scientific reports

Check for updates

OPEN

A Modified Deep Semantic Segmentation Model for Analysis of Whole Slide Skin Images

Muhammad Zeeshan Asaf[1,2], Hamid Rasul[1,2], Muhammad Usman Akram[1]✉, Tazeen Hina[1], Tayyab Rashid[1] & Arslan Shaukat[1]

Automated segmentation of biomedical image has been recognized as an important step in computer-aided diagnosis systems for detection of abnormalities. Despite its importance, the segmentation process remains an open challenge due to variations in color, texture, shape diversity and boundaries. Semantic segmentation often requires deeper neural networks to achieve higher accuracy, making the segmentation model more complex and slower. Due to the need to process a large number of biomedical images, more efficient and cheaper image processing techniques for accurate segmentation are needed. In this article, we present a modified deep semantic segmentation model that utilizes the backbone of EfficientNet-B3 along with UNet for reliable segmentation. We trained our model on Non-melanoma skin cancer segmentation for histopathology dataset to divide the image in 12 different classes for segmentation. Our method outperforms the existing literature with an increase in average class accuracy from 79 to 83%. Our approach also shows an increase in overall accuracy from 85 to 94%.

Keywords Whole slide image segmentation, Semantic segmentation, U-Net, EfficientNet-B3, Ensemble model

Image segmentation is partitioning of an image into different clusters on basis of distinctive features. It is done to better explain the global context of an image[1]. Segmentation separates all of the objects in an image based on their instances, where each instance belongs to a class. Image segmentation has many types, but research shows that for medical images, semantic segmentation has proved to be a successful approach for image analysis and understanding. Semantic Segmentation can be used to differentiate various parts of the tissues and help in highlighting different anomalies.

Medical imaging plays an important role in disease diagnosis, treatment planning and clinical monitoring[2]. Semantic segmentation is mostly used to segment medical images on the basis of different features and its goal is to label each pixel of an image with the corresponding class of the objects. Semantic segmentation of clinically relevant structures in different biopsy images plays a key role in automated diagnosis systems. Deep learning based algorithms learn from data to distinguish objects between different classes in images. The evolution of semantic segmentation started with the advent of computer vision applications. Depending on the quality of data, type of model output, learning techniques, etc., different deep learning models are used in computational pathology. This procedure is gaining increased attention as it reduces the labor required by different specialists and the algorithms also helps in processing a framework that achieves consistent results across different labs for better investigation of various abnormalities. Currently, the incidences and prevalence of skin cancer are increasing worldwide. Early detection, appropriate treatment, and prevention of recurrence are major challenges for researchers today[3]. A biopsy is the only way to confirm a skin cancer diagnosis. A skin biopsy is a procedure that involves removing a small piece of skin tissue and examining it under a microscope to diagnose skin conditions and diseases. But determining precise skin disease from brightfield microscope images using manual technique requires considerable experience, time, complex screening and could still be erroneous. Whereas automated methods face many other challenges like presence of hair, inconspicuous lesion margins, low contrast on dermoscopic images, and variability in skin lesion color, texture, and shape[4]. Advanced computation and optimized code can be used to extract some meaningful information from brightfield images that may not be readily perceived by humans[5]. The tissue extracted during a biopsy undergo investigation after the process of staining, where different anomalies

[1]Department of Computer and Software Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan. [2]These authors contributed equally: Muhammad Zeeshan Asaf and Hamid Rasul. ✉email: usmakram@gmail.com

are investigated. This investigation is again influenced by the type of processing it has received, leaving a lot of ambiguities even after quite a laborious work. It has been highlighted that due to these challenges, pathologists disagree on up to 60% of cases[6]. This indicates that research in related computations and segmentation algorithms to help in quick and uniformed outcome for easy diagnosis of different anomalies present in the tissue is needed.

The objective of this article is to address the problems of current computational histopathology works, which include subpar accuracy, complications in feature extraction, and the unsatisfactory performance of different techniques. Our solution involves utilizing EfficientNet-B3, a proficient encoder architecture, to allocate the image into layers and designate each layer to a class. This methodology resulted in improved outcomes with enhanced accuracy and reliability. Furthermore, our approach provides an upgraded structure for feature extraction and facilitates smoother predictions. We also scrutinized the efficiency of various backbones on a U-Net for semantic segmentation duties and achieved a 95% accuracy, surpassing the inadequacies of prior research.

