Medical Image Analysis 68 (2021) 101915

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer

Simon M. Thomas[a,b], James G. Lefevre[a], Glenn Baxter[b], Nicholas A. Hamilton[a,*]

[a] Institute for Molecular Bioscience, University of Queensland, 306 Carmody Road, St Lucia, Australia
[b] MyLab Pathology, 11 Hayling Street, Salisbury, Australia

ARTICLE INFO

Article history:
Received 4 December 2019
Revised 9 November 2020
Accepted 16 November 2020
Available online 21 November 2020

Keywords:
Deep learning
Machine learning
Computational Pathology
Skin cancer
Segmentation
Classification

ABSTRACT

We apply for the first-time interpretable deep learning methods simultaneously to the most common skin cancers (basal cell carcinoma, squamous cell carcinoma and intraepidermal carcinoma) in a histological setting. As these three cancer types constitute more than 90% of diagnoses, we demonstrate that the majority of dermatopathology work is amenable to automatic machine analysis. A major feature of this work is characterising the tissue by classifying it into 12 meaningful dermatological classes, including hair follicles, sweat glands as well as identifying the well-defined stratified layers of the skin. These provide highly interpretable outputs as the network is trained to represent the problem domain in the same way a pathologist would. While this enables a high accuracy of whole image classification (93.6-97.9%), by characterising the full context of the tissue we can also work towards performing routine pathologist tasks, for instance, orientating sections and automatically assessing and measuring surgical margins. This work seeks to inform ways in which future computer aided diagnosis systems could be applied usefully in a clinical setting with human interpretable outcomes.

1. Introduction

Non-melanoma skin cancers comprise over 90% of all skin cancer diagnoses, compared to melanoma which comprises approximately 4% (Geller and Annas, 2003). The most common forms of non-melanoma skin cancer include basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) which are among the most common forms of cancer in Australia and the United States (Rogers et al., 2015; AIHW, 2016). BCC comprises approximately 60% of all skin cancer diagnoses, with SCC comprising a further 30% (Staples et al., 2006). Another non-melanoma skin cancer is intraepidermal carcinoma (IEC), otherwise known as squamous carcinoma *in situ* or Bowen's disease. Between 3-5% of IECs develop into an invasive squamous cell carcinoma (Patterson, 2016). The developing field of digital pathology has a demonstrated interest in incorporating "smart" algorithms into a fully-digitised diagnostic pipeline (Acs and Rimm, 2018; Niazi et al., 2019; Tizhoosh and Pantanowitz, 2018). In an increasingly demanding environment for pathologists (Lundberg, 2019; Metter et al., 2019; Royal College of Pathologists of Australia, 2018), it is expected that deep learning methods can serve to improve the efficiency of pathologist workflows, such as in the case of diagnosing non-melanoma skin cancer.

1.1. Interpretable models

Deep Learning algorithms have been shown to outperform expert humans in some medical image diagnosis problems (Liu et al., 2019) but remain highly criticised for being "black boxes". This is seen as a major short-coming in the application of the technology to high-stakes decisions (Rudin, 2019). Indeed, addressing the problem of interpretability is an explicit goal in recent work in the analysis of histopathology images (Bychkov et al., 2018; Campanella et al., 2019; Faust et al., 2019; Zhang et al., 2017; Zhang et al., 2019) and machine learning in general (Adebayo et al., 2018; Carter et al., 2019; Kim et al., 2018; Olah et al., 2018, 2017; Rudin, 2019). Interpretability methods tend to be at the level of visualising filters to compare with human-meaningful features (Olah et al., 2018, 2017) or concordance between similar images (Campanella et al., 2019; Esteva et al., 2017; Hegde et al., 2019). However, an alternative approach to these interpretability methods is to implement models in such a way for them to be naturally interpretable (Chen et al., 2018; Rudin, 2019). Although what constitutes such a model is open to discussion, it can generally be thought of as a model which explicitly characterises a problem the same way a human would. In the case of histopathology, a pathologist brings to bear a wealth of knowledge in assessing

* Corresponding author at: The University of Queensland, Institute for Molecular Bioscience, 306 Carmody Road, St Lucia 4072, Australia.
E-mail addresses: simon.thomas@uq.edu.au (S.M. Thomas), j.lefevre@uq.edu.au (J.G. Lefevre), n.hamilton@uq.edu.au, n.hamilton@imb.uq.edu.au (N.A. Hamilton).

https://doi.org/10.1016/j.media.2020.101915
1361-8415/

a specimen. Importantly, they are doing much more than cancer or non-cancer classification. Instead, they can characterise the entire tissue, identifying what and where the cancer is in context (as will be described in the next section). Therefore, a naturally interpretable model would be capable of performing the same task. Such a model would learn the full-visual context of the problem, and its ability to do so would be explicit by its output. It is from this rich output that we can make inferences about the quality of the model's representational system. We believe that semantic segmentation methods are well suited to achieving this and aim to demonstrate this approach in the context of non-melanoma skin cancer diagnosis.

1.2. Skin cancer diagnosis

In the process of assessing skin cancer, a pathologist generally sees multiple slides on which multiple tissue sections show different histomorphological features. This is because the ability for a pathologist to perform an efficient and accurate diagnosis, as well as inform further clinical intervention, is closely linked to the sample preparation process. Intact excision specimens are often received oriented in relation to the site of excision within the patient, and this is maintained during specimen processing and examination using blue and black ink to distinguish the different margins. The region of interest in the specimen is sliced into 3 mm thick transverse sections, revealing the full extent of the surgical margins. The sections are processed in various solvents, impregnated with paraffin, sliced into transparent 3μm sections, and, most commonly, stained with hematoxylin and eosin dye (H&E) to differentiate nuclei, cytoplasm, stroma and other structures (Fischer et al., 2008; Titford, 2006). The slides are analysed to characterize and assess the lesion, which includes type, degree of differentiation, 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.

It is evident that skin cancer diagnosis is a multi-faceted problem. In fact, the minimum scope of what a pathologist does in reporting BCC and SCC cases, detailed by the Royal College of Pathologists (UK) in their reporting criteria (Saldanha et al., 2003; Slater and Barrett, 2019), expands considerably on the steps just described. One aspect is cancer subtyping, which can depend on cellular morphologies, macroscopic features and growth patterns or metrics such as depth of invasion. Another aspect is identifying evidence for lymphovascular or perineural invasion, which is critical to informing treatment options, particularly in the case of metastasis. Indeed, sometimes what is important is not what the cancer is (which may be obvious) but in what context it lies.

When considering these requirements to produce a report useful to the clinician, is it clear that binary classification approaches are limited in their ability to address them, despite that being the most common approach of previous work (Arevalo et al., 2015; Campanella et al., 2019; Halicek et al., 2019, 2017; Hekler et al., 2019; Jiang et al., 2020; Lu and Mandal, 2015; Rosado et al., 2003). Importantly, interpretability has not been a central component of previous work in skin cancer, despite being a clinically high-stakes decision domain. In contrast, a pipeline built on multi-class semantic segmentation provides a promising means to accomplish many of these routine steps as well as be naturally interpretable.

2. Methods

2.1. Data collection

MyLab Pathology provided access to their pre-existing collection of skin cancer slides. We selected 290 H&E slides representing typical cases of BCC (140), SCC (60) and IEC (90). Here *typical* refers to clear and unambiguous diagnostic features, in contrast to difficult cases requiring a differential diagnosis. The set included shave (100), punch (58) and excision (132) biopsies. Each slide was hand-annotated by a pathologist to indicate which tissue section was most representative of the cancer class. The slides were produced (using xylene processing) and imaged over four months in late 2017 and early 2018. The slides correspond to patients between the ages of 34 and 96, with a median age of 70 years with male and female proportions of 2/3 and 1/3, respectively. These statistics closely reflect the prevalence of non-melanoma skin cancer in the population (Staples et al., 2006).

The slides were imaged using a DP27 Olympus microscope camera at 10x magnification with the condenser attached. At this magnification, only a portion of the tissue section can be imaged at once and so the whole tissue section was captured by combining overlapping tiles into a single high-resolution mosaic. This was performed using bespoke software created for this project, available at https://github.com/smthomas-sci/HistoImageStitcher. The resulting images had a resolution where 1 pixel width corresponds to 0.67µm in length. Since the resolution was the same across the three biopsy types, there was a large variation in the overall dimensions producing a range between 11 million and 500 million pixels.

