests that they are not necessary if the network is robust to adversarial attacks[129,130]. This new hypothesis posits that networks have previously learned non-robust features which imperceptibly correlate with class predictions. Visualising these learned representations results in what appears to us as high-frequency noise and colour artefacts which actually helps to maximise a filter. With a network that has learned robust features, feature visualisation works without any regularisation techniques and has more natural colours[131].

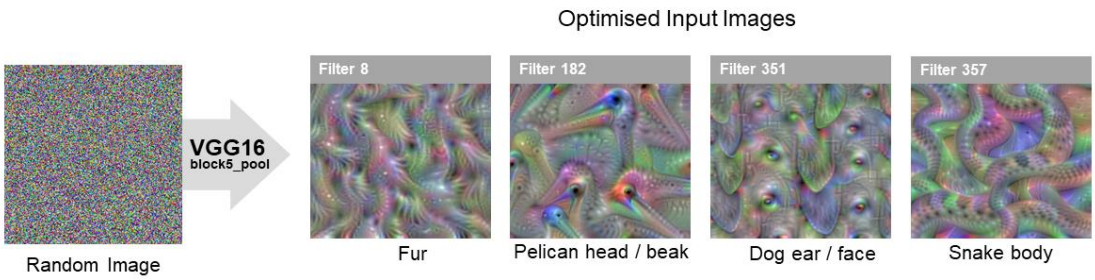

Figure 2.5.5: The feature visualisation algorithm optimises a random input image to maximise a filter for a given layer. During training of the network each filter was optimised to detect different features in an image, and this method enables these features to be independently inspected. At later layers the filters detect high-level concepts such as fur, pelican head, dog ear and snake body.

Maximising the activation of individuals filters is very useful to understand how a network works. However, we really want to interpret what a network is actually seeing when processing an image. In this case, spatial regions contain different responses to an image, and each response includes interactions between multiple filters. Feature visualisation at this resolution was first performed in 2018 and still stands as the state-of-the-art[132]. For a given activation feature map A of dimension $i \times j \times f$, the dot product of spatial region $A_{i,j}$ (length f) for both the real image feature vector and the generated image feature vector is maximised via optimisation. The resulting visualisation represents the information that the network is actually capturing at that region for a given layer (Figure 2.5.6). For deep layers, *e.g.* MIXED5A, representations can capture concepts such as dog head, dog body and cat face. This technique can be coupled with attribution techniques such as Grad-CAM and TCAV, but instead using the now interpretable network representations. For instance, it can be determined how each high-level concept influences a given class prediction. Furthermore, these visualisations can then be combined with t-SNE or UMAP to show the relationships between the concepts the network has learned to detect, not just the correlations they have with real images[133].

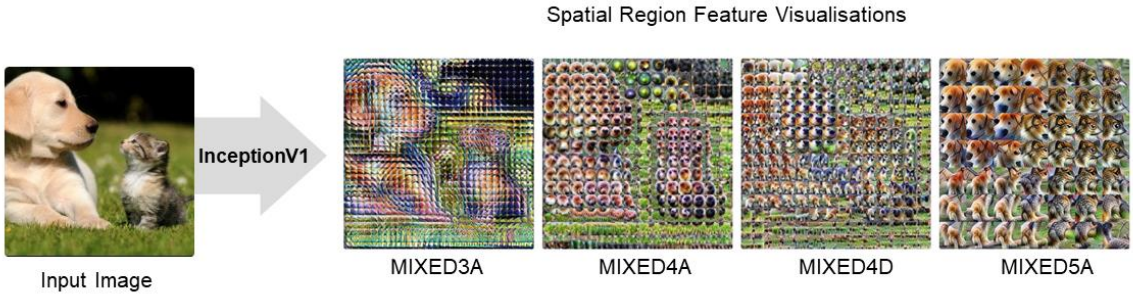

Figure 2.5.6: Example of visualising a networks interpretation of an input image for a given region. This can be done for different layers, where later layers are detecting dog heads, dog bodies and cat faces. The figure is adapted from results produced by Olah *et al*.[132] under Creative Commons Licence.

The major criticism of feature visualisation and the state-of-the-art examples[128,132,133] is that they cannot be reproduced reliably, or at all, for other networks. Indeed, work in Figure 2.5.6 and from refs.[128,132,133] are all produced using the InceptionV1 (GoogLeNet)[134] network which is not widely used in transfer learning. Recent discussion on learning robust features suggests that the discrepancy in reproducibly may be overcome. This is motivated by the case of neural style transfer which has previously only been shown to work effectively using VGG networks[135]. Importantly, weights for such robustly trained networks need to be made available. However, this only addresses features for ImageNet and so networks adapted to different domains may still acquire non-robust features during fine-tuning. An alternative could be to use existing weights but perform fine-tuning using injected noise to learn domain specific robust features[136]. In any case, combining visualisation with dimensionality reduction and attribution techniques demonstrates a powerful way to interpret deep learning systems and sets a new minimum standard to work towards for high-stakes decisions.

2.6 Digital Pathology

With a foundational understanding of deep learning for image classification and segmentation, as well as image generation and model interpretability, their application to the histological diagnosis of cancer will now be discussed. The field of digital pathology will first be introduced, and then recent contributions to the field will be explored. Having identified limitations to the current literature, the area of skin cancer diagnosis will then be reviewed and areas for meaningful contribution will be identified.

2.6.1 An Emerging Field

Digital pathology is a multidisciplinary field that combines biomedical imaging and histopathology, the microscopic identification and study of disease. Histopathology as a discipline began over 180 years ago and has continually evolved alongside technological progress, with much of the workflow now automated[137]. However, the tissue analysis and diagnosis have historically required a pathologist to look at microscopic sections of tissue on a glass slide, typically between 4x and 100x magnification on a compound microscope. The tissue is usually dyed with hematoxylin and eosin (H&E) which helps discriminate different tissue types and abnormalities[138], although immunohistochemical staining is now also used[137].

In 1997, the conception of the "virtual microscope"[139] moved the field in the direction of digitising slides. The capabilities of such a system were favourable given that it would improve workflow ergonomics and allow digital slides to be transmitted over distances quickly, enabling sharing of slides for second opinions (telepathology) and education. By 2010, digital pathology was increasingly used in education, but the limitations and cost of slide scanning technology as well as the resulting storage requirements of whole slide images (WSI) slowed adoption[140]. In 2016, moving to a complete digital system was still projected as a future possibility, with the cost of scanners cited as a major limitation [141]. Critically, the regulatory status of the scanners posed a challenge to adoption since they were not approved for clinical use. Progress was made in 2017 when the U.S Food and Drug Administration (FDA) released guidelines which enabled manufactures to work towards common standards[142]. In May 2019, the FDA approved the Leica Biosystems Aperio AT2 DX System for clinical diagnosis in the United States[143], which for the first time enables pathologists to diagnose slides away from a microscope. With the expectation that more manufactures will receive approval soon, it can be said that the age of digital pathology has finally arrived.

The recent success of AI in computer vision tasks has directed researchers to work on problems in digital pathology and the development of computer-aided-diagnosis systems[144]. In fact, results using deep learning for histological diagnosis of prostate cancer[145], lead the company Paige.AI to receive "Break-through Device" designation from the FDA in March 2019[146]. In practice, this designation merely speeds up the assessment and review stages of the traditional statutory approval standards which protect and promote public health. However, it is clear AI will play a major role in the development of digital pathology systems. Therefore, the remainder of this section will discuss recent applications of deep learning to digital pathology.

2.6.2 Recent Work

Deep learning has been applied to classification[145,147–149], detection[150–157] and grading[158–163] tasks across a variety of problems in digital pathology within the last four years. The most

commonly studied areas are breast cancer[154,159,161], breast cancer metastasis to lymph nodes[145,151,155–157] and lung cancer[147,160,162,163], which are in a large part due to the creation of challenge datasets[42]. Other works have studied prostate cancer[158], colorectal cancer[149], ovarian cancer[148], and skin cancer[145,164]. Others have developed more general techniques such as lymphocyte detection[150] and nuclei detection[152,153,159] which are important for characterising breast cancer and melanoma. Interestingly, other unique applications of deep learning in digital pathology include DeepFocus[165] which improves auto-focus capabilities for slides scanners, as well as image correction tasks such as stain normalisation using GANs[166].

In terms of the deep learning algorithms, the majority of works utilise pretrained convolutional neural networks (CNNs) such as in the use of VGG16[149], ResNet50[167] and InceptionV3[168] for transfer learning. In the case of segmentation and detection tasks, architectural variations are added to these base networks as needed, such as the U-Net network with skip-connections[34,95], and sometimes the addition of recurrent neural network (RNNs) layers seen in references [145,149,169]. In all these settings, supervised learning is the predominant approach although an early work from 2015 utilised unsupervised representation learning using autoencoders for basal cell carcinoma classification[170].

Dataset Sizes

The amount of data used in these studies varies between problems. For example, for nuclei detection 47 images were used but included 17,700 ground-truth annotations[152], compared to lung cancer prognosis prediction where 2,186 WSIs were used[163]. Comparing recent studies, Zhang *et al.*[169] used 913 slides for bladder cancer analysis, whereas Campanella *et al.*[145] have so far collected the largest dataset for digital pathology problems; 24,859 slides for prostate biopsies, 9,962 slides of skin data (basal cell carcinoma) and 9,894 breast cancer metastases to lymph node slides. This work will likely influence future studies to work with ever-larger datasets. However, WSIs are gigapixel level images and so annotating and/or simply managing data of this size is a unique challenge to digital pathology. Indeed, Campanella *et al.*[145] acknowledge that when using such a large dataset it is infeasible to rely on supervised learning which requires expert annotation. Their approach instead used "weakly"-supervised learning, taking the slide-level label to develop a binary-classifier of cancer/non-cancer for classifying the enormous amount of tissue variation found in the WSI. Developing a scalable analysis pipeline is therefore a key consideration when working with such large datasets. Conversely, smaller data sets may enable more specific labelling to answer more complex and clinically relevant questions.

Clinical Utility

Campanella *et al.*[145] suggest their system could be used to improve the efficiency of pathologist workflows by ordering work by probability of cancer being present in the slide. However, this approach has limited ability to improve overall efficiency since in this use-case a pathologist will always view the slide regardless of whether there is cancer or not. Zhang *et al.*[169] utilise their much smaller dataset to produce interpretable outputs which produce regions of interest (ROIs) in bladder cancer slides as well as generate anatomical text descriptions which characterise the ROIs *e.g.* "Moderate pleomorphism is present. Moderate nuclear crowding is seen. Polarity is negligibly lost. Mitosis is rarely visible. Nucleoli are not observed or exceedingly rare. Low grade.". In this case, a pathologist is not only viewing a

specimen, but much of their routine analysis is performed or supported by the AI system. Indeed, much of the work previously mentioned works to this end. Other work on automatically measuring the extent of tumours[171] again is directed at improving the efficacy of routine work. Applications like these can arguably prove more useful to the pathologist workflow but require much more effort and expertise to develop which impacts scalability. Indeed, the problem of scalability is a persistent problem and is an important area of research. If clinically useful methods have been demonstrated on small datasets, what is the best way to replicate their results at scale? One possible way would be to employ weakly-supervised or unsupervised learning methods. This possibility the later will be discussed further in Section 3.

2.6.3 Digital Pathology and Skin Cancer

To date there has been little work done in the field of digital pathology for skin cancer. In 2015, stacked autoencoders were used for unsupervised feature learning which were then used for binary classification[164]. A recent work (2019) has looked at identifying BCC using histological images captured using smart-phones[172]. They used 6,610 images of BCC with binary annotations as BCC/non-BCC. They performed low resolution segmentation and as well as classification for the whole image. Another work (2019) performed binary classification for BCC/non-BCC using the standard patch-technique using 9,962 WSIs of BCC and non-BCC cases. They similarly used the patch technique alongside RNNs to classify and sort slides based on the probability of the presence of cancer[145]. None of these works use common data and so comparison between them is difficult. However, they demonstrate the effectiveness of deep learning systems for detecting this cancer type.

Computer-assisted-diagnosis systems have been developed for histological classification of melanoma since 2003[173]. Despite progress in the field, in 2015, one work argued that deep learning approaches where not applicable as melanoma requires specific diagnostic features to be considered rather than the presence or absence of patterns associated with other skin cancers[174]. They consequently utilised hand-crafted feature extraction processes to automatically classify melanocytic nevus, melanoma and normal skin cases using WSI data with ~90% accuracy. However, given further developments with deep learning technology it is likely that this argument is now, or possibly never was, valid. Indeed, recent work claims pathologist-level classification for melanoma base on the discordance rate between the model (19%) and inter-observer variation of pathologists (22-25%)[175]. This network was trained on 595 images of melanoma and nevi and further tested on 100. This work claims to be the first to apply deep learning techniques to melanoma classification. Associated work in digital pathology has explored whether z-stacked WSI data improves diagnostic accuracy for pathologist[176]. Their conclusion was that z-stack capability made no significant difference in diagnosing melanocytic lesions, which has positive implications for the current methodologies of deep learning systems in digital pathology.

Another recent work (2019) claims to be the first work to investigate SCC detection using CNNs[177]. They performed a binary classification tasks using the standard patch technique to identify cancer and non-cancer regions in 192 WSIs. Their stated aim was to demonstrate the feasibility of machine learning for this task. There is no known work on applying machine learning to the analysis of images of actinic keratosis or IEC.

Critically, no work demonstrates multi-class classification or segmentation in skin cancer, either for multiple cancer types or healthy tissue types. Further, no work has so far demonstrated meaningful application in a clinical setting. They still require a high level of input from a pathologist for a clinically useful assessment. Indeed, most work in digital pathology has so far lacked the necessary interpretability interfaces and assurances required for high-stakes decisions. It is my opinion that ref.[169] (discussed in Section 2.6.2) stands as the best example of what direction the field needs to move. Therefore, there is opportunity to significantly contribute to the field of digital pathology and skin cancer diagnosis.

2.7 The Biology of Skin Cancer

Skin cancer is divided into melanoma and non-melanoma skin cancers. Non-melanoma skin cancer largely refers to basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) which are the most common forms of cancer in both Australia and the United States[178,179]. BCC comprises approximately 60% of all skin cancer diagnoses, with SCC comprising 30%[180] and melanoma comprising approximately 4%[181]. Although skin cancer has a lower mortality rate compared to other cancers, two in three Australians will at some point be diagnosed[179] contributing to the increasing economic burden of disease[182]. Research on improving the accuracy and efficiency of skin diagnosis is therefore a promising use-case for deep learning technology.

2.7.1 Anatomy of the Skin

In order to understand skin cancer, it is valuable to place it in the context of healthy skin tissue. For a complete overview of the skin see Bolognia, Jorizzo & Rapini[183]. The skin is referred to anatomically as the integumentary system and is the body's largest organ, constituting 6% of the bodies overall mass and covering an area of $2m^2$. The skin is involved in the protection, thermoregulation, and metabolic function of the body, as well as serving as the main way in which we derive sensation from the environment. It is a multi-layer organ whose function is due to specialised mini-organs at various layers such as hair follicles, eccrine, sebaceous and apocrine glands, fingernails and mammary glands. The main layers (Figure 2.5.1) are divided into the stratum corneum, epidermis, dermis and hypodermis.

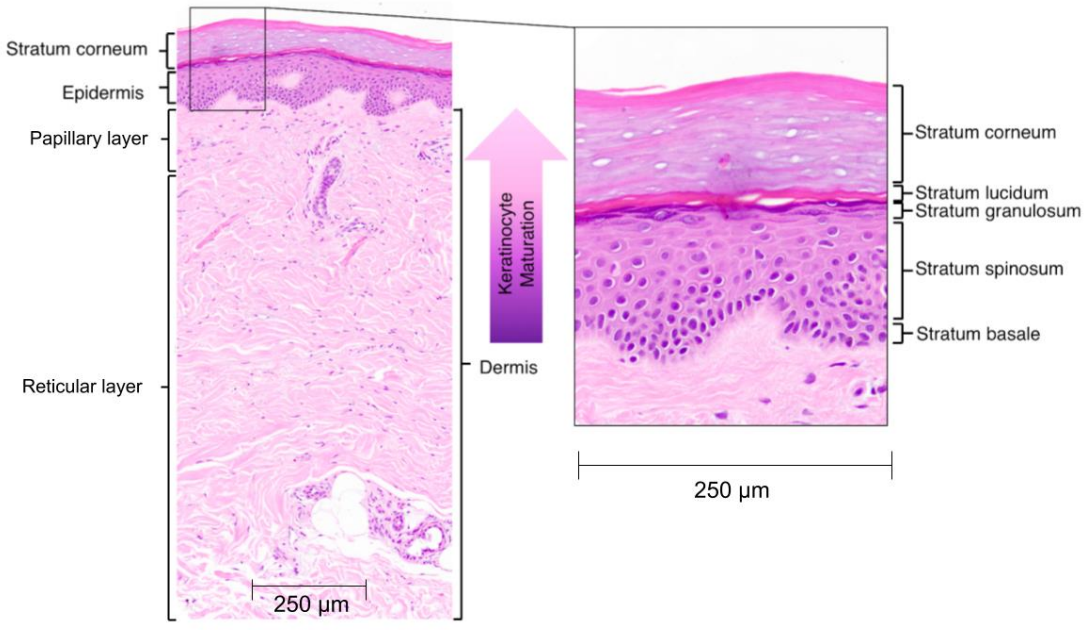

Figure 2.7.1. A histological section of skin showing the main layers; the stratum corneum, epidermis, dermis and hypodermis. Each layer has sublayers that contain specific features or define locality. The dermis extends well below the superficial surface and so the bottom layer, hypodermis, cannot be shown.