The rest of the paper is organized for relevant literature survey in "Literature review" section, while methodology and dataset is provided in "Materials and methods" section. "Experiments and results" section represents results and discussion, while conclusive remarks of this study will be provided in "Conclusion and future work" section.

Literature review

Existing work

Early detection of skin cancer can increase five year survival rate of patients from 18 to 98%. Automated diagnosis of skin disease consists segmentation, feature extraction, and its classification. Deep fully convolutional networks have achieved significant success in the task of semantic segmentation[7]. Segmentation has its own importance as feature extraction and classification rely on this part[8]. Adi Wibowo et al.[9] presented a light weight encoder-decoder for segmentation by treating variability with the augmentation. The method shows efficiency in terms of computation. UNet, introduced by Ronneberger et al. in 2015[10], is a deep learning architecture for semantic segmentation, which aims to accurately classify each pixel in an image to a certain category. The UNet architecture consists of an encoder and a decoder part.The encoder part extracts low level image features and downsamples the input[11], while the decoder part upsamples these features to the original image size, and then concatenates them with the corresponding features from the encoder to obtain the final segmentation map. The UNet architecture has shown impressive performance in various biomedical image segmentation tasks and has been widely adopted.

EfficientNet is another deep learning architecture that was proposed in 2019 by Tan and Le in their paper[12]. The EfficientNet architecture achieves state-of-the-art performance on various image classification tasks with much fewer parameters and Floating Point Operations per Second (FLOPS) compared to previous state-of-the-art models, by adopting a novel compound scaling method that efficiently scales up the depth, width, and resolution of the network. The authors also introduced a new scaling parameter called "compound coefficient" that uniformly scales all dimensions of the network based on a single scaling parameter. The EfficientNet architecture has been widely adopted in various computer vision tasks and has set a new standard for model efficiency. Utilizing these, Gouse Mohiddin et al.[13] has offered a theory of preprocessing of images for color consistency, hair removal, noise filtering and edge enhancement. Pre-processed image is then fed to convolutional network to get better results. Fatemah Bagheri et al.[8] proposed a two stage method to get the segmentation and masking separately to address different factors of dermoscopic images. This combination strategy has achieved the Jaccard Coefficient index of 80% for overall lesion segmentation. To address the boundry accuracy in leison segmentation, Lituan Wang et al.[14] has proposed Deep edge convolutional neural networks based on an encoder-decoder structure to focus more on the skin lesion boundary information. An edge information guided module is designed to introduce more information about the boundary. A new loss function including full loss, center loss and edge loss is proposed to pay more attention to boundary optimization.

Different researchers have worked for the purpose of skin segmentation with different ideas. Some of them have used complete image whereas others have argued that to get more precision, it is better to find the respective region of interests on the image and apply Deep Learning models on that specific region rather than using it for complete image. Hao Zheng et al.[7] has provided representative captions as an alternative for better image captioning. The proposed method relies on an unsupervised network that extracts features by directly targeting critical instances in the image followed by a fully connected trained supervised network to segment the image. The segmentation method selects the representative human annotations with reduced inter- and intra-cluster redundancy. Nooshin Moradi et al.[5] has put forward a multi class image segmentation instead of binary segmentation based on combining data from different feature spaces to build more informative structure.Two dictionaries are jointly learned using the K-SVD algorithm and then final segmentation is accomplished by a graph-cut method to distinguish background and foreground based on topological information of lesions and the learned dictionaries.

Medical data is scarce and collecting it is a difficult and time consuming process, while their annotation has to be performed by multiple specialists to ensure its validity[15]. Different deep learning models have shown varying degrees of achievement to date which has shown acceptability of the clinicians. Blind evaluation of these results by board-certified pathologists has also demonstrated similarities with gold standard[16]. The grouping of radiology through the application of information technology has led to its digitization. The images are digitally created, stored, rapidly transmitted over long distances, and consulted by medical professionals. New advancements in information technology has brought the possibility of 3D/ 4D in MRI[6] and further adding to the process of rapid diagnosis. Today, images are clearer, more detailed than ever and annotated[15], which allows healthcare teams to ultimately take a better approach towards patient care.