2.2. Segmentation ground truth

The ground-truth segmentations were created in ImageJ by painting over the full-resolution images with colors assigned to 12 classification categories: Glands (GLD), Inflammation (INF), Hair Follicles (FOL), Hypodermis (HYP), Reticular Dermis (RET), Papillary Dermis (PAP), Epidermis (EPI), Keratin (KER), Background (BKG), BCC, SCC, and IEC. One challenge in describing the epidermis is that a large amount of variation can be considered non-cancerous. But because hard labels were required, we chose to label *ideal healthy* epidermis as EPI, and include in the IEC class features from mildly to severely dysplastic keratinocytes (solar keratosis) as well as carcinoma. This allowed for the fact that often multiple abnormalities are present in a specimen, *e.g.* BCC is accompanied with regions of moderately dysplastic keratinocytes. The annotations were performed by a dermatopathology laboratory scientist in consultation with a pathologist. It took approximately 250 hours to segment the whole dataset at the original resolution.

The resulting segmentations represented pixel-level class distinctions which introduced several conceptual challenges. Firstly, humans use many levels of interpretation, both conceptually and visually, to determine class boundaries. A segmentation based on higher-level diagnostic features can result in a coarse and inaccurate segmentation at the pixel-level. However, for some classes, boundaries can be distinguished at high resolution, such as the basement membrane between the epidermis and papillary dermis layers, and background and non-background. For other classes, the boundary is inherently ambiguous such as between papillary and reticular dermis. Secondly, individual pixels do not carry information to be meaningful to a human on their own. Yet for segmentation tasks, it is common to measure accuracy on a per-pixel basis. This raises the question, what is the maximum achievable classification accuracy? Certainly, 100% ground truth accuracy is not a valid assumption, however, it is difficult to quantify exactly how much these various problems limit our ability to measure performance. To understand this, we created three other datasets that were down sampled by factors of 2, 5 and 10 (referred to as the 2x, 5x and 10x datasets). This enabled the ability to compare accuracy across the various scales and class-boundary contexts. A visual comparison between the 1x (no downsampling) and 2x datasets

revealed there was no meaningful difference in the tissue features to the human eye, and so the 2x dataset was used as the highest resolution for analysis. Conversely, downsample factors beyond 10 resulted in various tissue types being difficult for clear human distinction, undermining efforts towards interpretability (Section 1.1).

2.3. Model architectures and training

In 2015, U-Net (Ronneberger et al., 2015) was introduced to perform segmentation on greyscalebioimages, and has since become a popular choice for segmentation tasks. Another architectural advancement is the Residual Network, which improves the trainability of deep neural networks (He et al., 2016). Recent work in bioimage segmentation utilize a combination of these advances (de Geus et al., 2018; Estienne et al., 2019; Rad et al., 2020; Q. Zhang et al., 2017), demonstrating the effectiveness of the approach. Applying these recent advances, we implemented a U-Net-like architecture using Keras v2.1.6 (Chollet, 2015) and Tensorflow v1.12.2 (Abadi et al., 2016) in a transfer learning setting. A pre-trained headless ResNet50 (He et al., 2016) network was used as the base feature extractor / encoder and a U-Net-like decoder was added with skip connections from previous convolutional layers (Fig. 1a). The original U-Net architecture used a sequence of 512, 256, 128, 64 convolution kernels during the decoding phase. However, we found that this was over-parameterised for our task, and after a systematic search settled on 32 convolution kernels for the whole decoding phase. A dropout layer with a value of 0.5 was used in combination with the final 1×1 convolutional layer. Importantly, the batch normalisation layers in ResNet50 were unlocked so that a new population mean and variance could be learned. Code are available at https://github.com/smthomas-sci/SkinCancerSegmentation.

The original images were too large to feed directly into the network and so for training, overlapping patches (256×256 pixels for 10x and 5x datasets, and 512×512 pixels for 2x dataset) were created for each image. A pixel-level class count revealed that the classes BCC, SCC, IEC and FOL were substantially underrepresented. To manage class imbalances during training, patches containing these classes were augmented by flipping and rotating them, increasing their presence by 8 times.

The patch sizes were chosen to maximise the amount of tissue present in each patch while accommodating varying-sized WSIs and memory constraints. As noted, patches of size 256×256 and 512×512 were used. However, in terms of the features that the network is learning, it is the receptive field that is important, and training patch size has no effect on post-training predictions and performance (see later).

The networks were trained with two 16GB NVIDIA Tesla V100s in parallel, using an Adam optimizer with a learning rate of 1×10^{-4} for 20-50 epochs. The 5x and 10x datasets had a batch size of 24 and the 2x dataset had 12, all having an 80:10:10 training, validation and test split. Excisions, Shaves and Punches, as well as the various cancer types were distributed evenly across the splits. The contribution of each class to the categorical cross-entropy loss was weighted based on the number of pixels belonging to each class on a per-batch basis. This was necessary due to a persistent class-imbalance problem. Each run was replicated 3 times with random seeds of 1, 2 and 3. After the initial training phase the weights that produced the lowest validation loss were used to fine-tune the whole network with a learning rate of 1×10^{-6} for up to 5 epochs. Finally, the fine-tuned weights with the lowest validation loss were evaluated on the test set and the results were averaged.

2.4. Whole image segmentation

Whole image segmentation (Fig. 1b) was performed using the network with the lowest validation loss from the fine-tuned replicates. The available memory resources limited the input size of the network to less than $1{,}500 \times 1{,}500$ pixels, and so a suitable classification tile size was calculated for each image where the overlap was at least 100 pixels. To help improve tiling as well as minimise border effects the original image was first padded with white pixels (100 pixels on each edge), essentially extending the background class. The tile was then fed into the network and an argmax function applied class allocations to each pixel in the tile. The segmentation tile was then cropped to remove the overlap on each edge. The whole image segmentation was generated iteratively one tile at a time and was finally cropped to the original size. Each image for the 10x dataset was processed in a few seconds whereas the 2x images took up to a minute.

2.5. Whole image classification

Using the 10x dataset, the h x w x 12 output of the whole image segmentation workflow was saved for 290 cancer images, as well as a sub-sample of healthy crops (containing no cancerous tissue). These outputs are from here referred to as probability maps because each pixel represents a probability distribution across the 12 classes of skin structures. These probability maps were used to train and test another network (Fig. 1c) to perform whole image classification in a four-class setting. The original 80:10:10 splits were maintained for the classification tasks, with healthy crops taken from images within each split, respectively. For the test set n=47 with class counts as follows: Healthy (n=16), BCC (n=16), IEC (n=7), SCC (n=8). The median patient age was 67 with a range of 46-96, matching the population distribution. Males (23) and Females (24) were present in approximately equal proportion. The classification network, referred to as the CNN classifier, was trained for 5 epochs using an Adam optimizer with a learning rate of 1×10^{-4} and batch size of 1 (due to variable input size). Loss was calculated using categorical cross-entropy and training continued until validation loss stopped improving.

For comparison, the CNN classifier was benchmarked against three other classification algorithms. The normalized class counts from the generated segmentations were used as the input features and trained under the same 80:10:10 regime. The classifiers were as follows:

1. Random forest classifier using the default parameters (number of trees = 100) defined in the scikit-learn package (Pedregosa et al., 2011).
2. K-Nearest-Neighbors classifier using the default parameters (scikit-learn).
3. A Single-Layered Perceptron with 64 hidden neurons (relu activations) implemented in Keras.

2.6. Surgical margin clearance

We used the 10x segmentations to analyse the global context of the image and perform surgical margin clearance for 126 excision specimens (Fig. 5 in Section 3.4). Using the scikit-image Python library (Van der Walt et al., 2014), cancer regions (BCC, SCC and IEC) were combined and the distance to the whole tissue margins was calculated (Fig. 5a). Next, the superficial layers (keratin, epidermis, papillary dermis and cancer) were combined and the specimen was orientated by finding a line of best fit through all points in the region, and then the section was rotated to consistently orientate all samples. Further, the intersection of the superficial region perimeter with the entire tissue perimeter identified