2.7.1.1 Stratum Corneum

The stratum corneum is the most superficial layer of the skin. It is composed of interwoven layers of keratin, a structural and protective protein produced by keratinocytes. Keratin is all that remains at the end of the keratinocyte's maturation pathway. The cells are continually shed as new layers are formed beneath them. This layer ranges in thickness, being thinnest on facial areas and thickest on the hands and feet. Adaptive process enable the thickness to vary in areas of high use *e.g.* callouses on hands and fingers, however abnormal proliferation of this layer, termed hyperkeratosis, is associated with numerous skin conditions[184].

2.7.1.2 Epidermis

The epidermis, or epidermal layer is composed of stratified keratinocytes, and is also known as squamous epithelium. Keratinocytes at the base of the epithelium known as basal cells (found in the stratum basale) can divide and start the process of maturation, as cells move towards the skin surface (Figure 2.7.1). As maturation occurs, the keratinocytes become flatter (squamous cells), increase their production of the protein keratin, and eventually de-nucleate and undergo a specialised form of programmed cell death called cornification. The epidermis also contains non-keratinocytes such as melanocytes which produces the pigment melanin. Langerhans cells and Merkel cells are also present but are not visible under H&E staining. This is the layer where skin cancers originate, with melanoma a result of abnormal proliferation of melanocytes, and BCC, SCC and others from the abnormal proliferation of keratinocytes.

2.7.1.3 Dermis and Hypodermis

The dermis is comprised primarily of an interwoven network of connective tissue (collagen and elastin) which provides the skin with strength and elasticity. It is subdivided into a thin superficial layer known as the papillary dermis, and the remaining deeper layer known as the reticular dermis. The papillary layer provides the metabolic needs of the skin with a plexus of bloods vessels that deliver oxygen and remove waste. The reticular layer additionally contains hair follicles, eccrine (sweat), sebaceous (oil) and apocrine glands as well as their respective ducts. Within the dermis also are fibroblasts and dendritic cells and often in the presence of disease, an increased number of mast cells, lymphocytes, granulocytes, and macrophages. Below the reticular dermis is the hypodermis, or subcutis, which consists mainly of adipose tissue. The thickness of this layer also varies depending on location, with the back, belly and waist being significantly thicker than other areas.

2.7.2 The Histopathology of Skin Cancer

The following section will introduce the major histopathological features used to distinguish common skin cancers (and precursor lesions) that are associated with sun exposure. As discussed in Section 2.2, domain expertise is needed to guide the questions that are trying to be solved, to verify the quality of the data informing the development of algorithms and scrutinise the medical relevance of the results. With respect to medical expertise, the information presented here is in no way a complete coverage of the topic and is derived from the relevant authoritative texts[185–187]. What follows is a brief introduction to the histological indications of several common and important skin cancers; actinic keratosis, intraepidermal carcinoma, squamous cell carcinoma, keratoacanthoma and basal cell carcinoma.

2.7.2.1 Actinic Keratosis (AK) or Solar Keratosis

Actinic (solar) keratoses are a common clinical observation in sun-exposed areas of older and light-skin individuals. It is associated with childhood UV exposure which manifests as sunburn yet is also regarded as an occupational and environmental disorder. It has been proposed that actinic keratoses are precursors of squamous cell carcinoma. Although it is commonly seen, biopsy is usually undertaken as there can be overlap in its appearance and BCC, IEC or SCC. Actinic keratosis is usually characterised by parakeratosis (the retention of nuclei in the stratum corneum) and a loss of the underlying granular layer. The epidermis is slightly thickened with a loss of the normal orderly stratified arrangement of keratinocytes and a reduced degree of maturation. The parakeratotic growth may also come to form a cutaneous horn.

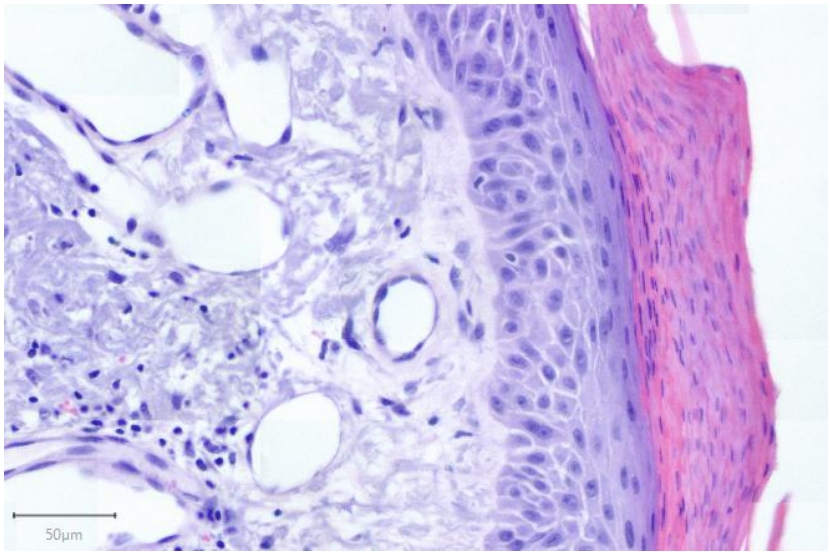

Figure 2.7.2 Example of actinic keratosis. Parakeratosis is present in the keratin layer and the granular layer has been loss. The epidermis is slightly disordered.

2.7.2.2 Intraepidermal Carcinoma (IEC)

Often referred to as squamous cell carcinoma *in situ*, IEC is characterised by the presence of full-thickness atypia compared to actinic keratosis which has partial thickness atypia. There is considerable disorder of cells within the epidermis as well as clumping of nuclei, dyskeratosis (keratinisation of individual cells within the stratum spinosum), highly atypical cells with large and hyperchromatic nuclei and pleomorphism (cells of varying size, shape and stain). The epidermis generally is acanthotic (thickened). The growth of atypical cells can extend down into the follicular infundibula (hair follicle shaft) and cause replacement of the follicular epithelium all the way to the sebaceous ducts. Between 3-5% of IECs develop into an invasive squamous cell carcinoma.

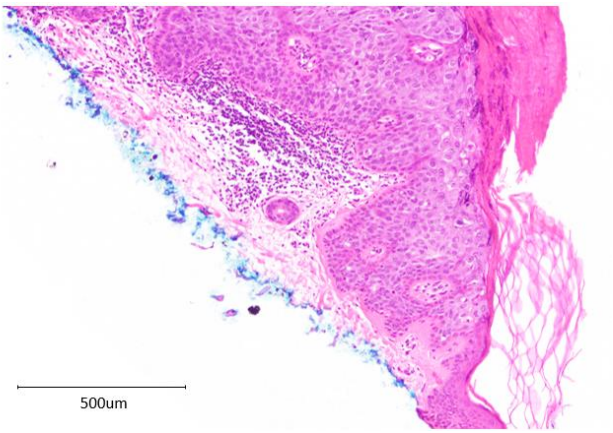

Figure 2.7.3 Example of intraepidermal carcinoma. The epidermis is thickened with full-thickness dysplasia. Hyperkeratosis is also present above the main lesion compared to the healthy basket weave keratin near the bottom right.

2.7.2.3 Squamous Cell Carcinoma (SCC)

Squamous cell carcinoma usually arises in actinic keratoses or sun-damaged skin, however the borderline between the classifications is somewhat arbitrary. This is supported by the lower interobserver concordance for these lesions than that for basal cell carcinoma. SCC consists of nests of squamous epithelial cells that have arisen from the epidermis and have invaded through the basement membrane running beneath the epithelium, extending into the dermis. The cytoplasm of the keratinocytes tends to be eosinophilic (pink) with large nuclei and individual cell keratinisation is often present. There is varying central keratinisation of the lesion and sometimes keratotic pearl formation depending on the degree of differentiation of the tumour. Differentiation is rather subjectively graded as ‘well’, ‘moderately’, and ‘poorly’, and is judged by the degree of anaplasia; deviations from normal in the context of nuclear pleomorphism, altered nuclear-cytoplasmic ratio, presence of nucleoli, high proliferation index with mitoses. There is often also a mild to moderate chronic inflammatory cell infiltrate at the periphery of the tumour. SCC occasionally invades the space surrounding a nerve (perineural invasion) which results in the cells spreading easily from the main body of the tumour to the surrounding local region. Therefore, the presence of perineural lymphocytes is an important clue for the presence of perineural invasion. SCC can also spread via lympho-vascular invasion, particularly in the head and neck region.

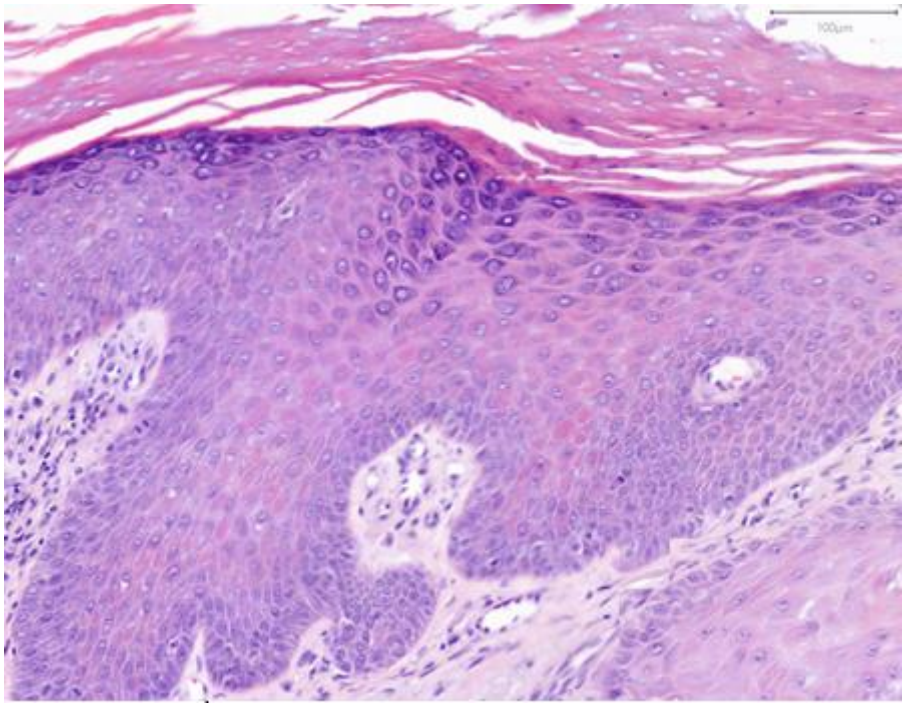

Figure 2.7.4 Example of squamous cell carcinoma. Mildly dysplastic solar keratosis overlies an isolated nest of enlarged, atypical keratinocytes, within the dermis (lower right).

2.7.2.4 Keratoacanthoma (KA)

Keratoacanthoma is increasingly being regarded as squamous cell carcinoma or a variant of it. However, some texts argue this as categorical error on both morphological and biological grounds and that it is a benign lesion[186]. Clinically it is characterised as a solitary, pink or flesh-coloured, dome-shaped nodule with a central keratin plug. It has a rapid growth period of 2-10 weeks which is then followed by a stationary period for similar duration. KAs then often spontaneously remit over a period between 8 and 50 weeks. Their persistence and size (1-2cm) means they can be locally destructive (particularly on the nose and eye) and so clinicians actively treat the lesion in the same manner as SCC.

Keratoacanthomas are an invaginating mass of keratinizing, well-differentiated squamous epithelium at the sides and bottom of the lesion formed around the central keratin plug. The plug is also seen to be buttressed by the lesion giving it a symmetrical appearance. The keratinocytes have a distinctive eosinophilic (pink) hue which can grow rapidly to become quite large, sometimes extending to the deep dermis and subcutis. They tend to form islands which mature central keratin pearls and associated neutrophils.

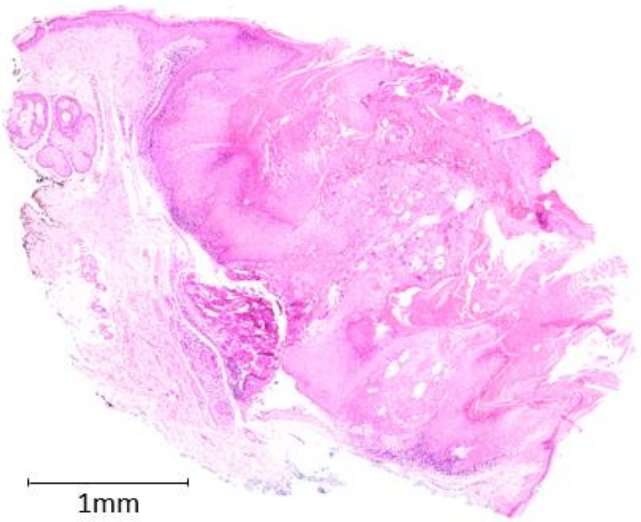

Figure 2.7.5 Example of keratoacanthoma. The lesion has a cup-shaped appearance consisting of a thickened layer of large keratinocytes with a large central keratin plug.

2.7.2.5 Basal Cell Carcinoma (BCC)

Basal cell carcinoma is the most common skin cancer and is found predominately in sun-exposed areas, particularly on light-skin individuals. Consequently, 80% of BCCs are found on the head and neck areas, with around 15% developing on the arms, shoulders, back or chest. BCC is generally slow growing but can be locally invasive and destructive if left untreated. Occasionally, BCC can spread to other regions as a result of perineural invasion, making resection of malignant tissue more difficult. Histologically it is characterised by basaloid cells proliferating from the basal layer of the epidermis or follicular infundibula within the dermis. Deviating from their normal path of maturation, basal cells continue to propagate in a form consisting mainly of the nucleus with little to no cytoplasm. Lobules of these cells can form and are frequently distinguished by the palisading of nuclei on the peripheral edges of the tumour. The lobules can sometimes separate from the stroma (surrounding connective tissue) and include the presence of mucin. The macroscopic growth pattern of BCC can be characterised into several variations including nodular, micronodular, cystic, basosquamous, keratotic, solid, superficial multifocal, and infiltrative sclerosing types. Their morphology sometimes can overlap with other lesions including SCC, IEC, trichoepithelioma, trichoblastoma, melanocytic nevus, fibroma, pseudolymphoma, melanoma and actinic keratosis.

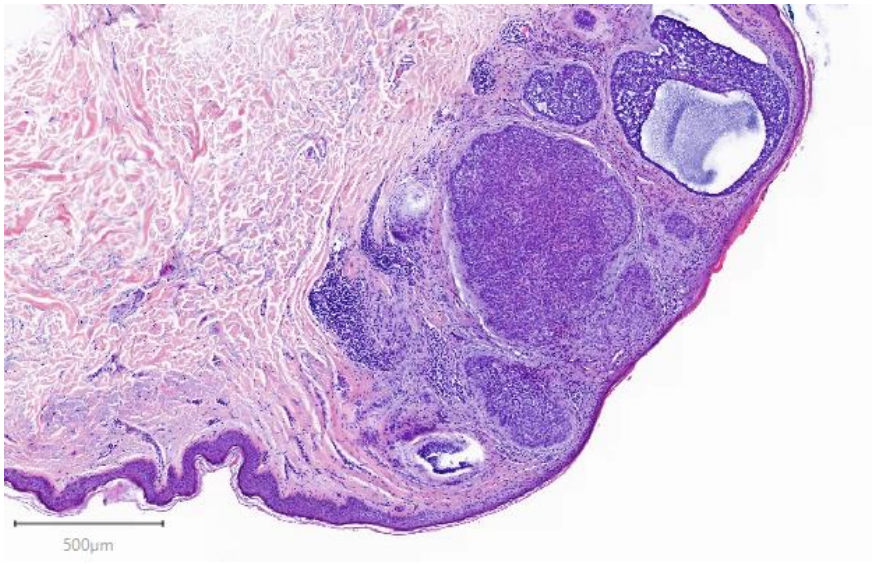

Figure 2.7.6 Example of basal cell carcinoma. Large nests of basaloid (blue cells) with peripheral palisading invade into the mid dermis. Lymphocytes (also blue) are present in the periphery.