The literature indicates that different algorithms which have been used for the image segmentation purposes and have provided good results. It has also been observed that combination of different algorithms either for pre-processing or post-processing, shows more promising result some extra computation. This factor has remained motivational factor and many segmentation processes are now being implemented with grouping of different algorithms targeting different features in the image for better results. Y Zhang et al.[17] has used an improved inception module in the encoder to efficiently extract and synthesize information from different receptive fields, followed by a new mesh synthesis strategy to gradually refine shallow features and further smooth the semantic gap for brain tumor segmentation. Similarly for breast cancer detection, KB Soulami et al.[18] has used U-Net followed by loss function to improve the accuracy of the model. For segmentation of retinal vessel, H Wu et al.[19] has first proposed a scale-aware feature synthesis module, which aims to dynamically adjust receptive fields to efficiently extract multi-scale features. An adaptive feature matching module is then designed to guide the efficient combination of adjacent hierarchical features to capture more semantic information

Kleczek et al.,[20] developed a method for tissue segmentation in H &E-stained skin specimens, which aimed to separate the foreground (tissue) from the background (slide) of the images. Their method combined statistical analysis, CIELAB color thresholding, and binary morphology for precise tissue segmentation and evaluated results using Jaccard index. Oskal et al.,[21] used a U-net based approach to separate the epidermis from the rest of the whole slide images for diagnostic analysis. The Dataset consisted of 380,000 image patches of size 512 x 512 pixels and was collected from 36 pathology images. The results were evaluated using mean Positive Predictive Value, Sensitivity, Dice Similarity Coefficient and Matthews Correlation Coefficient. Nofallah et al.,[22] proposed a two-stage segmentation pipeline using U-Net to segment skin biopsy images with coarse and sparse annotations. The team trained their model on a small region of the whole slide image, and generate segmentation masks of different skin tissue entities. In the first stage, they segment the image into four classes: stratum corneum, epidermis, dermis, and background. In the second stage, they train two more models and use masks generated from the first stage to remove the epidermis from the dermis input and dermis from epidermis input, the goal of second stage is to reliably segment the classes of dermal nests and epidermal nests. The results were evaluated using mean intersection over union (IoU), Dice coefficient and opinion score from three pathologists. Thomas et al.[23], developed an interpretable deep learning method for automatic diagnosis and analysis of the most common skin cancers. They collected the Non-melanoma skin cancer segmentation for histopathology dataset[24] and used a modified U-Net based architecture with Resnet50 encoder, U-NET like decoder and skip connections to classify histological images of skin tissue into 12 dermatological classes that represented the skin tissue structures and layers. They used mean class accuracy and uncertainty maps to interpret trained model and evaluate network performance, and showed the applicability of their segmentation method for dermatopathological tasks such as measuring surgical margins. Kriegsmann et al.[25] curated a dataset of 16 skin tissue classes and tumors from 386 cases. They trained an EfficientV2 based deep learning model on 129,364 image patches with a resolution of 395 x 395 pixels) for classifying anatomical tissue structures and neoplasms of the skin tissue. The model was evaluated using cross entropy loss, balanced accuracy and Matthews correlation coefficient. Table 1 shows the summary of all relevant articles discussed in this section.

Research gaps

The existing work suffers from several limitations, including inadequate accuracy, difficulties in extracting relevant features, and ineffective performance of different backbones on U-Net. Additionally, lower performance on some cancerous classes needed to be addressed especially with respect to segmentation. Another issue was the interpretation of overlap in patches of Whole Slide Images (WSIs) and the limited information of neighboring patches, which led to the problem of ragged patches.