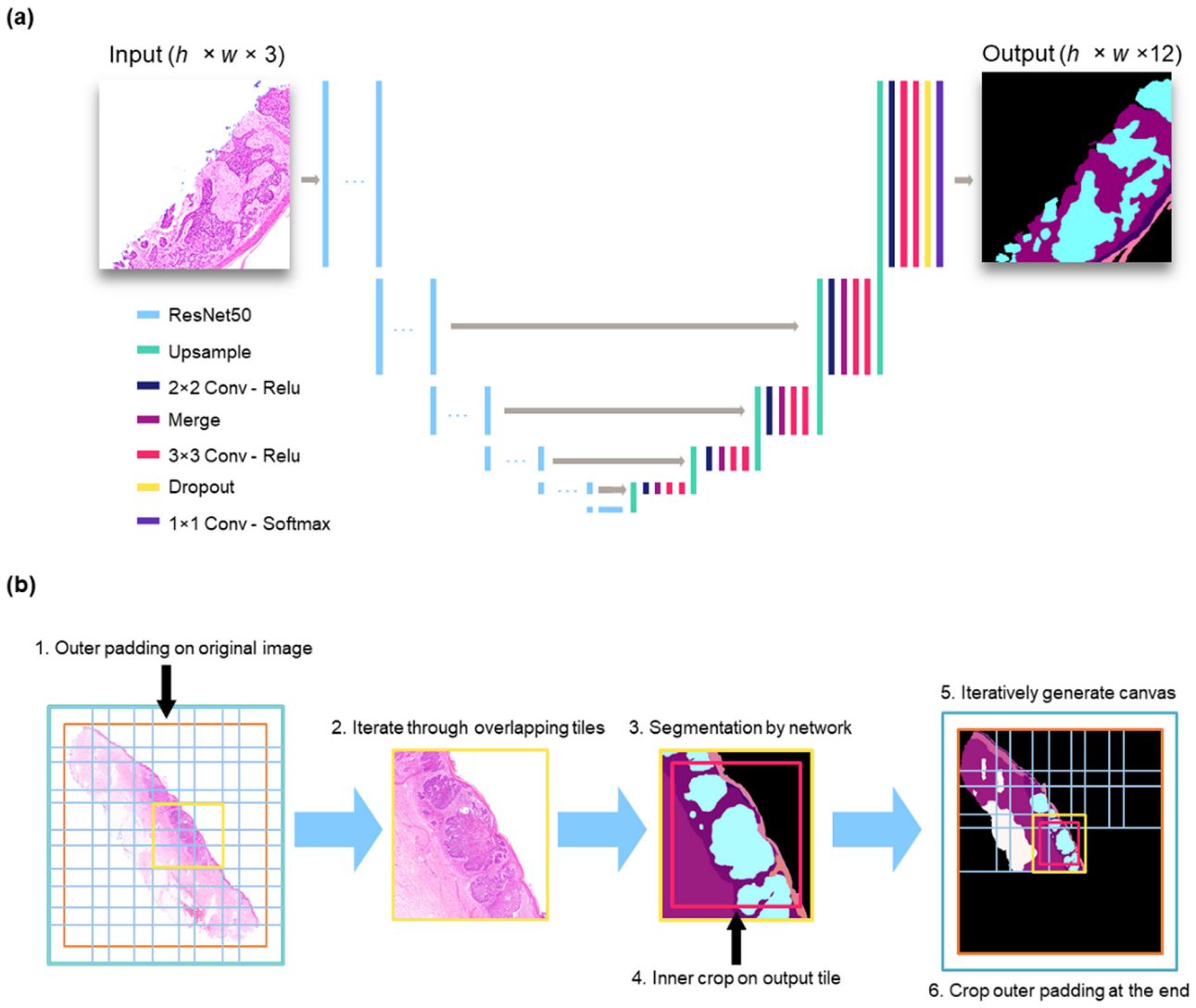

Fig. 1. An overview of the segmentation and classification network architectures. The segmentation network itself provides highly interpretability outputs, which are subsequently used in several downstream analyses such as whole-image classification, specimen orientation and surgical margin clearance (Sections 2.6&3.4). **(a)** A U-Net like architecture attached to ResNet50. The final softmax layer is a h x w x 12 tensor which used in combination with an argmax function along the last axis produced 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. **(b)** The Whole Image Segmentation pipeline. To minimise border effects and improve tile overlap, the image was first padded with white/background pixels. A tile was then fed into the network, where an inner crop was made to further minimize border effects. The segmented tile was added to the canvas which at the end is cropped to the original image size. **(c)** The whole image classification network. The h x w x 12 probability maps were fed into three consecutive 3 × 3 convolutional layers, each with 32 kernels. Global Average Pooling was used in contrast to Global Max Pooling, with the idea being that false positive cancer classes will be averaged out rather than signalled as being of high importance. Further, it seemed a reasonable method of accounting for potentially varied distributions as probability maps from the training, validation and test splits were combined for this task. Finally, two Dense layers were used to make the classification. Best viewed in color.

Table 1
Accuracy and average class accuracy for base and fine-tuned models on the test datasets. Values are the average from three replicates ± the standard deviation. Average class accuracy increased as the image size decreased.

Dataset	Base Accuracy	Base Average Class Accuracy	Fine-tuned Accuracy	Fine-tuned Average Class Accuracy
10x	0.8520 ± 0.0065	0.7373 ± 0.0065	0.8526 ± 0.0012	0.7420 ± 0.0054
5x	0.8472 ± 0.0078	0.7299 ± 0.0056	0.8571 ± 0.0005	0.7396 ± 0.0004
2x	0.8118 ± 0.0020	0.7254 ± 0.0017	0.8120 ± 0.0013	0.7261 ± 0.0013

Table 2
The range in test accuracy (%) across all 12 classes. The classes are keratin (KER), epidermis (EPI), papillary dermis (PAP), reticular dermis (RET), hypodermis (HYP), background (BKG), follicle (FOL), gland (GLD), inflammation (INF), BCC, SCC and IEC. The maximum for each class is bolded. The range is the difference between the minimum and maximum accuracy at different resolutions.

	KER	EPI	PAP	RET	HYP	BKG	FOL	GLD	INF	BCC	SCC	IEC
2x	**87.8**	74.6	77.3	68.5	94.4	**95.1**	52.9	**92.1**	**64**	80.7	75.6	65.2
5x	81.2	**84.9**	78.8	**81.8**	95.5	93.9	**61.6**	87.4	61.7	76.8	78.3	56.5
10x	84.6	83.1	**80.8**	70.2	**96.2**	95	61.5	87.3	57.4	**86.5**	**85.7**	**70.7**
	Range											
	6.6	10.3	3.5	13.3	1.8	1.2	8.7	4.8	6.6	9.7	10.1	14.2

the superficial lateral margins. This also enabled the deep / subcutaneous margin to be found (Fig. 5b). This deep margin was then scanned to find the minimum distance to the cancer. The surgical orientation of the specimen, with respect to the 3 o'clock margin, was found by thresholding the original histology image in the HSV color space for evidence of black ink (Fig. 5c). These methods were then combined to produce a visual summary of the results from automatic surgical margin clearance (Fig. 5d).

3. Results and discussion

3.1. Quantitative assessment of tissue segmentation

Initial experiments were performed using the standard U-Net-like decoder architecture and performance was measured as per-pixel accuracy. During training on the 2x dataset (high resolution), the network achieved a per-pixel accuracy of 86% and an average class accuracy of 85% on the *training data*. The corresponding validation accuracy and average class accuracy were 80% and 73%, respectively. Consequently, the parameters were systematically reduced until a noticeable decrease in overfitting was seen, and ultimately 32 kernels were settled on for all layers of the U-Net. This architecture was then used across the 2x, 5x and 10x datasets. The results on the unseen test sets are shown in Table 1. These results read favourably next to the benchmark of ~86% on the training set.

The highest performance (fine-tuned) was achieved on the 10x dataset, with a per-pixel accuracy and average class accuracy of 85% and 74% respectively. A trend can be seen where average class accuracy increases as the image resolution decreases and it is apparent that there was enough information in the 10x dataset to perform comparably to or better than the 2x dataset. This is unsurprising as it is possible for a human to recognise broadly all 12 classes of skin structures in the 10x images. However, the increase is relatively small, going from 72.61% to 74.2% for the average class accuracy. The accuracies for the 5x and 10x datasets suggest that the difference in performance between these two datasets is negligible. However, a practical consequence of the 10x dataset performing at least as well as the others is that it enables computational costs to be minimized without compromising quality.

Curiously, there is an apparent limitation of 72-74% for the average class test accuracy across all models and datasets. One explanation is that given the small sample size (n=290), the training set does not capture the full class variation. To understand this further, insight can be obtained from Table 2 and the confusion matrices in the appendix Fig. A1a-c. In the first instance, Table 2 compares the accuracies for the three datasets across all 12 classes. Background (BKG) and hypodermis (HYP) tend to have little variation and this is reflected in the similar accuracies of ~95% across all three datasets. Gland (GLD) is another class that seems to be predicted reliably between 87-92%. The two obvious underperforming classes are follicle (FOL) at 61.5% and inflammation (INF) at 64%, and these are skewing the average class accuracy. By excluding these two classes the average class accuracy increases to 84% for the 10x dataset, 81.5% for the 5x dataset and 81.3% for the 2x dataset.

The lower performance on the follicle and inflammation classes may be due to several factors. Humans often distinguish follicles by higher-level context rather than the microscopic structure. In contrast, the high-resolution images mean that a network focuses only on low level information and so is blind to high-level structure. Indeed, Table 2 clearly shows that some classes are more accurately classified at different resolutions and in the case of follicle, the difference between 2x and both 5x and 10x is close to 10%. The feasibility of combining different scales to improve performance is considered briefly in the next section. However, it is also the case that follicles, although prevalent in skin samples generally, are underrepresented in this dataset. Approximately half the specimens were shaves and punches, which due to their small size contain limited variety of skin structures. Further, only single sections were included, rather than whole slides. Under-performance may be a consequence of under-representation.