2.7.3 Dermatopathology Workflow

In the process of diagnosing skin cancer, a pathologist generally sees multiple slides on which multiple tissue sections show different histomorphological features (Figure 2.7.7). Intact excision specimens are often oriented using blue and black ink, which gives the pathologist a high-level point of reference regarding the orientation of margins. The region of interest in the specimen is sliced into 3 mm thick sections and processed in various solvents, impregnated with paraffin, sliced into transparent 3µm sections, and stained to differentiate nuclei, cytoplasm, stroma and other structures. The slides are analysed to characterise the lesion which includes type, extent of invasion and completeness of excision including the distance to the surgical margin. A final diagnosis is given by integrating all the evidence across multiple slides while also considering the accompanying clinical notes and the site of the specimen *e.g.* face, neck, back, chest *etc.*

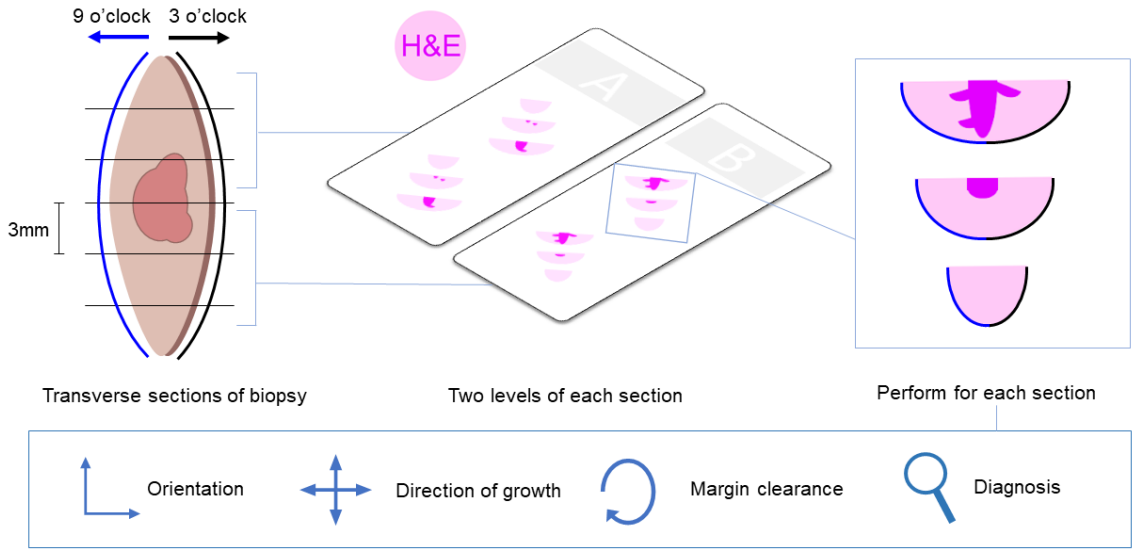

Figure 2.7.7 The dermatopathology workflow is aimed towards developing a single diagnosis and detailed characterisation which facilitates treatment and prognosis. A single biopsy is distributed across multiple slides which may have several sections, each shown at several levels (max 3). A pathologist then needs to orientate the section using ink markers, determine direction of growth, margin clearance, depth of invasion and other important features such as evidence of perineural invasion.

What is specifically included in a report is dependent upon the underlying diagnosis. The Royal College of Pathologists (RCP) published the minimum set criteria when reporting BCC[188]:

- Type of growth pattern
    - Nodular
    - Superficial
    - infiltrative / morphoeic
    - micronodular
    - others
- Type of differentiation
    - severely atypical
    - malignant squamous component present (basaosquamous)
- Perineural invasion (for infiltrative, micronodular and basosquamous)
- Excision margins
    - Distance to nearest peripheral and deep peripheral margins

Similarly, the RCP has published guidelines on the assessment and reporting of squamous cell carcinoma[189]. The requirements include subtyping *e.g.* basaloid-squamous cell carcinoma, grading (poorly, moderately, well differentiated), thickness/depth of the tumour, level of invasion (upper, mid, lower dermis), lymphovascular and perineural invasion, and surgical margin clearance.

The Royal College of Pathologists of Australasia (RCPA) has published requirements for the reporting of melanoma diagnoses[190]. The requirements include reporting the Breslow thickness (measured from the top of the granular layer of the epidermis to the deepest

invasive cell across the base of the tumour), involvement of the surgical margins, presence or absence of ulceration, the mitotic count per square millimetre of the invasive melanoma, the presence or absence of satellite metastatic tumour discontinuous from the primary tumour, lymphovascular invasion, the presence or absence of any desmoplastic melanoma component, perineural or intraneural invasion. In addition, a subtype should be recorded, belonging to one of the following:

- Superficial spreading melanoma (SSMM)
- Nodular melanoma (NM)
- Lentigo maligna melanoma (LMM)
- Acral-lentiginous melanoma
- Desmoplastic melanoma
- Melanoma arising from blue naevus
- Melanoma arising in a giant congenital naevus
- Melanoma of childhood
- Naevoid melanoma
- Persistent melanoma
- Melanoma, not otherwise classified

By understanding the workflow and reporting requirements of a pathologist better systems with more realistic contributions can be developed.

2.8 A Summary

The beginning of Section 2 explored the burgeoning interest in AI and the impressive progress that has been made in the last five years. With widespread talk of its adoption, there are major challenges that need to be addressed in this regard. However, these challenges are not insurmountable and by considering them in future research we can expect better outcomes as a result.

The area of medical AI is one such areas where risks exists, but when mitigated the benefits are enormous. This was evidenced in Section 2.2. which discussed the impressive results that have been produced across may areas of the healthcare sector. Arguably, results will continue to improve and because unlike other eras of artificial intelligence, the modern paradigm is driven by big data and deep learning algorithms. This new paradigm has demonstrated success in classifying various cancers, natural language processing of clinical records, gene targeting and protein structure prediction to name a few. Balancing the expectations of technologists and clinicians is however a challenge. For the technology to be successfully adopted clinicians need to be convinced, and perhaps even their patients, of its merits. Although it is largely accepted that AI will play some role in the future of medicine, it needs to be underpinned by clinical trials and demonstrated safety of the technology.

Section 2.3 demonstrated that part of the challenge of creating medical AI is having it conform to the multitude of expectations which define the boundaries of ethical medical practice. Having access to large datasets that are based on human data that covers the full diversity of end users is a necessary constraint. Further, models need to be interpretable by virtue of the fact that patients have a right to explanation for decisions that inform their

diagnosis, prognosis and treatment. By understanding how the technology itself works, we are better positioned to work towards these ends, and this was facilitated in Section 2.4.1.

The literature reviewed in Section 2.4.2 – 4 revealed that progress has continued in the areas of image classification, segmentation and generation. There is an increasing trend for more powerful models to have an increasing number of parameters. This is coupled with ever increasing dataset sizes. Although performance has improved, the memory costs may initially discourage uptake of better technology. However, there are still may ways in which popular models can be usefully applied, and their wide spread use means they are the best understood. Indeed, they are at the foundation of most interpretability methods which were discussed at length in Section 2.5, which will likely see them remain in high use. It was also mentioned that the major limitation of feature visualisation methods is they cannot be reliably reproduced. There is a trend however towards ever increasing realism in visualisation methods and this is likely a valuable avenue for future interpretability research, which is particularly motivated by the state-of-the-art results in Section 2.4.4.

It seems digital pathology (Section 2.6) has emerged at a time where machine learning can be a significant influence in its future direction. Recent work suggestions that histological diagnosis of many cancers is amenable to deep learning methods. Importantly, skin cancer diagnosis is an area where little work has been done. However, implementations of deep learning have so far lacked meaningful clinical application, few or any accompanying interpretability methods and have been limited in the learned context of the problem. To address this for future work, the full context of diagnosing the most common skin cancers was explored in Section 2.7. Importantly, major components of the dermatopathology workflow were identified to inform the clinical relevance of future applications.

It can be concluded that medical AI is indeed an area of great promise. When deep learning technology is considerately applied to the right problem, there is opportunity to improve the accuracy and efficiency of diagnosis. For the technology to be successfully adopted it also needs to be developed within a medical ethics framework and be informed and supported by medical experts. Skin cancer provides a unique opportunity to apply these ideas and set a new benchmark for machine learning in medicine, and this will serve as the basis for the thesis.

3 Aims and Significance

3.1 Motivation

Deep learning has a demonstrated ability to match human-level performance on large-scale image classifications problems *e.g.* ImageNet. Consequently, we know that these models have a capacity to capture knowledge across a diverse problem domain. Given the complexity of this problem there is reason to think that equally diverse problem domains such as the characterisation of cellular and tissue morphology, are similarly amenable to machine learning methods. Indeed, recent work in the field of digital pathology demonstrates that such features can be detected and used to classify healthy and disease states in histological images of various cancers. Importantly, the literature reveals that there is enormous opportunity to apply deep learning techniques to the histological diagnosis of skin cancer.

The application of deep learning to skin cancer diagnosis (and also cancer generally) has so far been done in ways which severely limit clinical translation. Firstly, previous work has performed either binary classification or binary segmentation, which is far removed from the real problem at hand. It is important to stress that a pathologist brings to bear a wealth of knowledge when assessing the presence or absence of a particular disease. They are not just able to identify diseased tissue but can place it in the context of the surrounding (and possibly unobserved) tissue. This context informs diagnosis, prognosis, treatment options *etc.* It also demonstrates that pathologists do more than just perform binary classification; they have knowledge of a diverse problem domain. It is clear then that the real-world problem of histological diagnosis of cancer is a multi-class problem, in terms of both the possible cancer types and the various healthy tissue types. Therefore, any meaningful application of deep learning to this problem must be from a multi-class perspective.

The benefit of working on skin cancer is that the full context of the problem is almost always presented to a pathologist. The stratified layers of the skin are easily identified along with associated organs such as glands, hair follicles, blood vessels *etc.* In this way, visual information is all that is required to understand a large contextual component of the problem *e.g.* orientation of the specimen, direction of growth, tissue abnormalities and resulting margin clearance. In this respect, skin cancer diagnosis is an ideal problem to apply deep learning because it is a diverse domain within a predictable structure. This predictable structure provides the opportunity to assist the pathologist in the routine tasks described, ideally improving the efficiency and accuracy with which they are performed. Concretely, this approach translates to meaningful clinical applications.

As a high-stakes decision domain, there is also a requirement to make the models interpretable. This has so far been an under-appreciated aspect of other work. Interpretably techniques are based on assessing the concordance of the network's and our own models of the world. It is likely that binary classification optimises a network to detect cancer features and ignore everything else (rather than the other way around). Instead, multi-class classification forces a network to learn the full visual context of the problem. In other words, every tissue type has a place in the model's representation, the same as a pathologist. This makes interpreting models and their decisions much easier because they are forced to not used hidden features, but ones which we ourselves use.

3.2 Overarching Hypothesis

Building upon what has been discussed, the overarching hypothesis of the project can be stated as follows:

> *Deep learning has the capacity to learn the full visual context of skin cancer diagnosis, and when applied in this way will lead to better and more general performance, be highly interpretable and translate to meaningful clinical applications.*

To constrain the scope, the focus will be primarily on non-melanoma skin cancer and the associated precursor morphologies as they constitute over 95% of skin cancer diagnoses in Australia[180,181]. This specifically includes actinic keratosis, intraepidermal carcinoma, squamous cell carcinoma, keratoacanthoma and basal cell carcinoma. Work on these cancers will take precedence over work done on melanoma.

3.3 Research Aims

The hypothesis contains three distinct components which form the major aims of this project.

Aim 1: To develop a system which can characterise the full-variety of tissue morphology present in the problem of non-melanoma skin cancer diagnosis.

Aim 2: To develop interpretability techniques which demonstrate that the model has learned the full context of the problem and make clear the capacity and the limitations of the model, and thereby fulfil our ethical obligations.

Aim: 3: To use the system to improve the efficiency and accuracy of routine pathologist tasks and thereby demonstrate the clinical usefulness of the technology.

3.4 Research Questions and Hypotheses

Beginning with the end in mind, Aim 3 significantly influences what approaches will be explored to answer Aims 1 and 2. Therefore, in Aim 3 I seek to assist the pathologist workflow by testing the feasibility of automating the following routine pathologist tasks:

- Perform whole tissue classification *e.g.* BCC, SCC, IEC, AK, KA, Non-Malignant
- Perform subtyping of BCC *e.g.* superficial, multi-focal, nodular, infiltrative
- Perform subtyping / grading of SCC *e.g.* poorly, moderately, well-differentiated
- Orientate a tissue section *e.g.* find superficial surface, identify orientation (3 o'clock)
- Measure surgical margin clearance
- Measure depth of invasion and thickness of tumour
- Identify lymphovascular and perineural invasion and measure clearance
- Identify mitoses and estimate mitotic rate
- Generate a body of evidence in the form of common morphological descriptions *e.g.* presence of pleomorphism, dysplasia, hyperchromatic nuclei, hyperkeratosis *etc.*

Given the variety of tasks, it is obvious that Aim 1 is a contingent step in achieving these ends. Therefore, in addressing Aim 1, the following questions will be explored:

- Can semantic segmentation methods be used to characterise whole tissue sections?

- How feasibly do multi-class semantic segmentation methods scale to real-world scenarios?
- Are unsupervised learning methods a suitable alternative to producing a full-context deep learning system?
- How do supervised and unsupervised methods compare in their performance?
- What aspects of Aim 3 are these methods capable of achieving, if any?

There is also a close link between Aim 1 and Aim 2. Therefore, approaches explored in Aim 1 will be assessed on their applicability to the following questions which address Aim 2:

- Are the methods developed inherently interpretable? *i.e.* do the outputs of the system provide an intuitive understanding of the system and decision processes?
- Do post-hoc interpretability methods sufficiently demonstrate that the problem domain has been learned?
- What aspects of the methods limit interpretability?

Importantly, the trend of interpretability is towards generating images of increasing realism. This naturally leads to the following hypothesis:

> *Learned representations which result in the ability to generate realistic images will enable classification and segmentation tasks to be performed in a naturally interpretable way.*

In essence high quality interpretable classification and realistic image generation are interchangeable. Therefore, it is important to consider the following questions:

- Can generative methods such as Variational Autoencoders or GANs produce realistic images across the full-context of the problem?
- Can these learned representations be used for classification or segmentation tasks?

Collectively these aims serve to test the overarching hypothesis; to determine the degree to which deep learning technology can assist in skin cancer diagnosis.

4 Summary of Work To Date

4.1 Characterisation and Classification of Non-Melanoma Skin Cancer

During my Master of Bioinformatics, I collected a dataset of 290 histological images of BCC, SCC and IEC. My thesis involved performing semantic segmentation of these images into nine tissue classes as well as whole image classification. I identified several limitations of this work; the segmentations were low resolution compared to the inputs (32x reduction), there were no ground-truth labels to assess accuracy of the segmentations, and several important tissue types were not included as classes. The first project of my PhD aimed to address these issues and demonstrate a machine learning approach that can be usefully applied in a clinical setting. Specifically, the aim was to generate high resolution semantic segmentations that characterise the image the same way a pathologist would. This addresses a short-coming in the literature in terms of interpretable models by forcing a network to learn the full context of the problem. The subsequent aim was to use these segmentations to perform whole image classification and also perform automated surgical margin assessment, the later providing a real-world application of the technology.

Tissue Segmentation

I hand-annotated the 290 images into twelve meaningful tissues types: Background (BKG), keratin (KER), epidermis (EPI), papillary dermis (PAP), reticular dermis (RET), hypodermis (HYP), glands (GLD), hair follicles (FOL), inflammation (INF), and BCC, SCC and IEC. Semantic segmentation was performed by training a ResNet50 with a U-Net-like decoder (Figure 4.1), which enabled pixel-level classification of images at the original input resolution. To understand how different image resolutions affect performance, three networks were trained with the dataset downscaled by factors 2, 5 and 10. The original resolution (1x) data contained information well above what was needed for a human to classify and so was not used for analysis.