Contributions

Our work aims to address above mentioned limitations and improve upon them. Our Efficient-Net model builds upon the works of Thomas et al.[24] by training the model to disburse the image with respect to different layers available in that specific image and assign each layer to a class. Overall, by experimenting we are able to:

Paper	Method	Dataset	Evaluation	Application
Kleczek et al.[20]	A combination of statistical analysis, CIELAB color thresholding, and binary morphology	60 high-resolution whole slide images from three laboratories	Jaccard index	Tissue segmentation in H &E-stained skin specimens
Oskal et al.[21]	A U-net based model	59 WSI images from University of British Columbia and 10 from University of Michigan	Mean Positive Predictive Value, Sensitivity, Dice Coefficient, Matthews Correlation Coefficient	Epidermal tissue segmentation in whole slide images
Nofallah et al.[22]	Two-stage segmentation using U-Net	240 H &E stained skin biopsy image slides	Mean intersection over union (IoU), Dice coefficient and opinion score from pathologists	Skin biopsy images segmentation with coarse and sparse annotations
Thomas et al.[23]	A modified U-Net with Resnet50 encoder and U-NET like decoder with skip connections	290 H &E images and their segmentation masks highlighting 12 classes	Mean class accuracy, Uncertainty maps	Skin tissue structures and diagnosis of non-melanoma skin cancers
Kriegsmann et al.[25]	EfficientV2 based classification learning model	A dataset of 16 classes of skin tissue and tumors from 386 cases	Cross entropy loss, accuracy, Mathew's correlation coefficient	Detection of anatomical tissue structures and neoplasms of the skin

Table 1. Comparative analysis of machine learning methods for skin tissue segmentation.

- Analyze different backbones on a UNet, resulting in an efficient and powerful encoder architecture based on EfficientNet-B3 and Ensemble model.
- Address limitations by training the model to disburse the image into layers and assign each layer to a class, leading to smoother predictions. Furthermore, uncertainty maps were used to show reliability of segmentation done by model.

Materials and methods

Dataset

The primary dataset utilized in this research was the Histopathology Non-Melanoma Skin Cancer Segmentation dataset, provided by Queensland University[24]. The dataset consists of 290 high-resolution images of skin cancer specimens, which were down sampled at various factors, including 1×, 2×, 5×, and 10×. The dataset includes three cancer classes: Basal Cell Carcinoma (*BCC*), Squamous Cell Carcinoma (*SCC*), and Intra-Epidermal Carcinoma (*IEC*), with a total of 290 images. The specimens were obtained from patients aged between 34 and 96 years, with a median age of 70 years. The gender distribution in the dataset was 2/3 female and 1/3 male. The ground-truth segmentation was created by color-coding the images into 12 classification categories, as shown in Fig. 1, including 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*. To determine the optimal down-sampling factor, the dataset was subjected to a series of experiments, which showed that the 5× and 10× down-sampling factors had minimal impact on the segmentation performance. Therefore, the 10× down-sampling factor was used in this study to ensure that the segmentation results were representative of the underlying data and that the performance of the method was not affected by external factors. Thus the Histopathology Non-Melanoma Skin Cancer Segmentation dataset, combined with the selection of an appropriate down-sampling factor, provided a reliable and rigorous foundation for the investigation of the proposed segmentation method in this research.

Methodology

Figure 2 presents the proposed model for the task at hand along with the various data processing steps involved. The figure illustrates the different stages of data processing involved in the model. The first step is pre-processing, where the input images are subjected to a series of preprocessing steps to prepare them for the CNN model. In this stage, the input images are first resized to patches of size 256×256. This is done to reduce the complexity of the model and to enable faster processing. After pre-processing, the patches are then fed to a CNN model for segmentation. The CNN model uses a set of convolutional layers to extract features from the input patches, followed by pooling layers to downsample the features and reduce the computational cost. The output of the CNN model is a segmented image that highlights the areas of interest. Finally, in the post-processing stage, the segmented patches are stitched together to form the final segmented image. This is done to obtain a complete and coherent representation of the input image. The post-processing step also involves the removal of any small artifacts or noise that may be present in the segmented image.