Considering inflammation, the confusion matrices in the appendix show that it was mostly confused with papillary dermis, BCC, and SCC, each approximately 10% of the time. This is not surprising as inflammation occurs mostly in the upper dermis and is strongly associated with abnormal tissue. Further, the confusion matrices (Fig. A1) conform with each other and show that much of the confusion is between proximal classes in skin layers or cancer classes, which is to be expected, for instance EPI↔KER, EPI↔IEC, RET↔HYP, INF↔PAP, INF↔BCC↔SCC↔IEC.

3.2. Qualitative assessment of tissue segmentation

Visual inspection of the segmentations reiterates the quantitative results, showing a high resemblance to the ground truth labels across the 2x, 5x and 10x datasets (Fig. 2a). In general, there is more noise than in the ground truth because of small-scale structures being incorrectly classified. Some are obvious errors inside larger sections of tissue *e.g.* keratin in the hypodermis. Interestingly, it appears that decisive boundaries have been learned for edges as there tend to be few broken or rough edges.

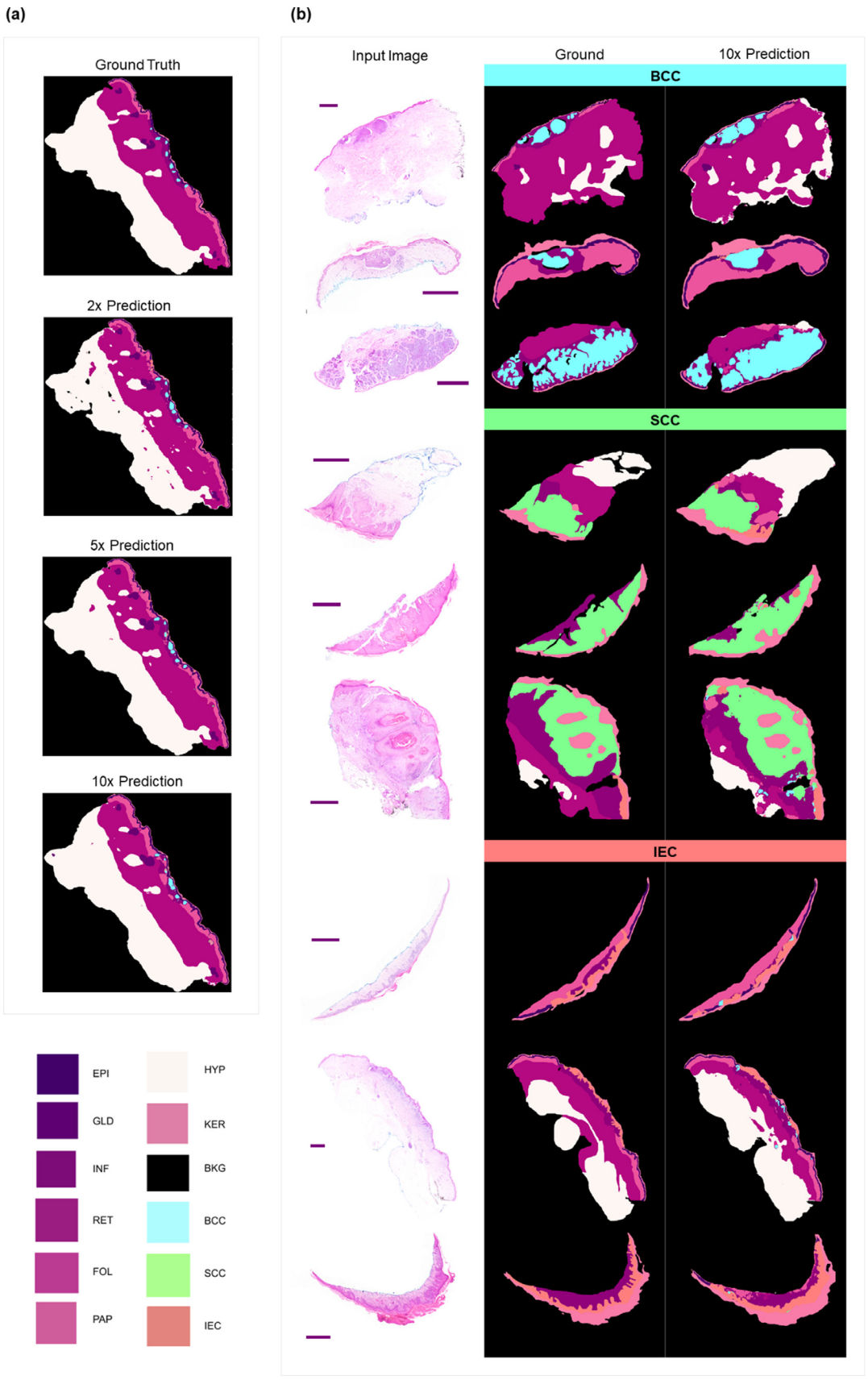

Fig. 2. Semantic segmentation of whole tissue sections. (a) A visual comparison of the segmentations for different image resolutions. As image resolution decreases there was a corresponding decrease in apparent noise. This reduction may account for the 2-4% difference in performance shown in Table 1.**(b)** Whole image segmentations from the 10x network for representative cases of BCC (Excision, Shave, Shave), SCC (Punch, Shave, Excision) and IEC (Shave, Excision, Shave). The purple bar represents 1mm. Color legend (bottom left). Best viewed in color.

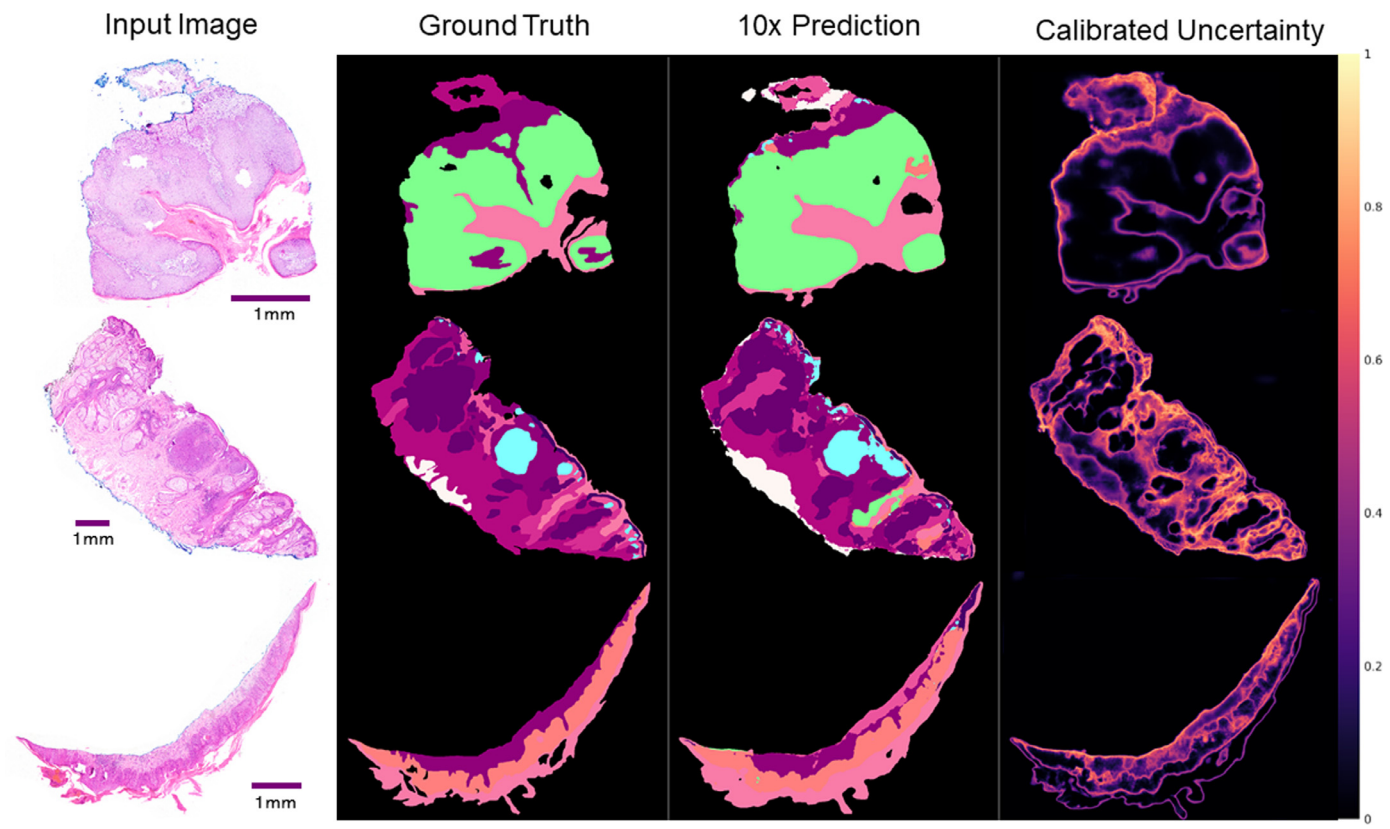

Fig. 3. Visualizing the model uncertainty. Visualizing the uncertainty (1 – maximum class probability) provides a highly interpretable way to evaluate network performance. Presenting calibrated uncertainty maps in a decision-support system could further provide a pathologist a means of locating areas which require more attention, particularly in large multi-section cases. For class colors see color legend in Fig 2.Best viewed in color.