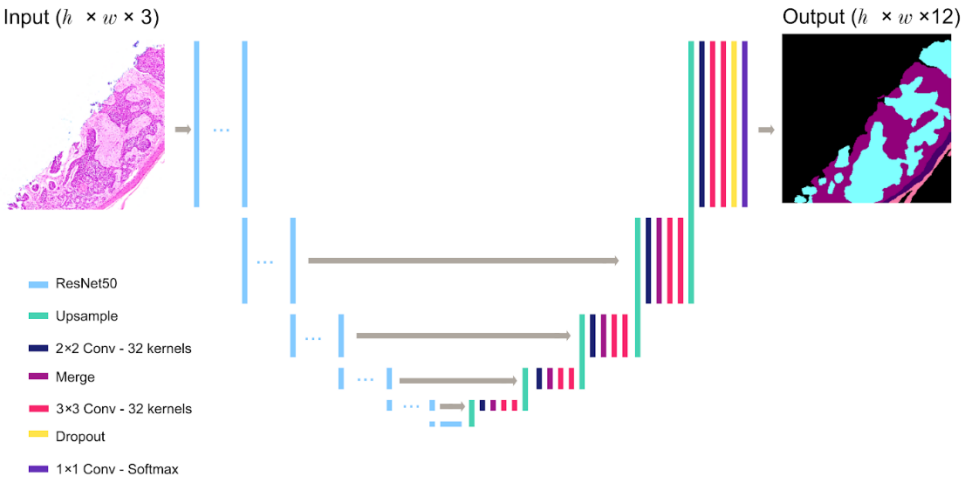

Figure 4.1: A U-Net like architecture attached to ResNet50. The final layer is a h × w × 12 tensor which used in combination with an argmax function along the last axis produces the 12-class segmentation seen above. A small variation on the original U-Net is that Upsampling and 2×2 Convolution layers were used in place of single Transposed Convolution layers.

The networks were trained on patches from the original images which are too large to process singly. Comparative performance of the three networks is shown in Table 4.1. The best performance was achieved on the 10x dataset with a weighted pixel-accuracy of 74.2%. Overall pixel accuracy was approximately 85% for both the 10x and 5x datasets. Whole image segmentation was performed using a sliding-window approach where a segmentation canvas was generated iteratively (Figure 4.2). Examples of the quality of the segmentations are shown in Figure 4.3.

Table 4.1: Loss, accuracy (Acc.) and weighted (W.) accuracy for base and fine-tuned (FT) models. Values are the average from three replicates ± the standard deviation. Weighted accuracy increased as the image size decreased. The resulting decrease in loss from fine-tuning doesn't translate late to a comparable increase in accuracy (~1%).

Dataset	Base Loss	Base Acc.	Base Weighted Acc.	FT Loss	FT Acc.	FT Weighted Acc.
10x	0.9803 ± 0.0359	0.8520 ± 0.0065	0.7373 ± 0.0065	0.9379 ± 0.01724	0.8526 ± 0.0012	**0.7420** ± 0.0054
5x	0.8747 ± 0.0283	0.8472 ± 0.0078	0.7299 ± 0.0056	0.8685 ± 0.0083	**0.8571** ± 0.0005	0.7396 ± 0.0004
2x	0.8612 ± 0.0030	0.8118 ± 0.0020	0.7254 ± 0.0017	0.8509 ± 0.0059	0.8120 ± 0.0013	0.7261 ± 0.0013

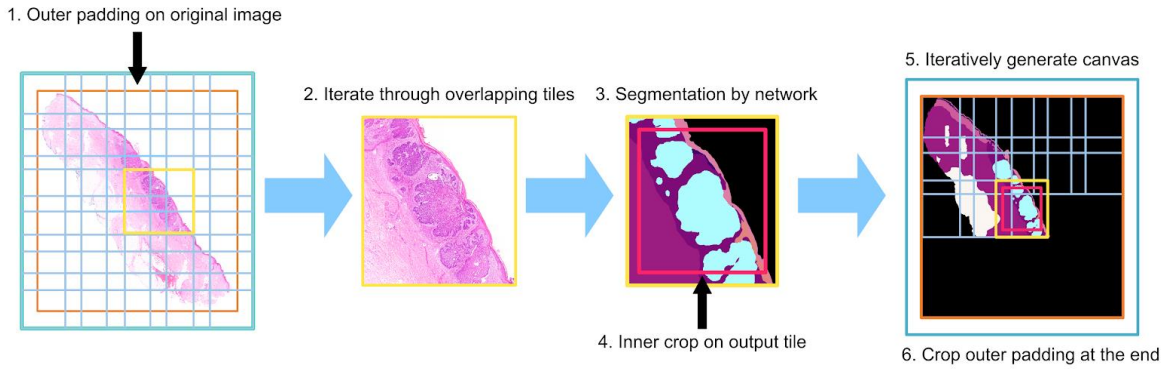

Figure 4.2: The Whole Image Segmentation pipeline. To minimise border effects and improve tile overlap, the image is first padded with white/background pixels. A tile is then fed into the network, where an inner crop is made to further minimize border effects. The segmented tile is added to the canvas which at the end is cropped to the original image size.

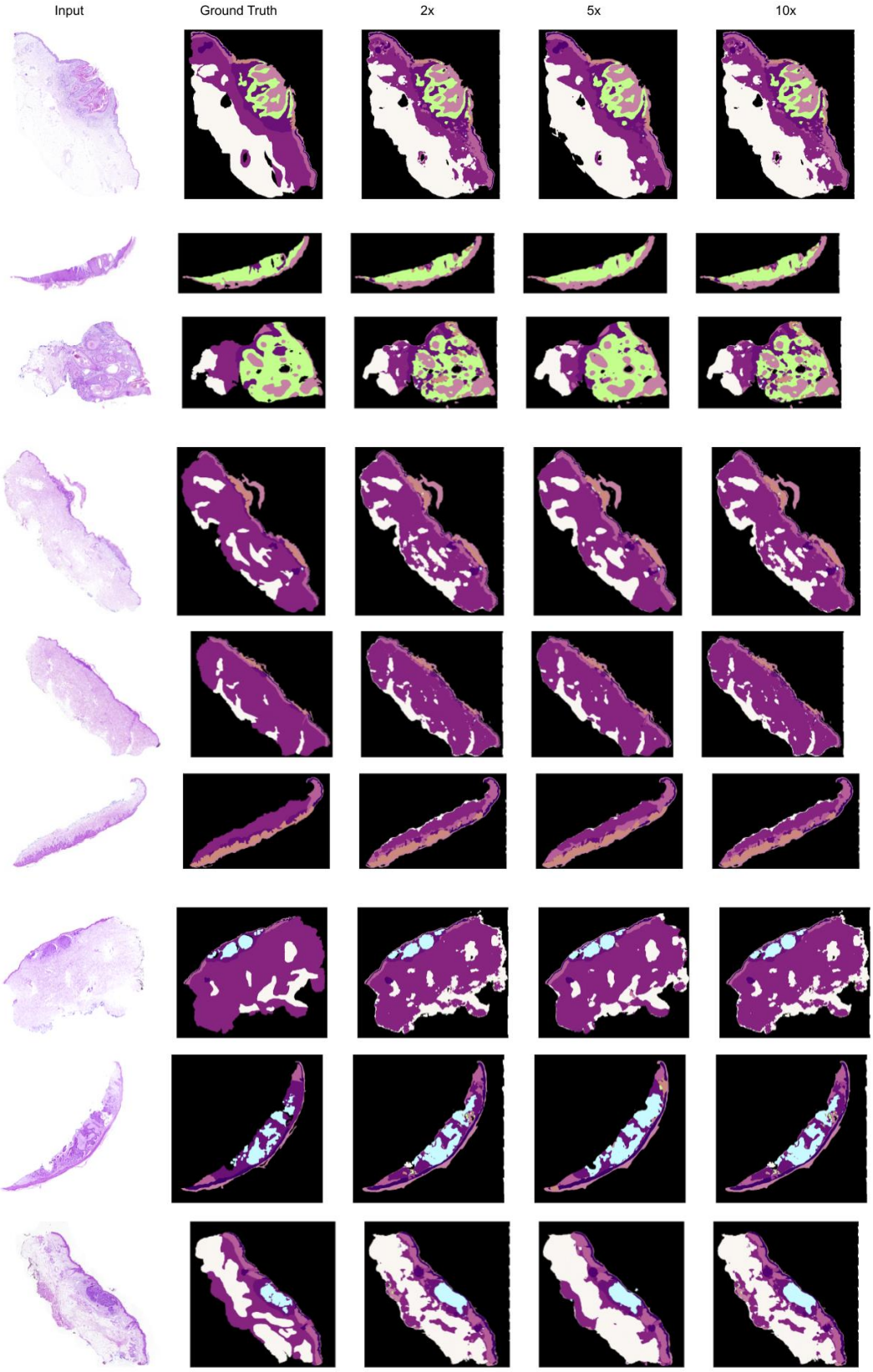

Figure 4.3: Whole image segmentations for representative cases of SCC (1. Excision, 2. Shave, 3. Punch), IEC (4. Excision, 5. Excision, 6. Shave) and BCC (7. Excision, 8. Shave, 9. Excision). SCC, IEC and BCC are labelled with green, orange and blue pixels respectively. For other colours see Figure A1.

Whole Image Classification

Whole image classification was performed using the raw output from the segmentation network to train another CNN to classify images as either BCC, SCC, IEC or healthy (Figure 4.4). This was performed for only the 10x dataset due to memory requirements of larger inputs *e.g.* $h \times w \times 12$ pixels where h and w are the original image size. The classification results are shown in Figure 4.5, revealing an overall accuracy **96.57%** with zero false negatives for the four-class problem, and **99.1%** for healthy versus cancer.

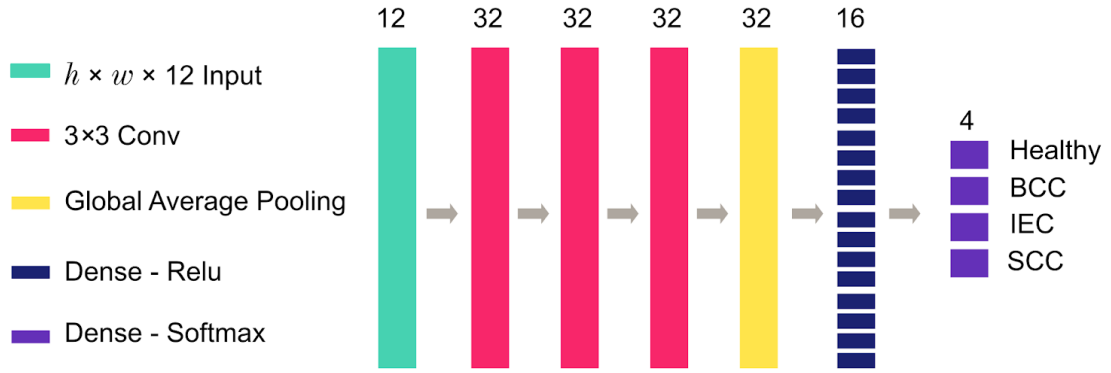

Figure 4.4: The whole image classification network. The h × w × 12 probability maps are fed into 3 consecutive 3x3 convolutional layers, each with 32 kernels. Global Average Pooling is used in contrast to Global Max Pooling, with the idea being that false positive cancer classes will be averaged out rather than signalled as of high importance. Finally, two Dense layers are used to make the classification.

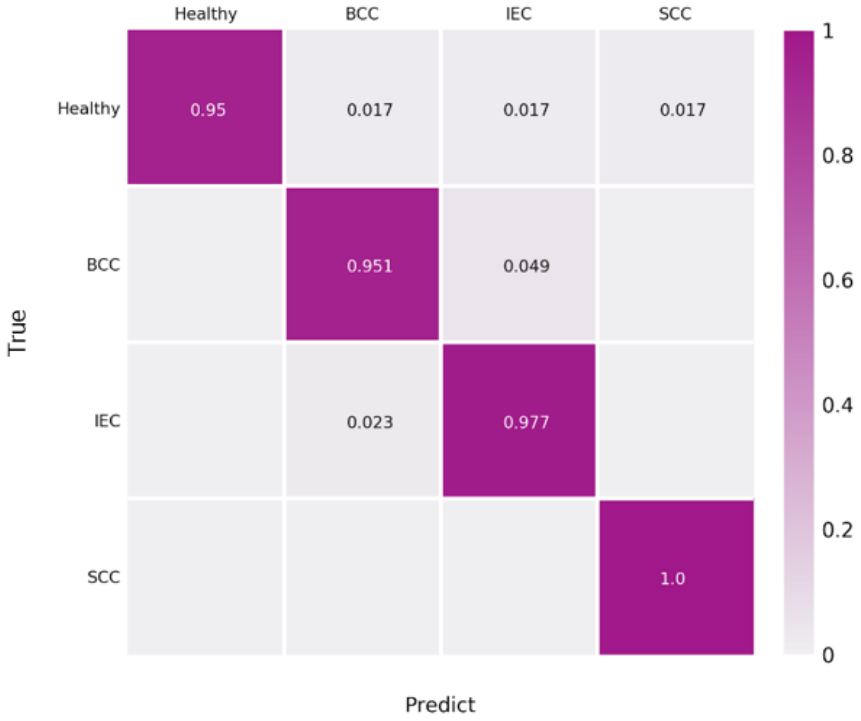

Figure 4.5: Confusion matrix for whole image classification using the 12-class probability matrices. Values read across columns for each row describe Recall *i.e.* true positive rate.

Model Uncertainty

To provide insight into the confidence the network had in its segmentations, we performed temperature-scaling to calibrate the output (Figure 4.6). With a good correspondence between confidence and accuracy we then generated confidence maps, showing which parts of the image the network found challenging (Figure 4.7). Areas of uncertainty compared to the ground truth correspond to FOL and INF classes which were previously identified as difficult classes based on their pixel level accuracies (Figure A2).

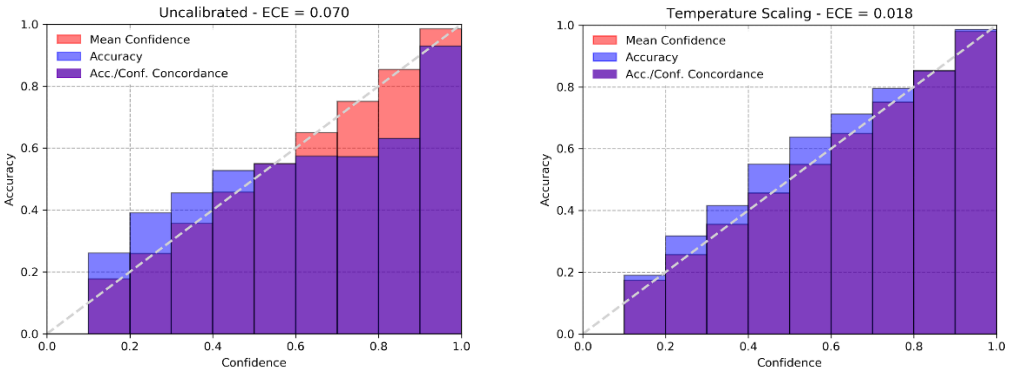

Figure 4.7: Calibration plots for the10x segmentation network. There should be a strong concordance between confidence and accuracy. Left) Uncalibrated network shows over and under-confidence. Right) A well-calibrated network after temperature scaling.

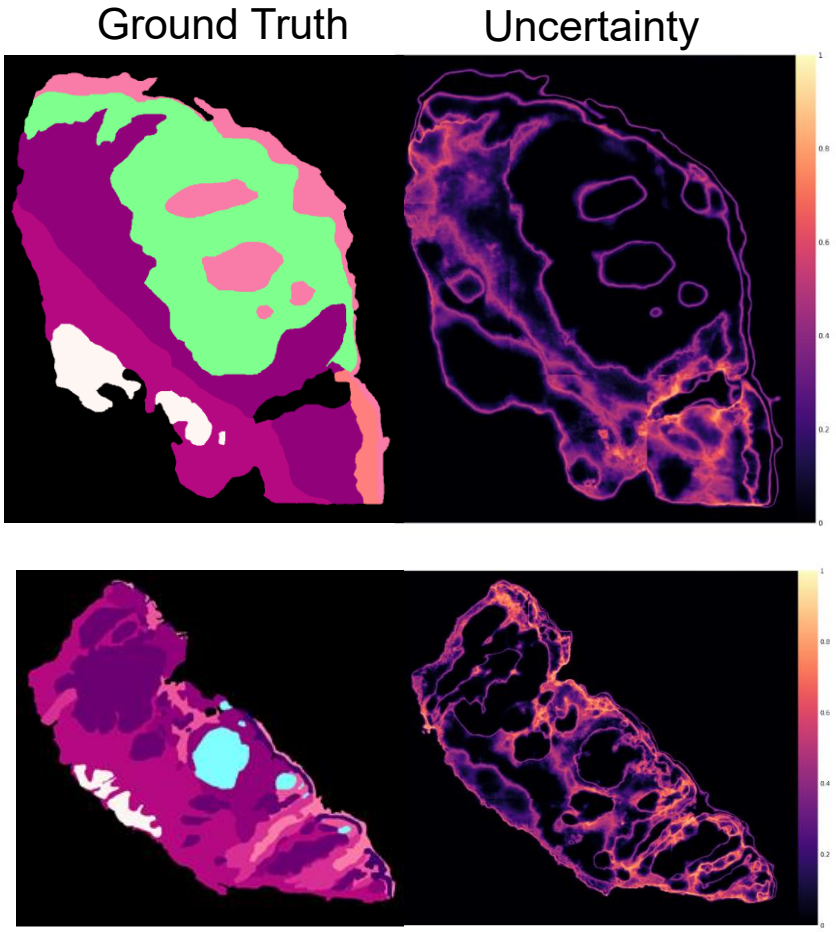

Figure 4.7: Image ground truth (Left) with prediction uncertainty (Right) visualised as the distance between 1 and the highest probability for each pixel. For class colours see Figure A1. The assumption is made that all images the network sees are within the training distribution and so confidence values depict the actual performance of the network.