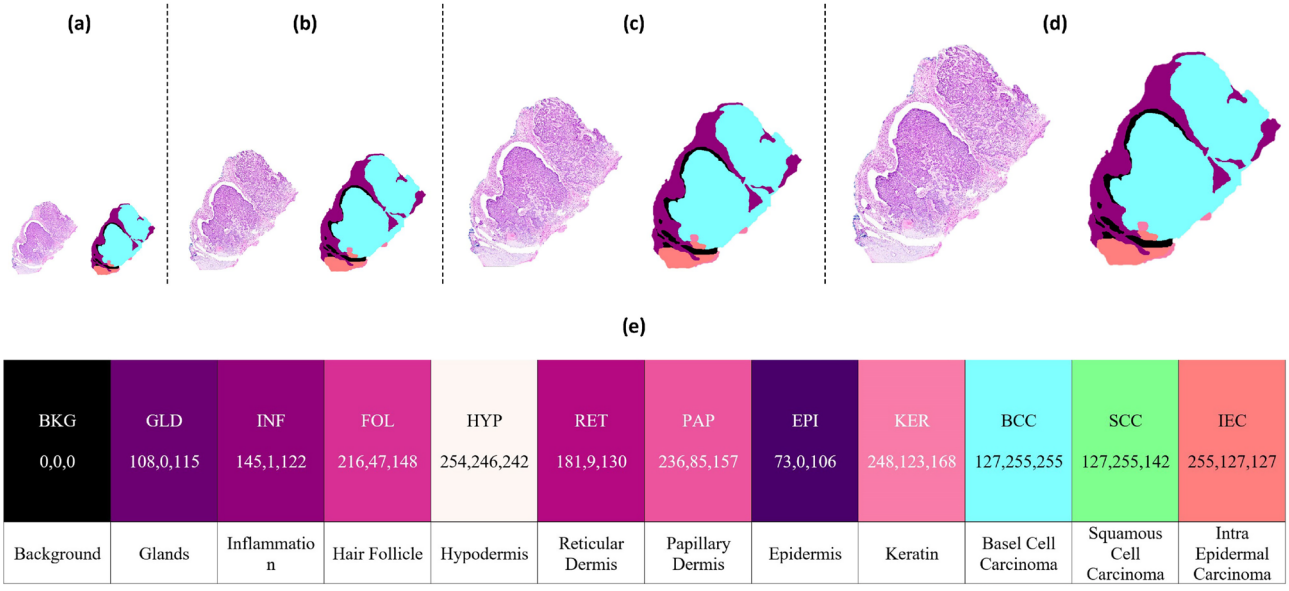


Figure 1. Skin tissue image with different levels of magnification. Image sample with its mask, downsampled at (**a**) 10×, (**b**) 5**x**, (**c**) 2**x**, (**d**) 1**x**, (**e**) shows the color palette for the mask, with the individual RGB values for different classes in the tissue.