Due to the similarity in quality, and indeed slightly better performance, a variety of segmentations for the 10x dataset are shown in Fig. 2b. Here it is evident that this method works well across the various cancer types and scales for excision, punch and shave biopsies.

To provide insight into the confidence the network had in its segmentations, temperature-scaling was performed to calibrate the output (Guo et al., 2017) (Fig. A2). With a good concordance between confidence and accuracy we then generated maps to visualize model uncertainty, which was given by 1 minus the maximum class probability for each pixel. In effect, the uncertainty maps show which parts of the image the network found challenging (Fig. 3). We note that uncertainty in this context is separate to its use in a Bayesian sense, where variance is the measure of uncertainty. It was found that areas of uncertainty often correspond with the underperforming follicle and inflammation classes (Table 2). The maps also reveal that borders between regions had reduced confidence to between 0.3 and 0.7, which is much greater than 0.083 for a random 12 class prediction. We also note from Fig. 3 that the high uncertainty regions (>0.7) correspond with complex regions involving the presence of BCC, inflammation, glands and hair follicles, where the 10x resolution is less able to clearly distinguish them. It is likely that the 2x and 5x networks could provide needed detail in cases of ambiguity and could therefore expect that improved performance could be achieved by combining the three segmentation scales. However, this comes at a significant computational cost. Given that such cases are a small proportion of this dataset, it is also possible that this limitation can be overcome with more data.

3.3. Whole image classification

High performance can be achieved for whole image classification tasks using the 10x dataset probability maps as learned representations. The CNN classifier achieved an overall classification accuracy across Healthy, BCC, SCC and IEC classes of 93.6%, outperforming the Random Forest at 87.2%. By truncating the results to a cancer versus non-cancer classification, the CNN and Random Forest performed comparably, each with an overall accuracy of 97.9%. The Nearest-Neighbor and Single-Layer Perceptron classifiers scored 80.9% and 85.1% for the four-class problem, and 85.1% and 89.4% for binary classification, respectively.

Focusing on the results of the CNN classifier, of the BCC images 1/16 were classified as IEC and 1/16 as Healthy. 1/7 of IEC were classified as BCC and all SCC images were classified correctly. A comparative performance of these classes is shown in the confusion matrix in Fig. A3. One explanation for the confusion between the IEC and BCC classes relates back to the segmentation network, which similarly shows confusion (~6%) between the two classes (Fig. A1). Indeed, these two categories sometimes require closer scrutiny by the pathologist to distinguish compared to the other categories included.

Given that the Random Forest and CNN have comparable performance for binary classification, a factor to consider further is that of model interpretability. Random forests can measure feature importance by way of ranking the features according to how they (on average) decrease the objective function. For our model this technique revealed that BCC, SCC, IEC, KER, INF and EPI were important for classification (Fig A4). Importantly, these features are the same explanatory variables a pathologist uses for diagnosis.

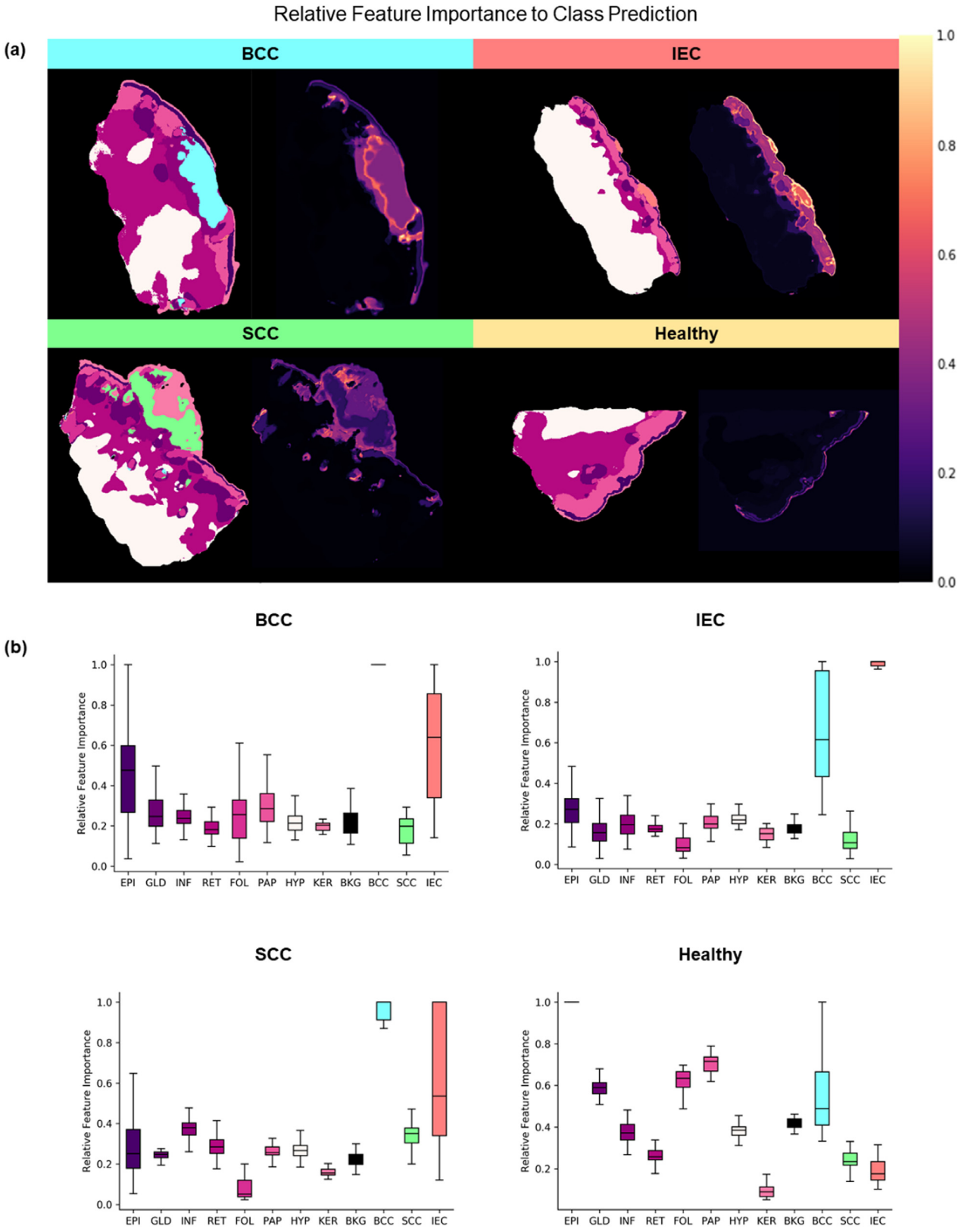

Fig. 4. Measuring the relative importance of each class in the CNN classifier. (a) Backpropagating the gradients to the input image reveals the diagnostically important regions of the image (right side of panels, scale at right) and hence reveals what features (tissue classes) are important for a given class prediction. **(b)** Measuring the relative feature importance across the whole dataset reveals that EPI, BCC, SCC and IEC are the most important classes. These are the same features a pathologist would use when diagnosing the specimen. The importance was calculated by firstly finding the gradient of the class prediction with respect to the input probability map (not the feature map as used in Selvaraju et al. (2017)), resulting in a $h \times w \times 12$ gradient tensor. The mean absolute value was then calculated on a per-pixel basis and binned according to the predicted class from the probability map. The values were normalized for each image and so reflect relative feature importance. A sanity check for feature importance is shown in Fig. A5.