Clinical Applications

Being able to produce high-quality segmentations provided an opportunity to assist in measuring surgical margin clearance. Figure 4.8 shows the results from using classical image processing algorithms to detect and measure margins from the segmentations. The information from the cancer region segmentation, whole image classification, and margin clearance can be presented as a useful summary to the pathologist in the context of the original histology image.

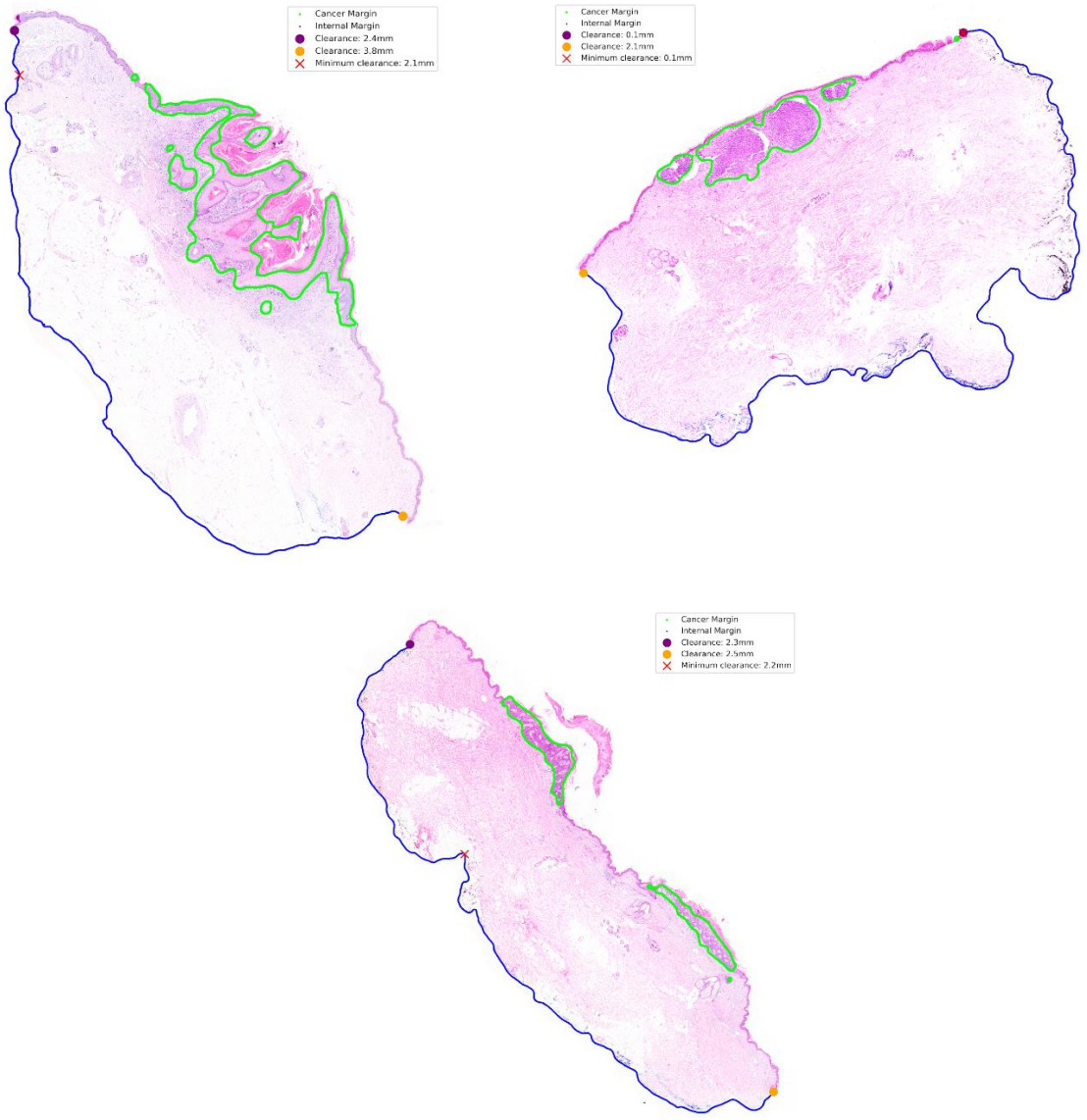

Figure 4.8: Examples of SCC (top left) BCC (top right) and IEC (bottom) with surgical margin clearance automatically measured in addition to regions of interest circumscribed.

In this work it has been demonstrated that high-quality segmentations can be utilised to perform reliable classification as well as provide useful tasks that could improve the efficacy of a pathologist. Efforts have also been made to make this model interpretable by forcing it to learn the full context of the problem and additionally generate confidence outputs to understand points of weakness. Further interpretability could be achieved by measuring the contribution of different tissue types to the whole image classification using a Grad-CAM/TCAV-like method. Thinking towards the future, there are issues with the scalability of this method. This small dataset under-represents the population of skin samples and so these results need to be validated at scale. However, it is not practical to hand-annotate

the thousands of images necessary to achieve this given the current research environment. On top of this, coarse labelling into BCC, SCC and IEC misses the nuanced variability in features which are important in other pathologist tasks such as subtyping. This information is likely captured by the trained networks however it remains elusive to us as classifying those features is not an explicit objective. Being able to present evidence of malignancy in terms of prescribed cellular morphologies *e.g.* hyperchromatism, pleomorphic, parakeratosis *etc.* might be a more robust way to characterise tissue generally as it scales to include other applicable skin diseases.

4.2 Investigating Interpretability Methods

The state-of-the-art interpretability methods[128,132,133] discussed in Section 2.5.3 have been criticised because the methods do not generalise to other networks. In order for these methods to be applied to other problem domains they must be replicated, and their applicability demonstrated. My work on this problem served as a learning process to develop my skills and knowledge base, as well as an attempt to replicate it. Specifically, I attempted to replicate some of the key methods involving feature visualisation and how they can be used to interpret the features that a network uses for classification. This lead to further visualisation techniques which have subsequently informed the direction of the project. In place of the seldom used InceptionV1 network I utilised the widely used VGG16 network.

Activation Maximisation

Using gradient decent I optimised input images to maximally activate the pooling layers of the network (Figure 4.9). The results are comparable to those produced by TensorFlow (https://storage.googleapis.com/deepdream/visualz/vgg16/conv5_3.html). The visualisations reveal that the network learns a hierarchy of features which increase in complexity and abstraction with depth. When looking at the filters individually it is interesting to note that some are optimised to detect salient features as discussed in Section 2.7.3. Conversely, many of the filters in the last layer (512 in total) do not represent human meaningful concepts by themselves. Thus, we want to know how these filters interact in the context of being activated by real images.

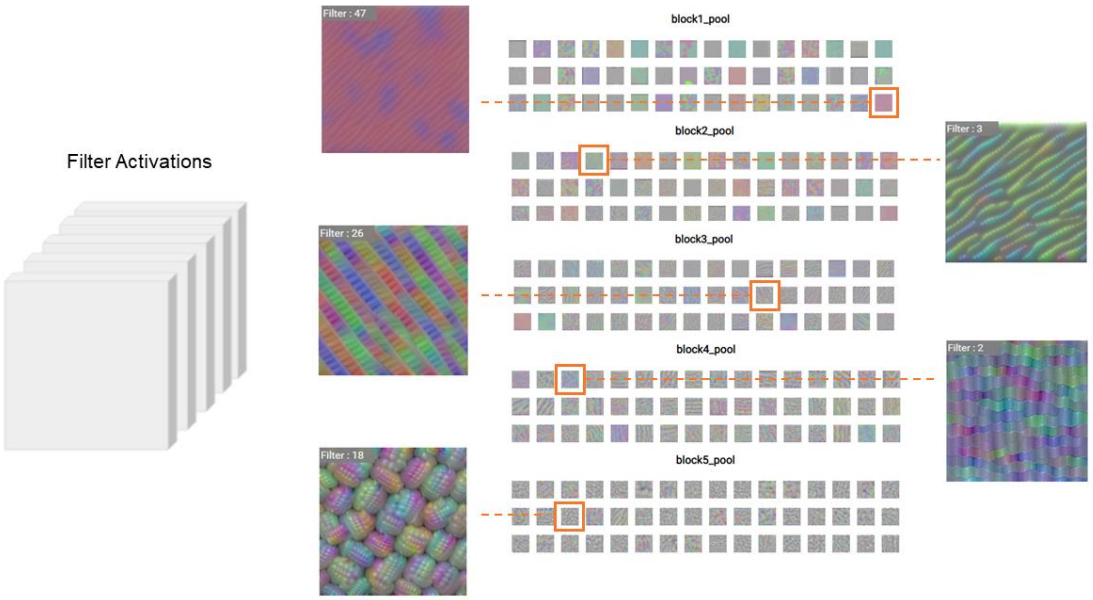

Figure 4.9: Feature visualisation was performed for all the pooling layers in VGG16. The above shows the first 48 filters for each block. The total number of filters for each layer were 64, 128, 256, 512 and 512. The visualisations show increasing complexity and abstraction. However, they all exhibit the classic rainbow colour scheme. An interactive version of this figure can be found at https://smthomas-sci.github.io/FeatureVisualisation/

For an input image of size $224 \times 224 \times 3$ the block5_pool layer of VGG16 outputs a $7 \times 7 \times 512$ tensor, A. To visualise what features are being detected at each region, *e.g.* $A_{i,j}$, the top 5 channel activations are selected. Gradient descent is then performed to visualise the network's representation of that region by maximising the weighted activation of those filters (Figure 4.10). A correspondence between what we see and what the network is optimised to detect serves as the basis for this interpretability method. The extent to which the feature was detected can be gauged by looking at the relative activations of each region. In the figure below it can be seen that the dog is the source of the

activation, likely because the network has learned to ignore background information common to most images.

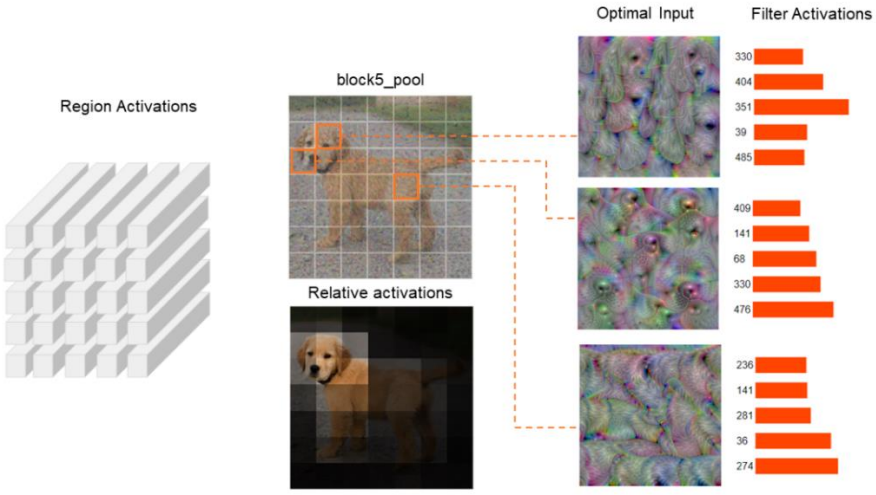

Figure 4.10: Feature visualisations that maximise the same filters that a given image does. The activations can be looked at for each separate region to visualise the interaction of features across channels. This shows that where we see a dog snout, dog ears or fur, the network detects the same. An interactive version of this figure can be found at https://smthomas-sci.github.io/FeatureVisualisation/

There may be features which extend beyond a single region and so instead we can look for combinations of spatial relationships and channel-wise features, which can be considered group activations. This can be done via non-negative matrix factorisation (Figure 4.11), where we optimise two matrices, with a common dimension k, so that the dot product approximates A. The number of concepts is determined by k and is chosen arbitrarily. In Figure 4.11, $k = 4$ which corresponds to the concepts of ears, face, fur and fur. In this case, the ideal number of concepts could be captured with $k = 3$. However, there is no way to know in advance how many independent concepts are present in a given image. Therefore, it is not clear how useful this technique is in practice.

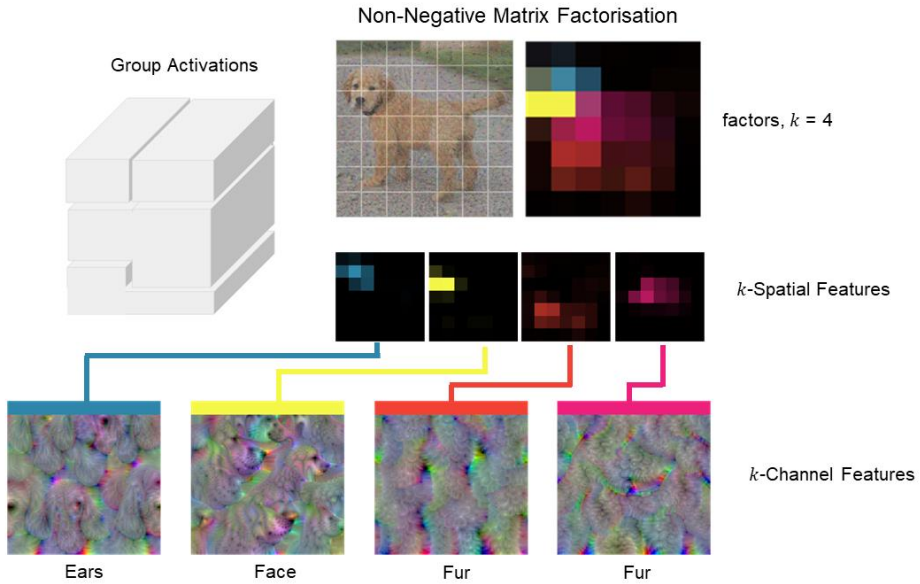

Figure 4.11: Feature visualisation performed on group activations for the block5_pool layer of VGG16, A. Groups are found through non-negative matrix factorisation which optimises two matrices (with a similar dimension k) whose dot product approximates A. A is then multiplied by the contribution of each group both spatially and channel-wise which can be visualised.

Having identified the concepts that the network has learned and that they do in fact correspond to human concepts, the state-of-the-art methods have further combined them with attribution techniques, similar to Grad-CAM and TCAV described in Section 2.4.2. However, that component is easily replicated and so was not performed in this work.

Optimising for Realistic Colours

A major limitation of the reproduced visualisations is that although they depict meaningful concepts, they lack the correct colour values. The reason for this is not clear. It might be due to non-robust features (likely), the multifaceted nature of most concepts, or the initialisation of the optimisation process[191]. In the later cases, taking ten images of a class and computing the average image to start the optimisation process produces colours that are more realistic (Figure 4.12). However, this introduces information outside the knowledge contained in the weights. This introduces the problem of regularisation – how much influence do we want to have on the quality of the visualisations? The start-of-the-art visualisations require regularisation such as jitter, blurring and good initialisation, as do my own visualisations. Therefore, it is common practice to regularise the generated image. However, to understand what the network has actually learned it is important to limit the amount of external bias introduced.

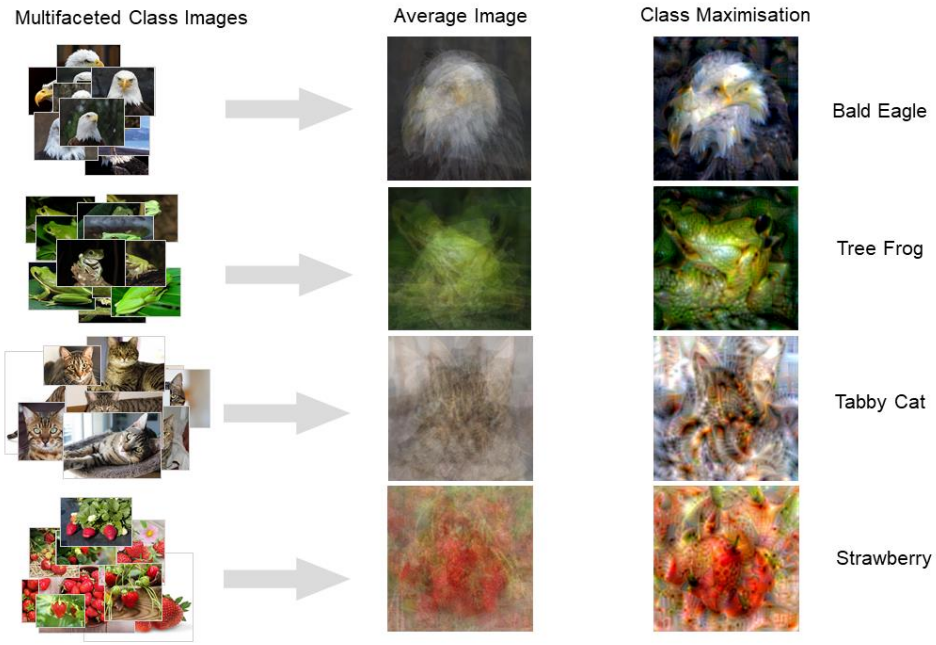

Figure 4.12: Taking ten images for a given class and averaging them produces a useful initialisation image. The optimisation process starts in a location close to where most images of this class reside. Performing class maximation results in an image which contains the features as well as the colours of that class. However, this biases the process because a significant amount of information is introduced by the initialisation. Although it was not done, a sanity check of optimising for a different class to the initialisation image could reveal a significant fault in this method.