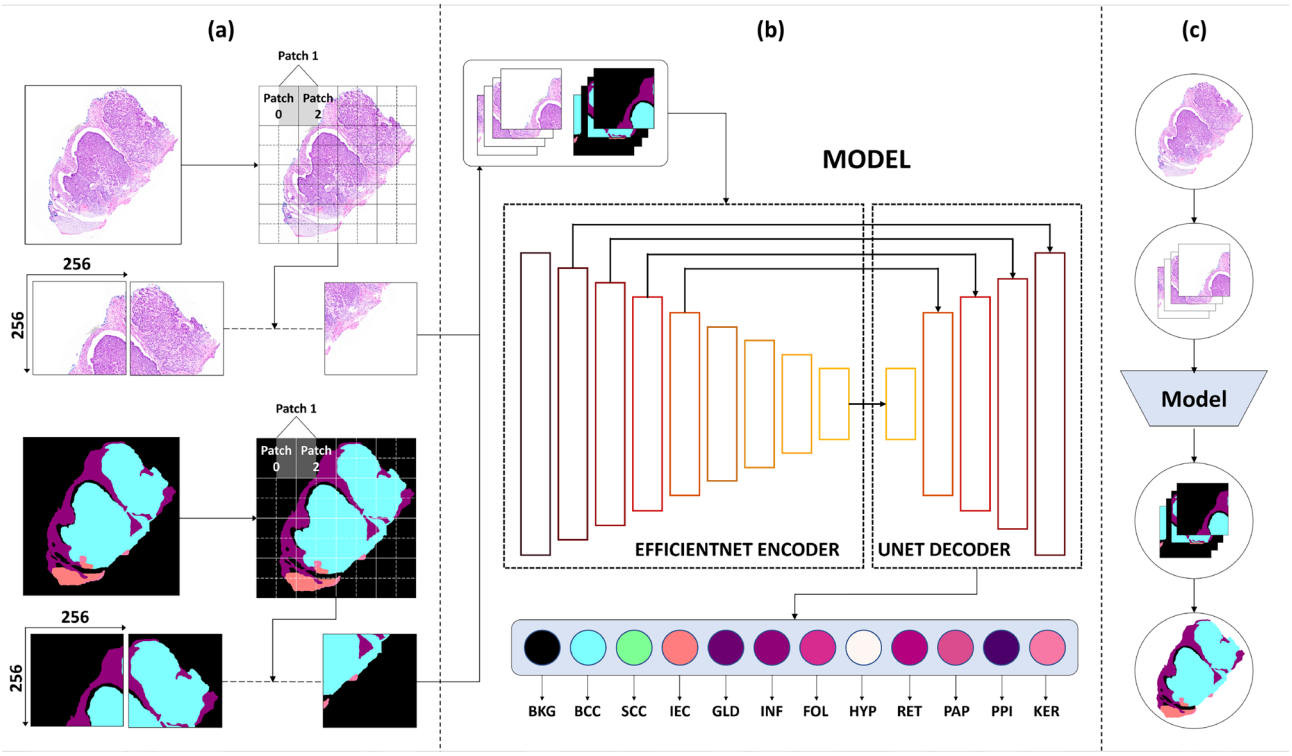


Figure 2. An overview of the entire working of the segmentation pipeline. (**a**) Shows the data input pipeline for training the model architecture, the input WSI and its corresponding mask is split into overlapping patches of 256 × 256. (**b**) The overlapping patches are fed in batches to the model. The model architecture is an EfficientNet-B3[12] encoder with a simple U-Net decoder[10]. The decoder transforms the 256 × 256 × 3 input to a 256 × 256 × 12, one hot encoded, output where each layer's probability is shown. Stage (**c**) shows the testing pipeline, which is same approach used by[26] where the input image is patched, rotated and augmented in and then fed to the inference model. This ensures smooth predictions and the network to achieve maximum confidence by.

U-NET

U-Net, an effective architecture for image segmentation[27], it includes context detection followed by correct localization. Usually it is difficult to have a large amount of training data in the field of biomedical, so a good deal about this algorithm is that it can be trained end to end with fewer images. UNet has a proven track record of multi-layer image segmentation by giving a specific class to each pixel in the image. The input data is down-sampled (encoded) to extract smaller features from the image and also to represent it for better segmentation, followed by upsampling (decoding) with concatenation of the skip connections from the downsampled layer to restore the original image. Pairing this helps the model get more details from the image while it was being encoded. To put it in simply, the downsampling process extracts the "what" from the image, and upsampling connects it to the "where" in that image[10].

In the first half of the architecture that consists of pre trained classification encoder network, we gave our input images of size 256 × 256 with 3 channels. It goes through couple of 3 × 3 convolutional layers that comprise of 64 filters, then maxpooling operation of 2 × 2 is performed to down samples the spatial dimensions of image to 128 × 128 with 128 filters. The steps were repeated to gradually increase the depth and reduce the image size from 128 × 128 with 128 filters to 16 × 16 with 1024 filters progressively.

Decoder network has 1024 filters with 16 × 16 dimension of lower resolution onto the pixel space. The up sampling is done by 2 × 2 transpose convolution by concatenation from corresponding blocks of encoder network, and to keep the spatial resolution intact ReLu function is performed. Here our 16 × 16 image with 1024 filter is up sampled to 32 × 32 with 512 filter along with the concatenation from its corresponding 512 features map from encoder. The depth is gradually decreased and image size is increased until we reach our final layer. In the end 1 × 1 convolution is used to map our desired 3 channeled 256 × 256 image to 12 classes for Segmentation.