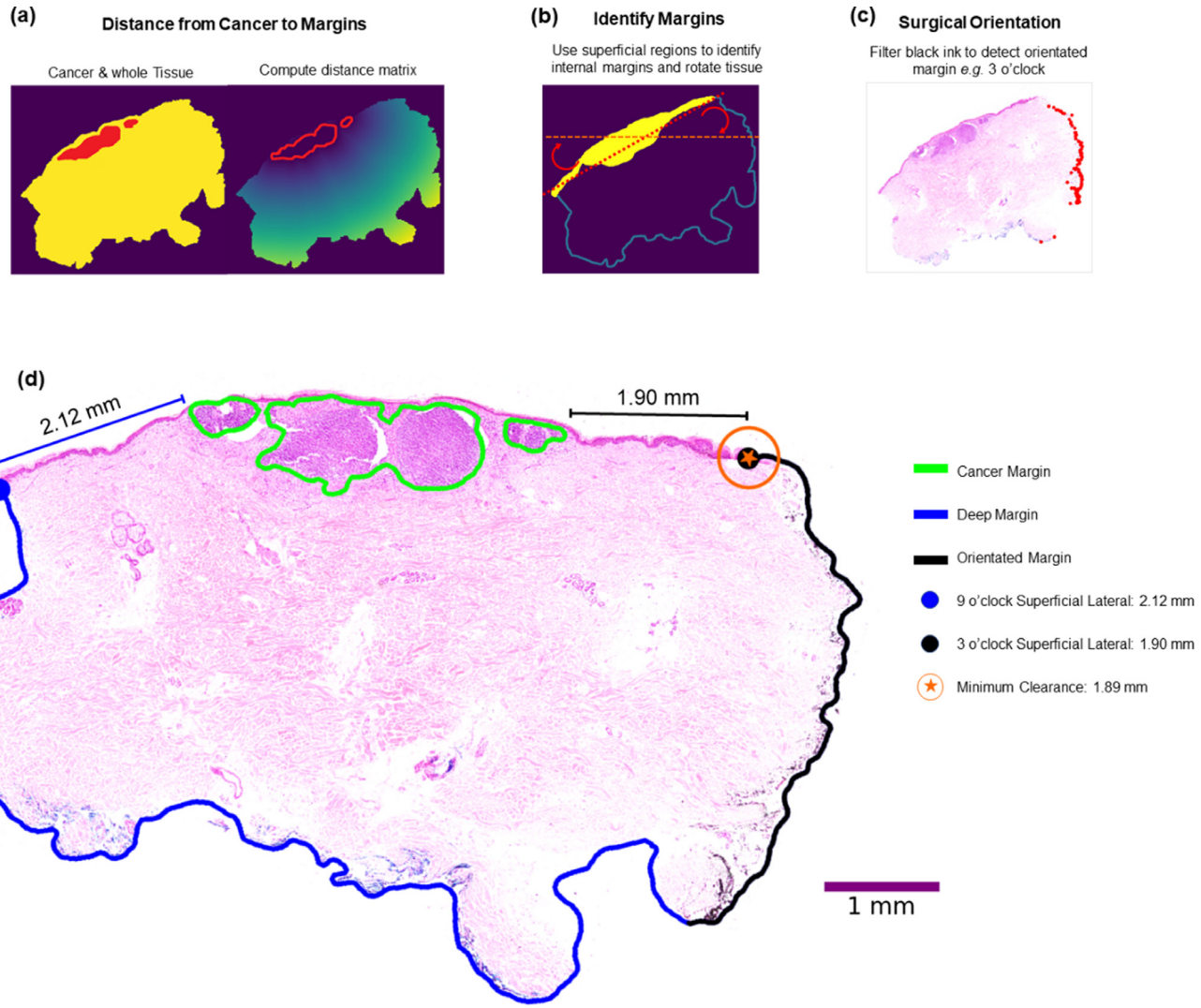

Fig. 5. The automatic surgical margin clearance pipeline performed using the 10x downsample dataset. **(a)** Cancer regions were combined and then a distance matrix was computed to the margins of the whole tissue. **(b)** The superficial regions (epidermis, keratin, papillary dermis and cancer) allowed the specimen to be rotated to the 0°/180° plane. The perimeters of these regions also allowed the internal margins to be located. **(c)** The surgical orientation was found by filtering the original histology image in the HSV color space for black ink. **(d)** An example of how the information from automatic surgical margin clearance assessment could be presented in a computer-aided-diagnosis system. The 3 and 9 o'clock superficial lateral margins have been identified, with the 3 o'clock identified in black. The distances to the nearest cancer region are shown alongside the predicted cancer regions. With no cancer found on the internal margin, the section could be labelled as "margins clear" within a specified distance. The result was generated using the predicted segmentation. Best viewed in color.

Despite conforming to expectations, such an approach is limited to the average case and does not provide insight to what features in a particular image are important for a given prediction. Using the CNN model, we can back-propagate gradients to the input image (see Selvaraju et al., 2017) creating a spatial-map to highlight the importance of each feature (Fig. 4a). Visualising the gradients in this way provides a means of showing directly what relative importance the network gives to each feature. Moreover, comparing the gradients for each class across the whole dataset reveals that the CNN, like the random forest, uses the most clinically relevant features for distinguishing each class (Fig. 4b). Notably, EPI is reliably important for Healthy predictions, as are BCC and SCC for their respective cancers. SCC has increased importance for its class, however, it appears the network is sensitive to the presence of BCC and IEC pixels in the image which are likely false positives, as seen in Fig. 2b.

We think that this approach highlights the value in summarising highly complex image features to high-level descriptions of clinical meaning. The segmentations enable accurate classification and provide a means to create and test parsimonious explanations for the completeness of the learned representations and behaviour of the classifier.

3.4. Surgical margin clearance

The quality of the segmentations provided an opportunity to perform automatic surgical margin clearance calculations. Fig. 5 shows an example of using classical image processing techniques to detect and measure distances to the surgical margins. We compared the algorithm for detecting the 3 and 9 o'clock superficial lateral margins to human ability on the 10x images, for a total of 252 margins. The median error was 0.081mm, and 90% of the predictions were within 0.27mm of the true margin. Significant errors occurred where the assumptions of the algorithm break down. For example, one assumption is that the margins always intersect superficial layers such as keratin and epidermis. However, in cases where the epidermis has been lost, either from ulceration or artefacts from tissue processing, the performance is poor with an error between 1mm and 3.5mm. Importantly, the same case given to a pathologist would require them to *infer* where the true

margin is, relying on information outside the image. Such cases constituted 10 out of the 252 margins, and although rare, are an expected occurrence. It would be possible to design an algorithm to flag such cases using the segmentation information.

With the ability to reliably predict the margins, distance measurements to the cancer allowed the robustness of automatic margin clearance to be assessed. Assuming perfect cancer segmentation (using the ground truth mask), the algorithm produced a median error as a proportion of the true distance of 0.965%. In real terms, 90% of the errors were within 0.163 mm of the true margin. However, the same task on the predicted segmentations produced a median error of 8.36%. Clearly the ability to perform this task successfully is dependent on the quality of the segmentation. It was found that sometimes the minimum clearance estimate was zero, resulting from false positive cancer predictions along the deep margin, possibly due to the presence of ink or debris. This limitation could be overcome by training further on the deep margin regions, which currently constitute a small portion of the training set. Alternatively, it might be possible to flag such predictions on uncertainty, or context such as size and distance from the main cancer body.

Encouragingly, these results demonstrate that this method can provide robust surgical margin clearance which in theory could extend to perform other routine tasks. Indeed, the depth of invasion could be calculated given the current segmentations. Further, if the number of classes were extended to include vascular and nervous tissue, then distances from the cancer to those regions could be calculated easily. This would provide decision support / automatic assessment of lympho-vascular and perineural invasion.

4. Conclusion

Our results show that BCC, SCC and IEC, which constitute over 90% of skin cancer diagnoses, are amenable to deep learning methods. We have demonstrated that the full-visual context of high-resolution histological images can be characterised using semantic segmentation. By framing this problem in a multi-class setting, a network learned high-level representations of the image content which correspond to human-meaningful concepts. The outputs of the network are therefore naturally interpretable and can be visualised in several ways to discern its capabilities. This provides robustness and overcomes limitations in other studies where representations are not constrained. Treating the segmentation network outputs as learned representations, they were then used to classify the whole image with high accuracy. Importantly, the segmentations could be manipulated in a logical way to perform several routine pathologist tasks such as automatic surgical margin clearance assessments. This method could be easily extended to perform other essential tasks such as assessing degree of differentiation and hence aggressiveness, depth of invasion, direction of growth, and assessing for the presence of lymphovascular and perineural invasion. Further, this method is applicable to other skin diseases, and given such potential, may justify the effort and expense of acquiring densely labelled data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Simon M. Thomas: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **James G. Lefevre:** Conceptualization, Supervision, Writing - review & editing. **Glenn Baxter:** Conceptualization, Investigation, Resources, Supervision, Writing - review & editing. **Nicholas A. Hamilton:** Conceptualization, Supervision, Writing - review & editing, Resources, Project administration.