Taking inspiration from Mahendran and Vadaldi[125] and Nguyen *et al*[192] I hypothesised that since hidden layers contain sufficient information to reconstruct the original image, a network that was trained to invert the features of real images would additionally be well-regularised to generate realistic images. An inversion network (Figure 4.13 – Step 1) was trained to invert the features from block2_pool ($56 \times 56 \times 128$) of VGG16 using the Cats & Dogs dataset[193]. This dataset contains 25,000 images of cats and dogs and contains a large variety of natural image settings. After training, the inversion network is locked and used to generate images that maximise an activation objective (Figure 4.13 – Step 2). The results of this can be seen in Figure 4.14 with comparisons between the original method and the state-of-the-art method. The visualisations from the inversion network not only show the features but do so with colours that correspond to what is in the original image, unlike

the other two methods. Importantly, no additionally regularisation tricks were used (unlike the other two methods). Comparable or higher quality visualisations work immediately via stochastic gradient descent. Further, the inversion network can produce similar visualisations when connected to the VGG19 network for class maximisation (Figure 4.15 & 4.16), albeit of lesser quality.

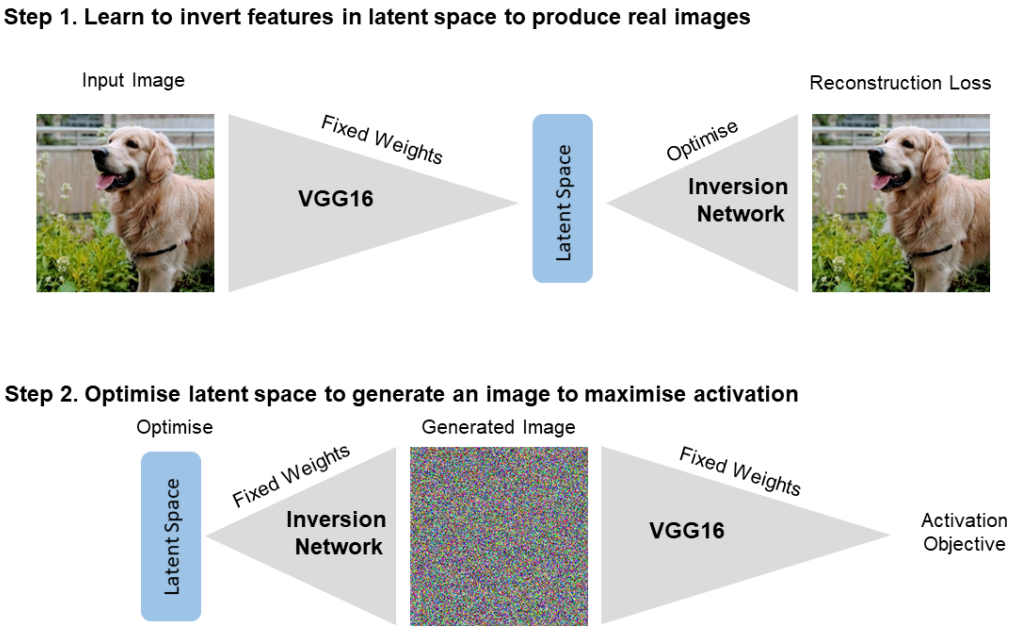

Figure 4.13: Step 1) An inversion network was trained to invert the features from the block2_pool of VGG16 to reproduce the original image. Training was performed on the Cats and Dogs data set containing 25,000 images in natural image scenes. Step 2) The inversion network was then used to generate images using the block2_pool layer as a latent space. Gradient descent was then used to optimise the latent space to generate an image which maximised a given activation objective.

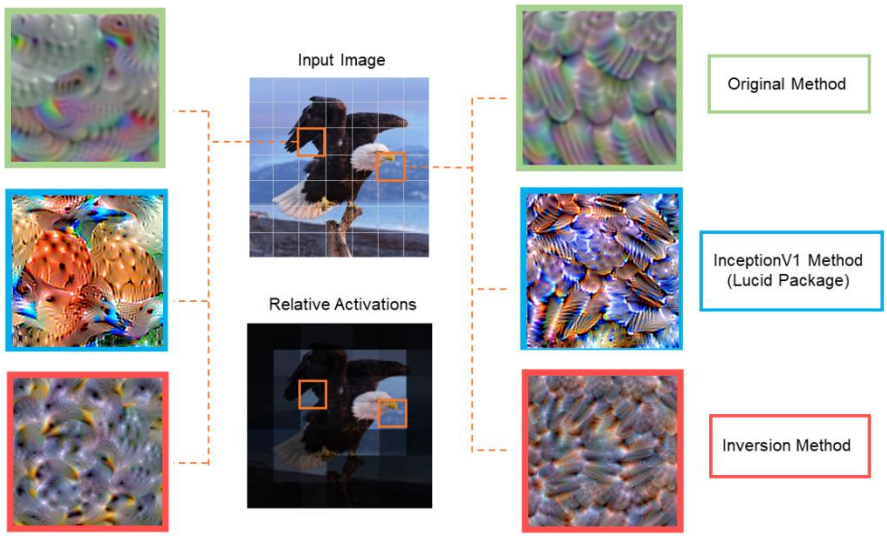

Figure 4.14: Feature visualisations for spatial regions using three different methods. The original method shows mostly correct features but incorrect colours. The Lucid library produces correct features but incorrect colours. The inversion network produces mostly correct features and much more realistic colours, all without other regularisation tricks. An interactive version of this figure can be found at https://smthomas-sci.github.io/FeatureVisualisation/InversionNetwork/

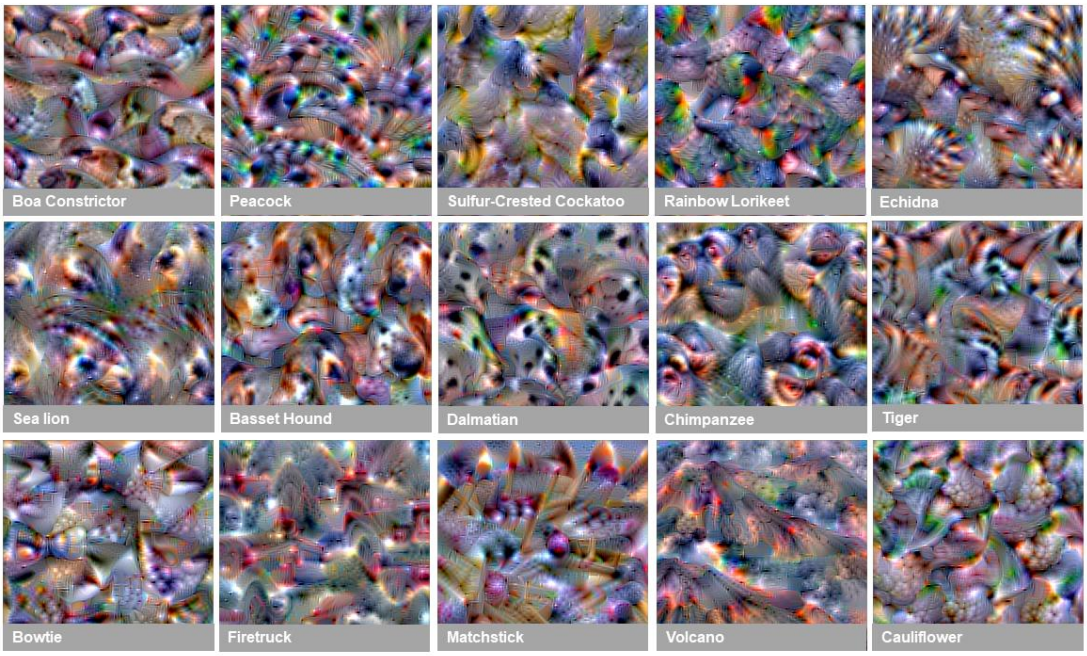

Figure 4.15: Feature visualisation of class activations using the inversion network with VGG16. In this case, the generated images represent evidence for a specific classification. In this way we are asking, what image maximises the prediction of peacock? The features and colours that are most representative of the class are visualised and largely correspond to human concepts associated with the class. Others are more abstract or of poorer quality such as Bowtie and Fire truck.

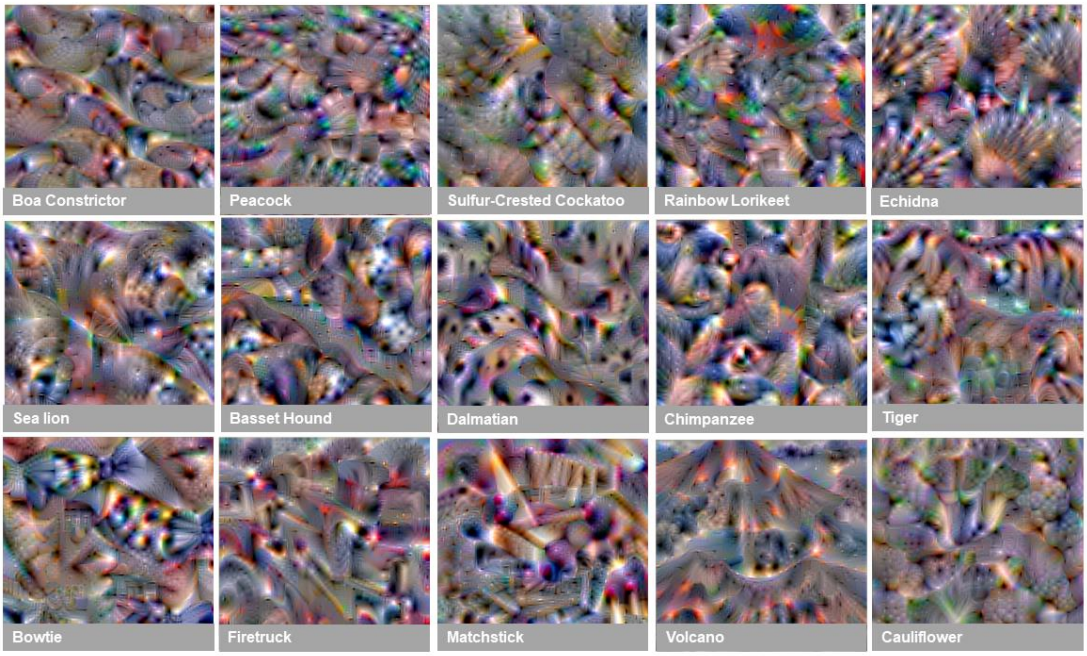

Figure 4.16: Feature visualisation of class activations using the inversion network with VGG19. In comparison to Figure 4.15, the inversion network appears to generate images with colours that correspond to the class but are of lesser quality than for VGG16. However, the edges and textures are generally representative of the concepts associated with each class which is normally achieved with regularisation tricks.

Latent Space Structure

The feature inversion network seems to be a promising alternative to producing feature visualisations. Of course, there are many parallels to the *structured latent space* that are characteristic of generative methods. Interestingly, the recent work on robust features suggests that robust networks are already well-regularised to generate high quality visualisations[131] and so this hints at the idea that their latent space is much more structured than non-robust models. As discussed in Section 2.4.4, generative methods such as Style-GAN[111] can place test images in a latent space and regenerate them with high fidelity, demonstrating the GAN has learned the correct representation. A demonstration of how this technique can be used to visualise what the VGG16 network is seeing is shown in Figure 4.17. Compared to the optimising pixels individually, the inversion network makes finding a realistic image a much easier optimisation problem and produces more recognisable results (though still not realistic). There is obviously more structure in the latent space than in pixel space, where a small shift can result in unnatural looking images. From this perspective it is clear why it is an easier optimisation problem.

Finding images in the latent space of network and then generating them is useful technique to validate that a model has meaningful representations. This further emphasises the trend of increasing the quality of visualisations, suggesting that state-of-the-art generative methods could be a useful avenue to explore. Importantly, the investigation into interpretability methods was successful in its aims. The state-of-the-art interpretability interfaces were successfully reproduced and indicate practical relevance to the problem of skin cancer.

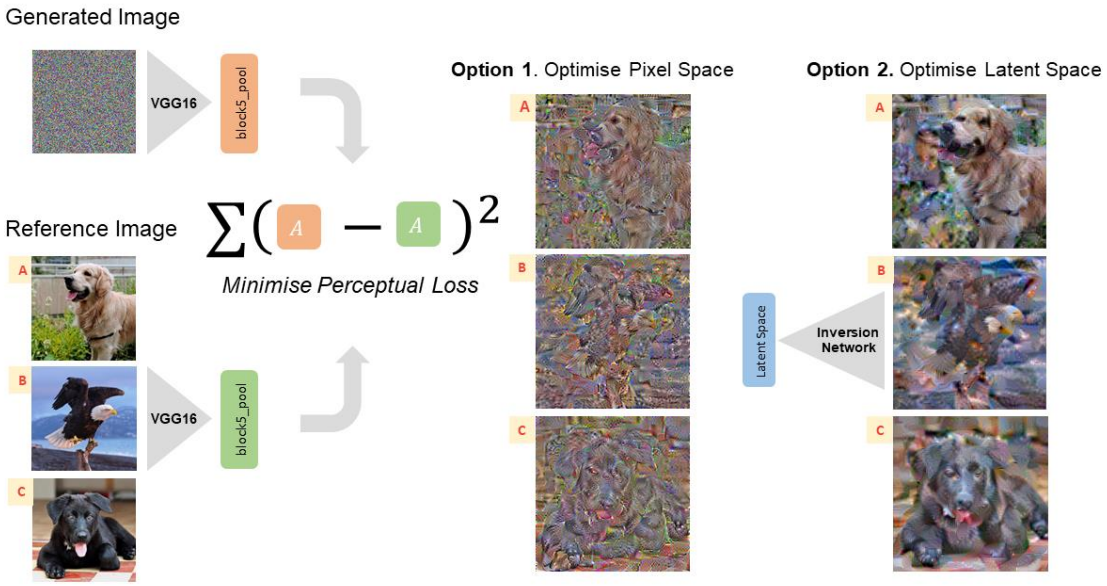

Figure 4.17: The perceptual loss can be used to find images which cause a similar activation to a reference image. Gradient descent can be used to minimise the sum of the squared differences between the activations. This is a difficult optimisation process in image space (Option 1). However, optimising the latent space of the inversion network (Option 2) leads to more realistic images which more closely resemble the reference image. This indicates that the space within a network for a given activation is representative of an image with specific features. *i.e.* it doesn't represent just any image in the same place.

4.3 Digital Pathology Interface

Many publications in the field of digital pathology and machine learning include a comparison of their algorithms with human performance. They have additionally relied upon expert annotations or analysis to help understand the problem or label their data. Being able to collect expert annotations is valuable, however, sharing data and providing interfaces by which they can do so poses big challenges. Although it is not a central aim of the project, to address a possible need to interact with and share data with pathologists, I developed a prototype web-interface, available at https://dermopath.uqcloud.net/. This utilises the OpenSlide software to serve gigabyte size WSI data, through an interface built using HTML, CSS and Javascript. It is similar to Google Maps where it utilises image pyramids to load local image data at high resolution. It further allows simple annotation, measurement and saving of segmentations (Figure 4.18).

The interface is very bare and at this stage is not suitable for general use. However, it may be useful in later stages of the project to securely share data with pathologists and collect annotations. Alternatively, it could simply serve as a prototype interface showcasing how the developed machine learning system could work in practice.