EfficientNet-B3

ConvNet provides better result if scaled up in width, depth or image resolution, but after achieving a certain level, the improvement in results are diminishing. Scaling some or all components has also been tested, but requires a lot of manual fine-tuning[28] and consistently produces substandard accuracy and efficacy. Extending network depth is the most common means used by many ConvNets where the intuition is that the deeper ConvNet can capture richer and more complex functionality and generalize well to new tasks[29]. At the same time, deeper networks are more difficult to form due to the vanishing gradient problem. Network width scaling is often used for small models as larger networks tend to be able to capture features in more detail and are easier to train[30].

However, ultra-wide but shallow networks tend to have difficulty capturing higher-level peculiarities. With higher resolution input images, ConvNets is capable of capturing better segments[31], but accuracy decreases for very high resolutions. Efficient Net is a variety of ConvNet but with balanced scaling with respect to width, depth and Image resolution. It is observed that the different scale sizes are not independent. Intuitively, for higher resolution images we should increase the depth, so that larger receptive fields can help capture the identical features including more pixels in the large image. Accordingly, we should also increase the width when the resolution is higher, to capture finer samples with more pixels in the high resolution image. These perception suggest that we need to coordinate and balance different scale sizes rather than conventional uni-directional scaling[12].

Our Efficient Net model consisted of 10M parameters and consisted of 9 layers each in the encoder and decoder. The base layer had a total of 1536 features which then grew into the class wise mask segmentation. Model is trained to disburse the image with respect to different layers available in that specific image and assign each layer to a class. T

Ensemble

Ensemble is a popular technique in machine learning technique that involves combining the predictions of multiple models to improve the overall accuracy and robustness of the system. One popular type of ensemble is the weighted average ensemble, which involves combining the predictions of multiple models using a weighted average, where the weights are determined based on the relative performance of each model. In this particular case, we are interested in ensembling two different U-Net models, one with a vanilla UNet architecture and one with an EfficientNet-B3 backbone. The U-Net is a popular architecture for semantic segmentation tasks, while the EfficientNet-B3 is a state-of-the-art model architecture for image classification tasks. By combining these two models, the goal is to leverage the strengths of both architectures to improve the overall performance of our segmentation system. Table 2 shows number of parameters for all models used in this research.

Model training

The fine-tuning of the U-Net backbone models in the Queensland dataset was performed using TensorFlow and trained on a GeForce RTX 2070 8GB GPU. During training, the model parameters were adjusted using the Adaptive Moment Estimation (Adam) optimizer, with batch size set at 4 and maximum 50 epochs, based on the gradients of loss function. Adam optimizer's adaptive learning approach for individual parameters converges to a sharper minima faster, and shows superior training performance[32]. The parameters resulting in the minimum validation loss during training were used for testing the model. Early stopping and learning rate decay were used to improve model performance. The EfficientNet-B3 model showed a training loss of 0.3 and a training accuracy of 94, while a validation loss of 0.38 and a validation accuracy of 92. The loss function used in the training was the categorical cross-entropy loss. It was selected because it calculates the sum of loss over all classes. Using the negative sign in front of the sum ensures that the loss is minimized during training. The loss is defined as:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{1}$$

During the testing phase, the input images underwent a smoothing process that included a combination of rotations, mirroring, and windowing techniques. The transformations were applied to augment the images by generating eight different versions of each image at different angles, which were then combined to produce a more robust prediction. Additionally, a simple second-order spline window function was used for interpolation to blend the predictions together, with a default overlap of 50% between merged windows. This approach ensured that the predictions merged smoothly. Figure 3 shows EfficientNet-B3 training and validation curves for accuracy and loss.

Experiments and results

The proposed model was tested using skin biopsy labelled by pathologists.