Acknowledgements

We wish to acknowledge The University of Queensland's Research Computing Centre (RCC) for its support in this research. JL is supported by Australian Research Council Discovery Grant DP180101910. MyLab Pathology provided access to their archived histological collection. Financial support was provided by The Australian Government Research Training Program (RTP) and The Laurel Joy George Perpetual Scholarship.

Appendices

Fig. A1, Fig. A2, Fig. A3, Fig. A4, Fig. A5

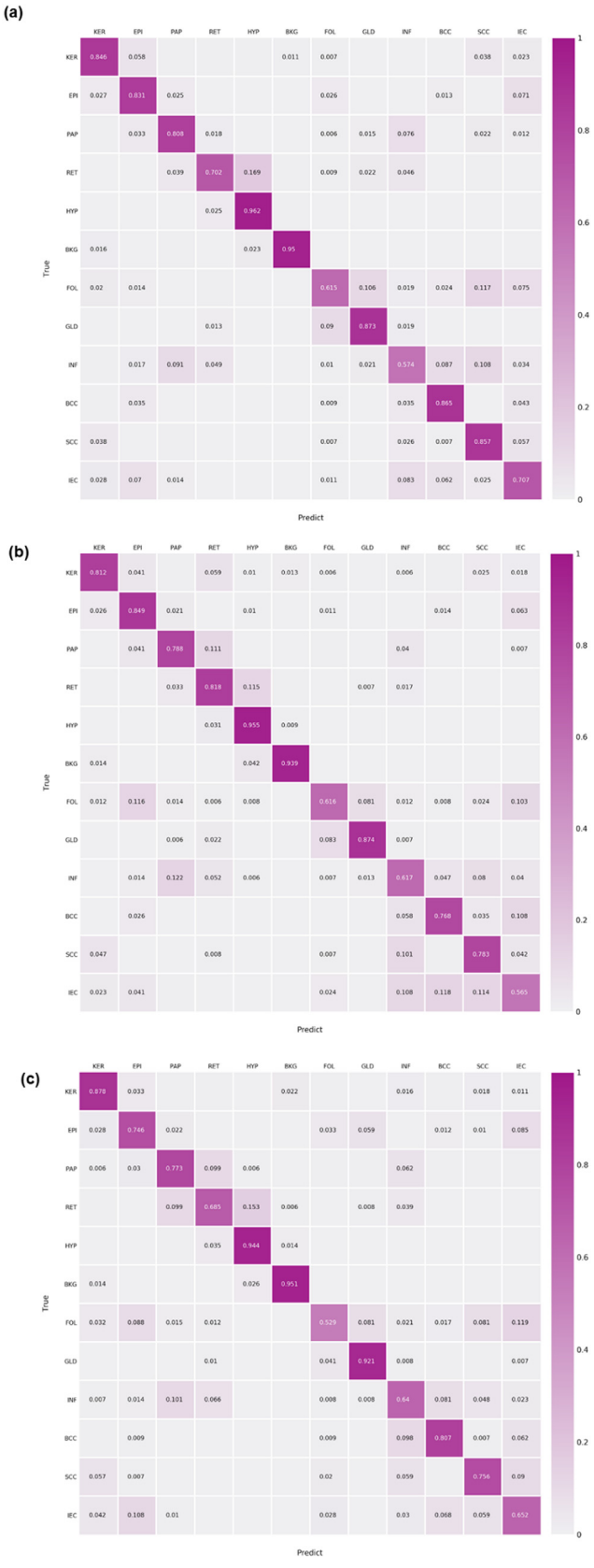

Fig. A1. Confusion matrix of test scores for 12 class segmentation across the three datasets. **(a)** 10x dataset, **(b)** 5x dataset, **(c)** 2x dataset. Values are read across columns so that the diagonal describes recall (true positive rate). The classes are keratin (KER), epidermis (EPI), papillary dermis (PAP), reticular dermis (RET), hypodermis (HYP), background (BKG), follicle (FOL), gland (GLD), inflammation (INF), BCC, SCC and IEC.

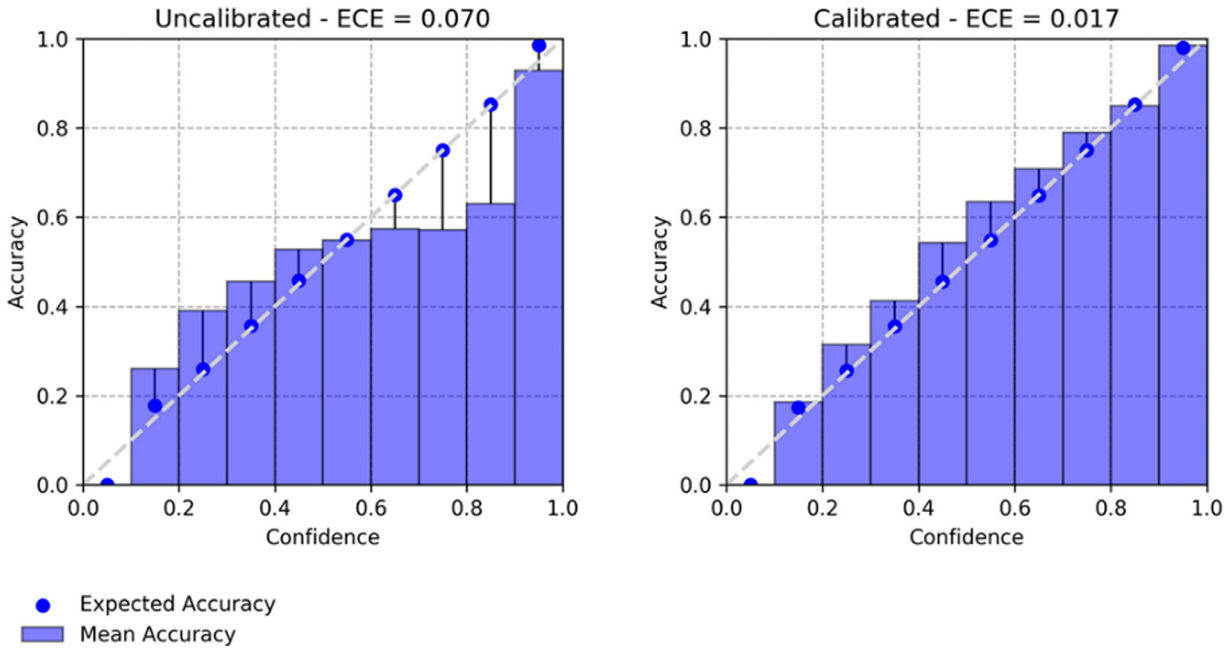

Fig. A2. Calibration plots for the 10x segmentation network. There should be a strong concordance between expected accuracy (how confident the network is) and accuracy. (**Left**) Uncalibrated network shows over and under-confidence. (**Right**) A well-calibrated network after temperature scaling, reflected by the reduced expected calibration error (ECE) and improved conformity to the "ideal" accuracy line.

Fig. A3. Multi-class whole image classification. The whole image classification was performed using the 12-class probability matrices from the 10x dataset. Images are classified as Healthy (n=16), BCC (n=16), IEC (n=7), SCC (n=8). Values are read across columns so that the diagonal describes recall (true positive rate).

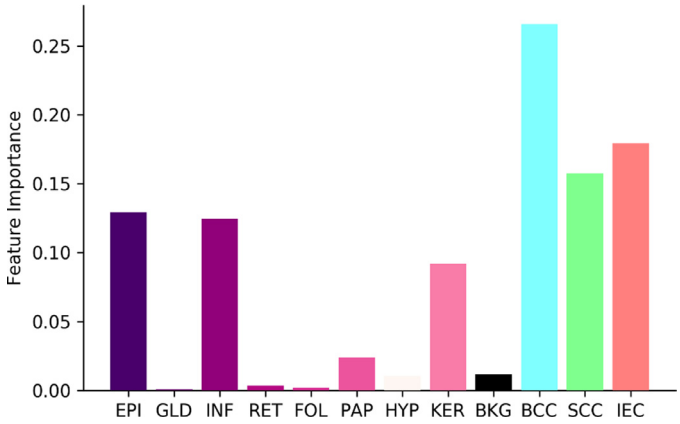

Fig. A4. Average feature importance from the random forest classifier. Feature importance is measured by ranking the features according to how they (on average) decrease the objective function, which by default is *gini impurity* in scikit-learn. The classifier reliably uses cancer tissue as class predictors, followed by epidermis, inflammation and keratin, all of which are informative pathologically.