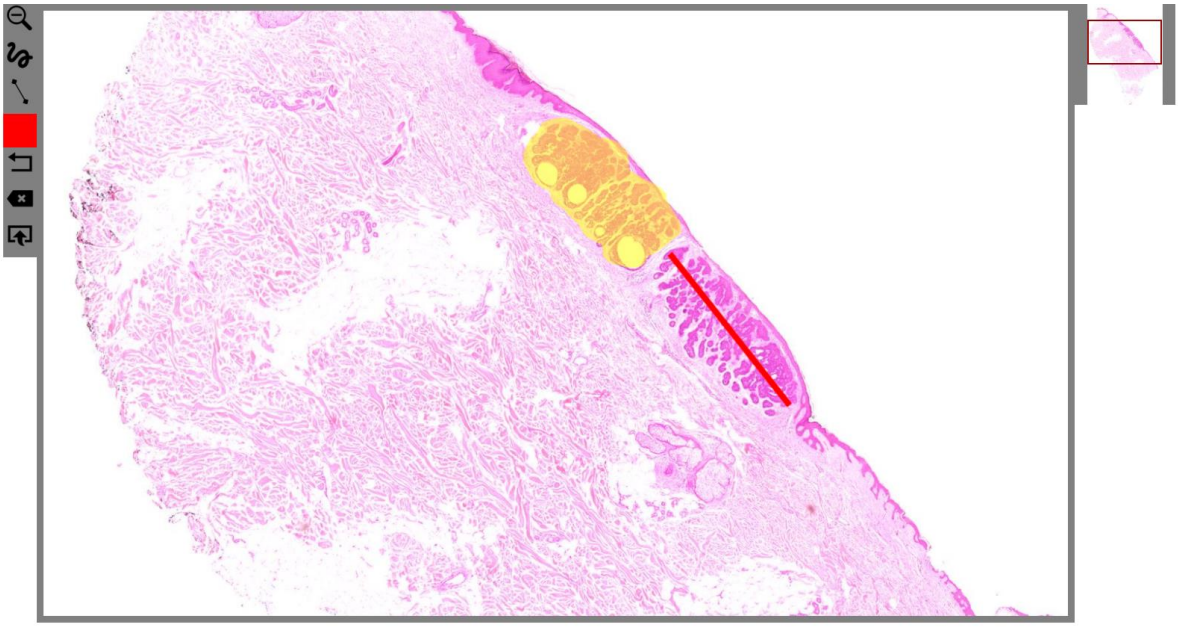

Figure 4.18: A prototype interface that could be used to share data with pathologists and collect annotations. It could further be used to demonstrate the use of the developed machine learning system in a clinical setting. The interface is available at https://dermopath.uqcloud.net .

5 Research Plan and Timeline

The table below outlines a prospective time for the project. The objectives are set out to complete and submit the thesis by 3 years.

	2019		2020				2021			
	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Overseas Conferences										
Local & National Conferences										
Publication of Current Work										
Data Collection										
Interpretability Problem • Unsupervised learning methods to learn the full-context of problem and demonstrate interpretability through visualisation										
Scalable Clinical Applications • Replicate and extend clinical tools developed in M1, using unsupervised models • Convert unsupervised models into segmentation networks to include: o Perineural invasion o Lymphovascular invasion • Convert unsupervised models into classification networks to include: o Subtyping o Grading										
M2 – Report, Interview & 2x Presentations		3MT		Main						
M3 – Report & Interview									Buffer	
Thesis writing										
Thesis Submission										
	Year 2				Year 3				Half	

6 Skills and Resources

6.1 Deep Learning and Image Analysis

In the last year I have developed a solid foundation in deep learning for image classification and segmentation. Further, I have implemented or utilised interpretability methods involving dimensionality reduction, attribution and feature visualisation techniques. I have built a good intuition for the suitability of these techniques to most 2D imaging problems across medical and non-medical domains. I have also spent considerable time working with the ImageJ software; automating workflows and creating ground truth segmentations for my datasets.

I have introductory knowledge of unsupervised generative models such as variational autoencoders and GANs but have not implemented these algorithms to any meaningful degree. Although I am confident in how they can be utilised for this project, it will be necessary to spend time developing the confidence to apply them suitably towards Aims 1 and 2. This won't require any further hardware resources, and I already have access to the necessary learning material online.

For prototyping models, I have used my desktop computer with a high memory GPU card. This has been sufficient (and expected to continue to be), for developing techniques and training small models *etc*. High performance computing is a necessary component of the project and I have heavily relied on the Wiener GPU cluster. This resource will continue to be central to the project.

When working with WSI data I have utilised the opensource framework called OpenSlide. Connected to Python, I have created pipelines to automatically access, segment, store and managed image files that are too big for memory. Unfortunately, it lacks a meaningful annotation interface, but I have since discovered QuPath, an opensource annotation software for digital pathology images. However, sharing images is the biggest challenge in this area given their size. Therefore, it is likely that if I am to securely share slides and annotations it is probably best done via a web interface (Section 4.3 – Figure 4.18). However, at this stage it is not clear if this will be needed given the aims of the project.

6.2 Skin Cancer

The dataset that my current work utilised consists of 300 images of BCC, SCC and IEC skin cancers. To scale the project to include more classes and diversity I have arranged data access agreements with MyLab Pty. Ltd. and Southern Sun Pathology Pty. Ltd., two histopathology laboratories that specialise in skin cancer. Southern Sun Pathology has shared 1000 slides of a variety of skin cancers along with the pathologist diagnosis and descriptions. MyLab has provided access to their archives as well as hardware to image them. Given the flexibility with collecting slides of our choosing and to exert control over quality, a major component of the project will be to image slides provided by MyLab. They will additionally provide access to the pathologist diagnosis and descriptions. My associate advisor, Dr Glenn Baxter, is also employed with MyLab as a pathologist and has already begun to facilitate slide collection.

In terms of skin cancer diagnosis, I need to develop a more thorough understanding of the histological diagnosis and characterisation of the various cancers. Having been working on this topic for almost two years now, I am broadly familiar with the features of each cancer type. However, to truly develop a system for clinical use it is necessary to understand more deeply the cellular morphology and variation within. Further to that, it is important to understand the pathologist workflow to be able to successful integrate the technology. To facilitate this, I will continue to be guided by Dr Glenn Baxter who has agreed to provide help on these matters. I additionally have access to histology and skin cancer textbooks through the UQ library.

In general, everything I need to complete the project is available or will be made available when and where required.

7 Current and Planned Publications

My current work on segmentation and classification is in preparation for publication. The working title is:

> *Towards Interpretable Deep Learning Systems for Multi-Class Segmentation and Classification of Non-Melanoma Skin Cancer.*

Possible journals include the following and are listed in order of ambition and reverse order of likelihood of acceptance:

- Nature Medicine – publishing a lot of binary classification in digital pathology
- Nature Scientific Reports – publishing a lot of binary classification in digital pathology (open access)
- JAMA – publishing some work in digital pathology
- Medical Imaging – standard medical imaging journal
- BMC Bioinformatics – calling for machine learning and image analysis (open access)
- Pathology Informatics – dedicated journal (open access)

Presentations and posters derived from the above work have been or will be presented at various symposiums, conferences and discussion groups.

Presentations:

- Two Minute Tease Talk - Translational Research Symposium at TRI, 6th August 2019 (2min)
- 18th July 2019 Australian Skin and Skin Cancer Seminar (ASSC) at TRI (30 min)
- Australian Society for Medical Research (ASMR) Student Conference at TRI – 22nd May 2019 (10 min) – **2nd Place Award ($150)**
- Practical Machine Learning Group at UQ – 15th May 2019 (1 hour)
- Skin Cancer Research Group at UQ – 1st March 2019 (15min)
- Practical Machine Learning Group at UQ – 10th October May 2018 (40 min)

Posters:

- Accepted - Biomed Link 2019, 5th November 2019 – Melbourne
- Pathology Visions 2019 – 5th October 2019, Orlando USA - **Travel Award Recipient ($3000 USD)**
- Translational Research Symposium at TRI, 6th August 2019 – **Student Category - 1st Place ($250)**
- Princess Alexandra Hospital (PAH) Health Symposium – 31st July 2019
- IMB Student Symposium – 21st June 2019

Additional Planned Publications

- Interpretability methods for multi-class skin cancer classification
- Lack of domain specificity for transfer learning
- Unsupervised learning to capture the variation of cellular morphology in skin cancer

8 Appendix

Figure A1: Class label colours for the segmentations. Used to interpret segmentations shown in Figures 4.1-4.3.

Figure A2: Confusion matrix for 10x data. Values read across columns for each row describe Recall *i.e.* true positive rate.

9 References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. in *Advances in neural information processing systems* 1097–1105 (2012).
3. Harari, Y. N. Reboot for the AI revolution. *Nat. News* **550**, 324 (2017).
4. Press Association. AI revolution 'at risk of being stifled in UK by fear-driven backlash'. *The Guardian* (2018).
5. Makridakis, S. The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures* **90**, 46–60 (2017).
6. Benaich, N. & Hogarth, I. *State of AI 2019*. (2019).
7. trends.google.com. Google Trends. (2019).
8. arXiv. arXiv usage statistics. (2019). Available at: https://arxiv.org/help/stats. (Accessed: 8th August 2019)
9. Yang, X., Li, Y. & Lyu, S. Exposing deep fakes using inconsistent head poses. in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 8261–8265 (2019).
10. Hambleton, S. J. & Aloizos AM, J. Australia's digital health journey. *Med. J. Aust.* **210**, S5–S6 (2019).
11. Mishra, S. Does modern medicine increase life-expectancy: Quest for the Moon Rabbit? *Indian Heart J.* **68**, 19–27 (2016).
12. United Nations. *[World population prospects 2019]. United Nations. Department of Economic and Social Affairs. World Population Prospects 2019.* (2019).
13. Department of Parliamentary Services. *Budget Review 2015*. (2015).
14. Krittanawong, C., Zhang, H. J., Wang, Z., Aydar, M. & Kitai, T. Artificial Intelligence in Precision Cardiovascular Medicine. *Journal of the American College of Cardiology* **69**, 2657–2664 (2017).
15. Perkins, B. A. *et al*. Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. *Proc. Natl. Acad. Sci.* **115**, 3686–3691 (2018).
16. Fleming, N. Computer-calculated compounds. *Nature* 5–7 (2018).
17. Patil, R. S. Artificial Intelligence Techniques for Diagnostic Reasoning in Medicine. in *Exploring Artificial Intelligence* 347–379 (1988). doi:10.1016/b978-0-934613-67-5.50013-7
18. Kahn, M. G. Artificial intelligence in medicine workshop. *Artif. Intell. Med.* **4**, 409–411 (1992).
19. Chandrasekaran, B. On Evaluating Artificial Intelligence Systems for Medical Diagnosis. *AI Mag.* **4**, 34–37 (1983).
20. Kuipers, B. New reasoning methods for artificial intelligence in medicine. *Int. J. Man. Mach. Stud.* **26**, 707–718 (1987).
21. L, M. P. The evaluation of artificial intelligence systems in medicine. *Comput. Methods Programs Biomed.* **22**, 5 (1986).

22. Horvitz, E. J., Breese, J. S. & Henrion, M. Decision theory in expert systems and {AI}. *Intl. J. Approx. Reason.* **2**, 247–302 (1988).

23. Schwartz, W. B., Patil, R. S. & Szolovits, P. Artificial Intelligence in Medicine. *N. Engl. J. Med.* **316**, 685–688 (1987).

24. Servan-Schreiber, D. Artificial intelligence and psychiatry. *Journal of Nervous and Mental Disease* **174**, 191–202 (1986).

25. Shortliffe, E. H., Axline, S. G., Buchanan, B. G., Merigan, T. C. & Cohen, S. N. An Artificial Intelligence program to advise physicians regarding antimicrobial therapy. *Comput. Biomed. Res.* **6**, 544–560 (1973).

26. Coats, P. Why expert systems fail. *Financ. Manag.* **17**, 77–86 (1988).

27. Chang, A. C. Big data in medicine: The upcoming artificial intelligence. *Prog. Pediatr. Cardiol.* **43**, 91–94 (2016).

28. Chen, X.-W. & Lin, X. Big data deep learning: challenges and perspectives. *IEEE access* **2**, 514–525 (2014).

29. Miller, D. D. & Brown, E. W. Artificial Intelligence in Medical Practice: The Question to the Answer? *Am. J. Med.* **131**, 129–133 (2018).

30. Chen, Y., Argentinis, E. & Weber, G. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clin. Ther.* **38**, 688–701 (2016).

31. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).

32. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).

33. Han, Z. *et al.* Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model. *Sci. Rep.* **7**, 4172 (2017).

34. Li, J. *et al.* A Multi-scale U-Net for Semantic Segmentation of Histological Images from Radical Prostatectomies. *AMIA ... Annu. Symp. proceedings. AMIA Symp.* **2017**, 1140–1148 (2017).

35. Havaei, M. *et al.* Brain tumor segmentation with Deep Neural Networks. *Med. Image Anal.* **35**, 18–31 (2017).

36. Trebeschi, S. *et al.* Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. *Sci. Rep.* **7**, 5301 (2017).

37. Che, N., Chen, D. & Le, J. Entity Recognition Approach of Clinical Documents Based on Self-training Framework. in *Recent Developments in Intelligent Computing, Communication and Devices* (eds. Patnaik, S. & Jain, V.) 259–265 (Springer Singapore, 2019).

38. KS, K. & Sangeetha, S. SECNLP: A Survey of Embeddings in Clinical Natural Language Processing. 1–45 (2019).

39. Chalapathy, R., Borzeshi, E. Z. & Piccardi, M. Bidirectional LSTM-CRF for Clinical Concept Extraction. (2016).

40. Liang, H. *et al.* Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat. Med.* **25**, 433–438 (2019).

41. Ching, T. *et al. Opportunities and obstacles for deep learning in biology and medicine : 2019 update Authors*. *J. R. Soc. Interface* **15**, (2018).

42. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical Image Analysis* **42**, 60–88 (2017).

43. Blease, C. *et al.* Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. *J. Med. Internet Res.* **21**, (2019).

44. Krittanawong, C. The rise of artificial intelligence and the uncertain future for physicians. *European Journal of Internal Medicine* **48**, e13–e14 (2018).

45. Hamet, P. & Tremblay, J. Artificial intelligence in medicine. *Metabolism.* **69**, S36–S40 (2017).

46. Bini, S. A. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *J. Arthroplasty* **33**, 2358–2361 (2018).

47. Buch, V. H., Ahmed, I. & Maruthappu, M. Artificial intelligence in medicine: current trends and future possibilities. *Br. J. Gen. Pract.* **68**, 143–144 (2018).

48. Larson, J., Johnson, M., American, S. B.-J. of the & 2014, undefined. Application of surgical safety standards to robotic surgery: five principles of ethics for nonmaleficence. *journalacs.org*

49. Cornet, G. Robot companions and ethiCs a pragmatiC appRoaCh of ethiCal dEsign. *J. Int. Bioethique* **33**, 49–58 (2013).

50. Mori, M. Bukimi no tani [the uncanny valley]. *Energy* **7**, 33–35 (1970).

51. Ross, C. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. (2018). Available at: https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/. (Accessed: 9th September 2019)

52. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* bbx044–bbx044 (2017). doi:10.1093/bib/bbx044

53. Jiang, F. *et al.* Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2**, 230–243 (2017).

54. Stead, W. W. Clinical Implications and Challenges of Artificial Intelligence and Deep LearningClinical Implications and Challenges of Artificial Intelligence and Deep LearningEditorial. *JAMA* **320**, 1107–1108 (2018).

55. Hajar, R. The physician's oath: Historical perspectives. *Hear. views Off. J. Gulf Hear. Assoc.* **18**, 154 (2017).

56. Antoniou, S. A. *et al.* Reflections of the Hippocratic Oath in modern medicine. *World J. Surg.* **34**, 3075–3079 (2010).

57. Gardner, J. M. & Allen, T. C. Keep calm and tweet on: Legal and ethical considerations for pathologists using social media. *Arch. Pathol. Lab. Med.* **143**, 75–80 (2019).

58. Shringarpure, S. S. & Bustamante, C. D. Privacy risks from genomic data-sharing beacons. *Am. J. Hum. Genet.* **97**, 631–646 (2015).

59. Harting, M. T., Dewees, J. M., Vela, K. M. & Khirallah, R. T. Medical photography: Current technology, evolving issues and legal perspectives. *International Journal of Clinical Practice* **69**, 401–409 (2015).

60. Dry, S. Who Owns Diagnostic Tissue Blocks? *Lab. Med.* **40**, 69–73 (2009).

61. Hakimian, R. & Korn, D. Ownership and Use of Tissue Specimens for Research. *JAMA* **292**,

2500–2505 (2004).

62. Abadi, M. *et al.* TensorFlow: A System for Large-Scale Machine Learning. in *OSDI* **16**, 265–283 (2016).

63. Paszke, A. *et al.* Automatic differentiation in PyTorch. in *NIPS-W* (2017).

64. Rutkin, A. Digital discrimination. *New Sci.* **231**, 18–19 (2016).

65. Goodman, B. & Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.* **38**, 50–57 (2017).

66. Rosenblatt, F. *The perceptron, a perceiving and recognizing automaton Project Para*. (Cornell Aeronautical Laboratory, 1957).

67. Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. in *Proceedings of the 27th international conference on machine learning (ICML-10)* 807–814 (2010).

68. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Int. Conf. Mach. Learn.* 448–456 (2015).

69. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 249–256

70. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *arXiv Prepr. arXiv1412.6980* (2014).

71. Caron, M., Bojanowski, P., Joulin, A. & Douze, M. Deep clustering for unsupervised learning of visual features. in *Proceedings of the European Conference on Computer Vision (ECCV)* 132–149 (2018).

72. Deng, J. *et al.* ImageNet: A Large-Scale Hierarchical Image Database. in *CVPR09* (2009).

73. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).

74. Chollet, F. Keras - GitHub repository. (2015).

75. Howard, A. G. *et al.* Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv Prepr. arXiv1704.04861* (2017).

76. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv Prepr. arXiv1409.1556* (2014).

77. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4510–4520 (2018).

78. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 4700–4708 (2017).

79. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778

80. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 1492–1500 (2017).

81. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception

Architecture for Computer Vision. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2016-Decem**, 2818–2826 (2016).

82. Chollet, F. Xception: Deep learning with depthwise separable convolutions. *arXiv Prepr.* (2016).

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

84. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. in *Thirty-First AAAI Conference on Artificial Intelligence* (2017).

85. Zoph, B., Vasudevan, V., Shlens, J. & Le, Q. V. Learning transferable architectures for scalable image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 8697–8710 (2018).

86. Touvron, H., Vedaldi, A., Douze, M. & Jégou, H. Fixing the train-test resolution discrepancy. *arXiv Prepr. arXiv1906.06423* (2019).

87. Mahajan, D. *et al.* Exploring the limits of weakly supervised pretraining. in *Proceedings of the European Conference on Computer Vision (ECCV)* 181–196 (2018).

88. Tan, M. & Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv Prepr. arXiv1905.11946* (2019).

89. Huang, Y. *et al.* Gpipe: Efficient training of giant neural networks using pipeline parallelism. *arXiv Prepr. arXiv1811.06965* (2018).

90. Everingham, M., Gool, L., Williams, C. K., Winn, J. & Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **88**, 303–338 (2010).

91. Cordts, M. *et al.* The Cityscapes Dataset for Semantic Urban Scene Understanding. in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

92. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. in *Proceedings of the {IEEE} conference on computer vision and pattern recognition* 3431–3440 (2015).

93. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. in *Proceedings of the IEEE international conference on computer vision* 2980–2988 (2017).

94. Javanmardi, M., Sajjadi, M., Liu, T. & Tasdizen, T. Unsupervised total variation loss for semi-supervised deep learning of semantic segmentation. *arXiv Prepr. arXiv1605.01368* (2016).

95. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9351**, 234–241 (Springer Verlag, 2015).

96. Iglovikov, V. & Shvets, A. TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. (2018).

97. Xia, X. & Kulis, B. W-Net: A Deep Model for Fully Unsupervised Image Segmentation. (2017).

98. Kamnitsas, K. *et al.* Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **36**, 61–78 (2017).

99. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. in *Proceedings of the European*

conference on computer vision (ECCV) 801–818 (2018).

100. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2017).

101. Berman, M., Rannen Triki, A. & Blaschko, M. B. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4413–4421 (2018).

102. Lin, T.-Y. *et al.* Microsoft coco: Common objects in context. in *European conference on computer vision* 740–755 (Springer, 2014).

103. Sun, C., Shrivastava, A., Singh, S. & Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. in *Proceedings of the IEEE international conference on computer vision* 843–852 (2017).

104. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv Prepr. arXiv1312.6114* (2013).

105. Goodfellow, I. *et al.* Generative adversarial nets. in *Advances in neural information processing systems* 2672–2680 (2014).

106. Kodali, N., Abernethy, J., Hays, J. & Kira, Z. On convergence and stability of gans. *arXiv Prepr. arXiv1705.07215* (2017).

107. Odena, A., Dumoulin, V. & Olah, C. Deconvolution and checkerboard artifacts. *Distill* **1**, e3 (2016).

108. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein gan. *arXiv Prepr. arXiv1701.07875* (2017).

109. Zhang, H. *et al.* Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. in *Proceedings of the IEEE International Conference on Computer Vision* 5907–5915 (2017).

110. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv Prepr. arXiv1710.10196* (2017).

111. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4401–4410 (2019).

112. Brock, A., Donahue, J. & Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv Prepr. arXiv1809.11096* (2018).

113. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

114. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv Prepr. arXiv1802.03426* (2018).

115. Wattenberg, M., Viégas, F. & Johnson, I. How to use t-SNE effectively. *Distill* **1**, e2 (2016).

116. Erhan, D., Bengio, Y., Courville, A. & Vincent, P. Visualizing higher-layer features of a deep network. *Univ. Montr.* **1341**, 1 (2009).

117. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv Prepr. arXiv1706.03825* (2017).

118. Selvaraju, R. R. *et al.* Grad-CAM: Why did you say that? *arXiv Prepr. arXiv1611.07450*

(2016).

119. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* 3319–3328 (JMLR. org, 2017).

120. Dabkowski, P. & Gal, Y. Real time image saliency for black box classifiers. in *Advances in Neural Information Processing Systems* 6967–6976 (2017).

121. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. in *Proceedings of the IEEE International Conference on Computer Vision* 618–626 (2017).

122. Adebayo, J. *et al.* Sanity checks for saliency maps. in *Advances in Neural Information Processing Systems* 9505–9515 (2018).

123. Ghorbani, A., Abid, A. & Zou, J. Interpretation of neural networks is fragile. in *Proceedings of the AAAI Conference on Artificial Intelligence* **33**, 3681–3688 (2019).

124. Kim, B. *et al.* Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv Prepr. arXiv1711.11279* (2017).

125. Mahendran, A. & Vedaldi, A. Understanding deep image representations by inverting them. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **07-12-June-2015**, 5188–5196 (IEEE Computer Society, 2015).

126. Szegedy, C. *et al.* Intriguing properties of neural networks. *arXiv Prepr. arXiv1312.6199* (2013).

127. Nguyen, A., Yosinski, J. & Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 427–436 (2015).

128. Olah, C., Mordvintsev, A. & Schubert, L. Feature visualization. *Distill* **2**, e7 (2017).

129. Ilyas, A. *et al.* Adversarial examples are not bugs, they are features. *arXiv Prepr. arXiv1905.02175* (2019).

130. Engstrom, L. *et al.* Learning Perceptually-Aligned Representations via Adversarial Robustness. *arXiv Prepr. arXiv1906.00945* (2019).

131. Santurkar, S. *et al.* Computer Vision with a Single (Robust) Classifier. *CoRR* **abs/1906.0**, (2019).

132. Olah, C. *et al.* The building blocks of interpretability. *Distill* **3**, e10 (2018).

133. Carter, S., Armstrong, Z., Schubert, L., Johnson, I. & Olah, C. Activation atlas. *Distill* **4**, e15 (2019).

134. Szegedy, C. *et al.* Going Deeper with Convolutions (GoogLeNet). *Comput. Res. Repos. (CoRR), abs/1409.4842* (2014).

135. Engstrom, L. *et al.* A Discussion of'Adversarial Examples Are Not Bugs, They Are Features': Discussion and Author Responses. *Distill* **4**, e00019-7 (2019).

136. Rakin, A. S., He, Z. & Fan, D. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. *arXiv Prepr. arXiv1811.09310* (2018).

137. Titford, M. A Short History of Histopathology Technique. *J. Histotechnol.* **29**, 99–110 (2006).

138. Fischer, A. H., Jacobson, K. A., Rose, J. & Zeller, R. Hematoxylin and eosin staining of tissue

and cell sections. *CSH Protoc.* **2008**, db.prot4986 (2008).

139. Ferreira, R. *et al.* The Virtual Microscope. *Proc. a Conf. Am. Med. Informatics Assoc. AMIA Fall Symp.* 449–453 (1997).

140. Pantanowitz, L. Digital images and the future of digital pathology. *J. Pathol. Inform.* **1**, (2010).

141. Indu, M., Rathy, R. & Binu, M. P. 'Slide less pathology': Fairy tale or reality? *J. Oral Maxillofac. Pathol.* **20**, 284–288 (2016).

142. Abels, E. & Pantanowitz, L. Current state of the regulatory trajectory for whole slide imaging devices in the USA. *J. Pathol. Inform.* **8**, 23 (2017).

143. Leica Biosystems. Leica Biosystems Receives FDA 510(k) Clearance to Market a Digital Pathology System for Primary Diagnosis. (2019). Available at: https://www.leicabiosystems.com/news-events/news-details/article/leica-biosystems-receives-fda-510k-clearance-to-market-a-digital-pathology-system-for-primary-diag/News/detail/. (Accessed: 23rd August 2019)

144. Tizhoosh, H. R. & Pantanowitz, L. Artificial Intelligence and Digital Pathology: Challenges and Opportunities. *J. Pathol. Inform.* **9**, 38 (2018).

145. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, (2019).

146. Business Wire. FDA Grants Breakthrough Designation to Paige.AI. (2019). Available at: https://www.businesswire.com/news/home/20190307005205/en/FDA-Grants-Breakthrough-Designation-Paige.AI. (Accessed: 23rd August 2019)

147. Teramoto, A., Tsukamoto, T., Kiriyama, Y. & Fujita, H. Automated classification of lung cancer types from cytological images using deep convolutional neural networks. *Biomed Res. Int.* **2017**, (2017).

148. Wu, M., Yan, C., Liu, H. & Liu, Q. Automatic classification of ovarian cancer types from cytological images using deep convolutional neural networks. *Biosci. Rep.* **38**, BSR20180289 (2018).

149. Bychkov, D. *et al.* Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Rep.* **8**, (2018).

150. Chen, J. & Srinivas, C. Automatic lymphocyte detection in H&E images with deep neural networks. *arXiv Prepr. arXiv1612.03217* (2016).

151. Bejnordi, B. E. *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA - J. Am. Med. Assoc.* **318**, 2199–2210 (2017).

152. Lu, C. *et al.* Multi-Pass Adaptive Voting for Nuclei Detection in Histopathological Images. *Sci. Rep.* **6**, 1–18 (2016).

153. Sornapudi, S. *et al.* Deep Learning Nuclei Detection in Digitized Histology Images by Superpixels. *J. Pathol. Inform.* **9**, 5 (2018).

154. Cruz-Roa, A. *et al.* Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci. Rep.* **7**, 46450 (2017).

155. Liu, Y. *et al.* Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists. *Arch. Pathol. Lab. Med.* (2018).

156. Litjens, G. *et al.* 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *Gigascience* **7**, giy065 (2018).

157. Liu, Y. *et al.* Detecting Cancer Metastases on Gigapixel Pathology Images. (2017).

158. Lee, G. *et al.* Nuclear Shape and Architecture in Benign Fields Predict Biochemical Recurrence in Prostate Cancer Patients Following Radical Prostatectomy: Preliminary Findings. *Eur. Urol. Focus* **3**, 457–466 (2017).

159. Lu, C. *et al.* Nuclear shape and orientation features from H&E images predict survival in early-stage estrogen receptor-positive breast cancers. *Lab. Invest.* **98**, 1438–1448 (2018).

160. Corredor, G. *et al.* Spatial Architecture and Arrangement of Tumor-Infiltrating Lymphocytes for Predicting Likelihood of Recurrence in Early-Stage Non-Small Cell Lung Cancer. *Clin. Cancer Res.* **25**, 1526–1534 (2019).

161. Mungle, T. *et al.* MRF-ANN: a machine learning approach for automated ER scoring of breast cancer immunohistochemical images. *J. Microsc.* **267**, 117–129 (2017).

162. Wang, X. *et al.* Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital H&E images. *Sci. Rep.* **7**, 1–10 (2017).

163. Yu, K.-H. *et al.* Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7**, 12474 (2016).

164. Arevalo, J., Cruz-Roa, A., Arias, V., Romero, E. & González, F. A. An unsupervised feature learning framework for basal cell carcinoma image analysis. *Artif. Intell. Med.* **64**, 131–145 (2015).

165. Senaras, C., Niazi, M. K. K., Lozanski, G. & Gurcan, M. N. DeepFocus: Detection of out-of-focus regions in whole slide digital images using deep learning. *PLoS One* **13**, e0205387 (2018).

166. Xu, Z., Moro, C. F., Bozóky, B. & Zhang, Q. Gan-based virtual re-staining: A promising solution for whole slide image analysis. *arXiv Prepr. arXiv1901.04059* (2019).

167. Han, S. S. *et al.* Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *J. Invest. Dermatol.* **138**, 1529–1538 (2018).

168. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).

169. Zhang, Z. *et al.* Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat. Mach. Intell.* **1**, 289–289 (2019).

170. Romo-Bucheli, D., Moncayo, R., Cruz-Roa, A. & Romero, E. Identifying histological concepts on basal cell carcinoma images using nuclei based sampling and multi-scale descriptors. in *2015 {IEEE} 12th International Symposium on Biomedical Imaging ({ISBI})* 1008–1011 (2015).

171. Rosenbaum, B. E. *et al.* Computer-assisted measurement of primary tumor area is prognostic of recurrence-free survival in stage IB melanoma patients. *Mod. Pathol. an Off. J. United States Can. Acad. Pathol. Inc* **30**, 1402–1410 (2017).

172. Jiang, Y. Q. *et al.* Recognizing Basal Cell Carcinoma on Smartphone-Captured Digital Histopathology Images with Deep Neural Network. *Br. J. Dermatol.* (2019). doi:10.1111/bjd.18026

173. Rosado, B. *et al.* Accuracy of computer diagnosis of melanoma: a quantitative meta-analysis. *Arch. Dermatol.* **139**, 361–367 (2003).

174. Lu, C. & Mandal, M. Automated analysis and diagnosis of skin melanoma on whole slide histopathological images. *Pattern Recognit.* **48**, 2738–2750 (2015).

175. Hekler, A. *et al.* Pathologist-level classification of histopathological melanoma images with deep neural networks. *Eur. J. Cancer* **115**, 79–83 (2019).

176. Sturm, B. *et al.* Validation of whole-slide digitally imaged melanocytic lesions: Does z-stack scanning improve diagnostic accuracy? *J. Pathol. Inform.* **10**, (2019).

177. Halicek, M. *et al.* Detection of squamous cell carcinoma in digitized histological images from the head and neck using convolutional neural networks. in *Medical Imaging 2019: Digital Pathology* **10956**, 109560K (International Society for Optics and Photonics, 2019).

178. Rogers, H. W., Weinstock, M. A., Feldman, S. R. & Coldiron, B. M. Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the U.S. Population, 2012. *JAMA dermatology* **151**, 1081–1086 (2015).

179. Australian Institute of Health and Welfare. *Skin Cancer in Australia*. **Cat. no. C**, (2016).

180. Staples, M. P. *et al.* Non-melanoma skin cancer in Australia: the 2002 national survey and trends since 1985. *Med. J. Aust.* **184**, 6–10 (2006).

181. Geller, A. C. & Annas, G. D. Epidemiology of melanoma and nonmelanoma skin cancer. *Semin. Oncol. Nurs.* **19**, 2–11 (2003).

182. Shih, S. T., Carter, R., Heward, S. & Sinclair, C. Economic evaluation of future skin cancer prevention in Australia. *Prev. Med.* **99**, 7–12 (2017).

183. Bolognia, J. L., Jorizzo, J. L. & Rapini, R. P. *Dermatology*. (Gulf Professional Publishing, 2003).

184. Phulari, R. G., Rathore, R., Talegaon, T. P. & Shah, A. Cutaneous horn: A mask to underlying malignancy. *J. Oral Maxillofac. Pathol.* **22**, S87--S90 (2018).

185. Rapini, R. P. *Practical Dermatopathology*. (Elsevier Health Sciences, 2012).

186. Patterson, J. W. Tumors of the epidermis. in *Weedon's Skin Pathology* (ed. Patterson MD, FACP, FAAD, J. W.) 783-835.e29 (2016). doi:http://dx.doi.org/10.1016/B978-0-7020-5183-8.00031-X

187. Elder, D. E. *Lever's histopathology of the skin*. (Lippincott Williams & Wilkins, 2014).

188. Saldanha, G., Fletcher, A. & Slater, D. N. Basal cell carcinoma: a dermatopathological and molecular biological update. *Br. J. Dermatol.* **148**, 195–202 (2003).

189. Slater, D. & Barrett, P. *Dataset for histopathological reporting of primary invasive cutaneous squamous cell carcinoma and regional lymph nodes*. (2019).

190. Royal College of Pathologists of Australasia. *Primary cutaneous melanoma structured areporting protocol*. (2014).

191. Nguyen, A., Yosinski, J. & Clune, J. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv Prepr. arXiv1602.03616* (2016).

192. Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T. & Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. in *Advances in Neural Information Processing Systems* 3387–3395 (2016).

193. Elson, J., Douceur, J. J. D., Howell, J. & Saul, J. Asirra: a CAPTCHA that exploits interest-aligned manual image categorization. (2007).