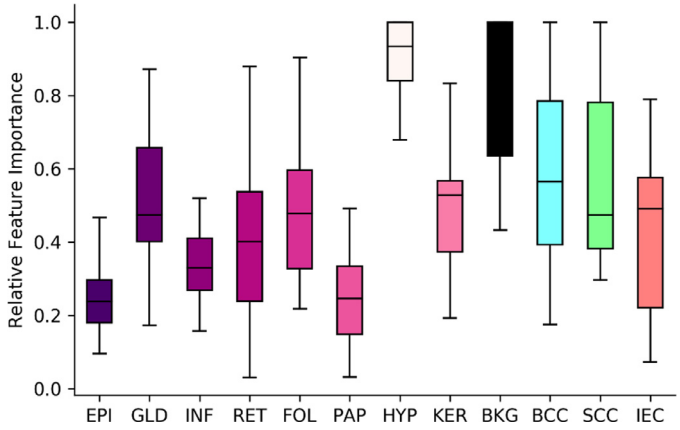

Fig. A5. Relative feature importance for a randomly initialized classifier as a sanity check. Using random weights disrupts the expectations we have for a trained network. In this case, we see an increased importance of HYP, BKG, FOL, GLD which themselves are not informative pathologically. Indeed, classes such as EPI and INF, considered important in the trained network (Fig. 4b) are now among the least important. The importance was calculated by firstly finding the gradient of the class prediction with respect to the input probability map (10x downsample dataset), resulting in a $h \times w \times 12$ gradient tensor. The mean absolute value was then calculated on a per-pixel basis and binned according to the predicted class from the probability map. The values were normalized for each image and so reflect relative feature importance across all classes in the training set.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: A system for large-scale machine learning, in: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. OSDI 2016 16, 265–283.

Acs, B., Rimm, D.L., 2018. Not just digital pathology, intelligent digital pathology. JAMA Oncol. 4, 403–404.

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

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

Australian Institute of Health and Welfare, 2016. Skin cancer in Australia. AIHW, Canberra.

Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P.E., Verrill, C., Walliander, M., Lundin, M., Haglund, C., Lundin, J., 2018. Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci. Rep. 8. https://doi.org/10.1038/s41598-018-21758-3.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25. https://doi.org/10.1038/s41591-019-0508-1.

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., Olah, C., 2019. Activation atlas. Distill 4, e15.

Chen, C., Li, O., Tao, C., Barnett, A.J., Su, J., Rudin, C., 2018. This looks like that: deep learning for interpretable image recognition. arXivPrepr. arXiv1806.10574.

Chollet, F., 2015. Keras - GitHub repository.

de Geus, D., Meletis, P., Dubbelman, G., 2018. Panoptic segmentation with a joint semantic and instance segmentation network.arXivPrepr. arXiv1809.02110.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118. https://doi.org/10.1038/nature21056.

Estienne, T., Vakalopoulou, M., Christodoulidis, S., Battistela, E., Lerousseau, M., Carre, A., Klausner, G., Sun, R., Robert, C., Mougiakakou, S., 2019. U-ReSNet: Ultimate coupling of registration and segmentation with deep nets. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 310–319.

Faust, K., Bala, S., van Ommeren, R., Portante, A., Al Qawahmed, R., Djuric, U., Diamandis, P., 2019. Intelligent feature engineering and ontological mapping of brain tumourhistomorphologies by deep learning. Nat. Mach. Intell. 1, 316.

Fischer, A.H., Jacobson, K.A., Rose, J., Zeller, R., 2008. Hematoxylin and eosin staining of tissue and cell sections. CSH Protoc. 2008 db.prot4986.

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

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks, in: 34th International Conference on Machine Learning. ICML 2017, 2130–2143.

Halicek, M., Lu, G., Little, J.V, Wang, X., Patel, M., Griffith, C.C., El-Deiry, M.W., Chen, A.Y., Fei, B., 2017. Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. J. Biomed. Opt. 22, 60503.

Halicek, M., Shahedi, M., Little, J.V, Chen, A.Y., Myers, L.L., Sumer, B.D., Fei, B., 2019. Detection of squamous cell carcinoma in digitized histological images from the head and neck using convolutional neural networks. Medical Imaging 2019: Digital Pathology. International Society for Optics and Photonics.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hegde, N., Hipp, J.D., Liu, Y., Emmert-Buck, M., Reif, E., Smilkov, D., Terry, M., Cai, C.J., Amin, M.B., Mermel, C.H., 2019. Similar image search for histopathology: SMILY. npj Digit. Med. 2, 56.

Hekler, A., Utikal, J.S., Enk, A.H., Berking, C., Klode, J., Schadendorf, D., Jansen, P., Franklin, C., Holland-Letz, T., Krahl, D., 2019. Pathologist-level classification of histopathological melanoma images with deep neural networks. Eur. J. Cancer 115, 79–83.

Jiang, Y.Q., Xiong, J.H., Li, H.Y., Yang, X.H., Yu, W.T., Gao, M., Zhao, X., Ma, Y.P., Zhang, W., Guan, Y.F., Gu, H., Sun, J.F., 2020. Recognizing basal cell carcinoma on smartphone-captured digital histopathology images with a deep neural network. Br. J. Dermatol. 182, 754–762. https://doi.org/10.1111/bjd.18026.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., 2018. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning, pp. 2673–2682.

Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit. Heal. 1, e271–e297.

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

Lundberg, G.D., 2019. How many pathologists does the United States need? JAMA Netw. Open2, e194308 –e194308.

Metter, D.M., Colgan, T.J., Leung, S.T., Timmons, C.F., Park, J.Y., 2019. Trends in the US and Canadian Pathologist Workforces From 2007 to 2017. JAMA Netw. Open2, e194337 –e194337.

Niazi, M.K.K., Parwani, A.V., Gurcan, M.N., 2019. Digital pathology and artificial intelligence. Lancet Oncol.. https://doi.org/10.1016/S1470-2045(19)30154-8.

Olah, C., Mordvintsev, A., Schubert, L., 2017. Feature visualization. Distill 2, e7.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A., 2018. The building blocks of interpretability. Distill 3, e10.

Patterson, J.W., 2016. Tumors of the epidermis. In: Weedon's Skin Pathology. Elsevier, pp. 783–835 e29https://doi.org/http://dx.doi.org/10.1016/B978-0-7020-5183-8.00031-X.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Rad, R.M., Saeedi, P., Au, J., Havelock, J., 2020. Trophectoderm segmentation in human embryo images via inceptioned U-Net. Med. Image Anal. 62, 101612.

Rogers, H.W., Weinstock, M.A., Feldman, S.R., Coldiron, B.M., 2015. Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the U.S. Population, 2012. JAMA Dermatol.151 1081–1086. https://doi.org/10.1001/jamadermatol.2015.1187.

Ronneberger, O., Fischer, P., Brox, T., 2015. 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). Springer Verlag, pp. 234–241.

Rosado, B., Menzies, S., Harbauer, A., Pehamberger, H., Wolff, K., Binder, M., Kittler, H., 2003. Accuracy of computer diagnosis of melanoma: a quantitative meta-analysis. Arch. Dermatol. 139, 361–367.

Royal College of Pathologists of Australia, 2018. Australian Pathologist Workforce Study. RCPA 2018, 1–9.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Saldanha, G., Fletcher, A., Slater, D.N., 2003. Basal cell carcinoma: a dermatopathological and molecular biological update. Br. J. Dermatol. 148, 195–202. https://doi.org/10.1046/j.1365-2133.2003.05151.x.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.

Slater, D., Barrett, P., 2019. Dataset for histopathological reporting of primary invasive cutaneous squamous cell carcinoma and regional lymph nodes. Royal College of Pathologists, London.

Staples, M.P., Elwood, M., Burton, R.C., Williams, J.L., Marks, R., Giles, G.G., 2006. Non-melanoma skin cancer in Australia: the 2002 national survey and trends since 1985. Med. J. Aust. 184, 6–10.

Titford, M., 2006. A short history of histopathology technique. J. Histotechnol. 29, 99–110. https://doi.org/10.1179/his.2006.29.2.99.

Tizhoosh, H.R., Pantanowitz, L., 2018. Artificial intelligence and digital pathology: challenges and opportunities. J. Pathol. Inform. 9, 38. https://doi.org/10.4103/jpi.jpi_53_18.

Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., 2014. scikit-image: image processing in Python. PeerJ 2, e453.

Zhang, Q., Cui, Z., Niu, X., Geng, S., Qiao, Y., 2017. Image segmentation with pyramid dilated convolution based on ResNet and U-Net. In: International Conference on Neural Information Processing. Springer, pp. 364–372.

Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., Ahmad, N., Khalil, F.K., Dickinson, S.I., Shi, X., Liu, F., Su, H., Cai, J., Yang, L., 2019. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat. Mach. Intell. 1, 289. –289 https://doi.org/10.1038/s42256-019-0062-z.

Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L., 2017. Mdnet: A semantically and visually interpretable medical image diagnosis network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6428–6436.