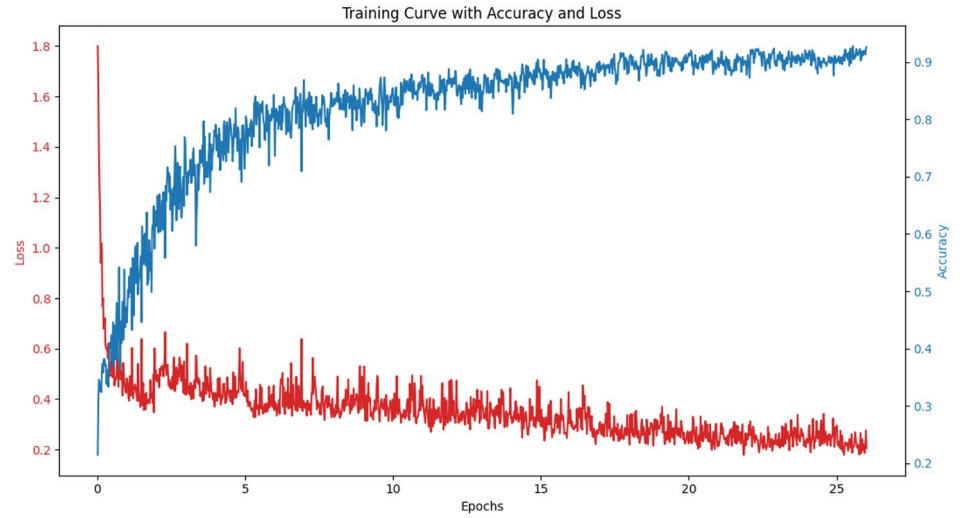
Evaluation metrics

The final segmented whole slide images were compared to the ground-truth (GT) images. The testing results in terms of class-wise recall are shown in Table 1. To define these metrics, the individual pixels in both sets of GT and model outputs were categorized into four classes: True-Positive (TP), True-Negative (TN), False-Positive (FP), and False-Negative (FN) as follows:

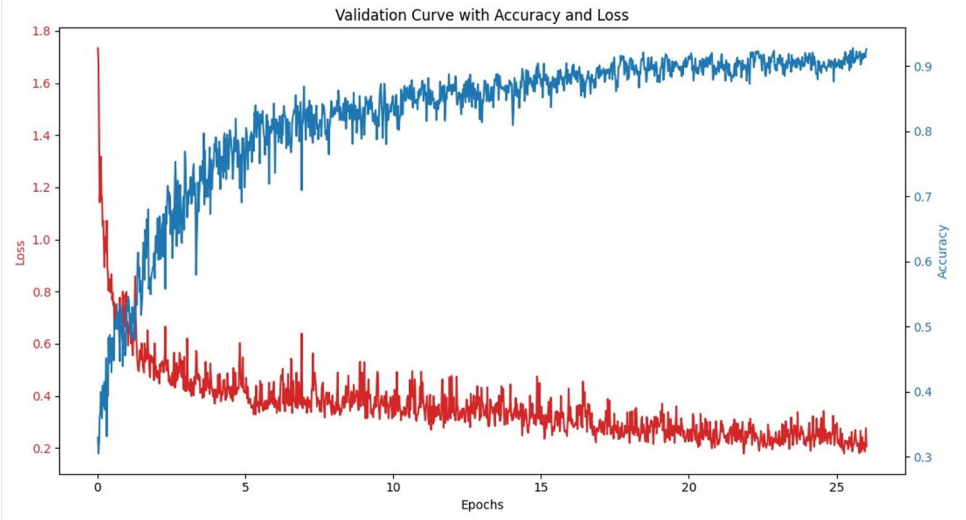
- TP: Pixels correctly segmented as the target class by the model.

Sr#	Model	Parameters
1	U-Net	8 M
2	EfficientNet-B3	10 M
3	Ensemble	18 M

Table 2. Total number of parameters.



(a)



(b)

Figure 3. Accuracy (blue) and loss (red) curves. (**a**) Training curves (**b**) validation curves.

- TN: Pixels correctly identified as not belonging to the target class by the model.
- FP: Pixels incorrectly identified as belonging to the target class by the model.
- FN:Pixels incorrectly identified as not belonging to the target class by the model.

The class-wise accuracy is defined as the recall:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Another useful metric is the Dice F-1 Score which shows intersection of a said class in the terms of the total area:

$$F1 - Score = \frac{TP}{TP + 1/2(FP + FN)} \tag{3}$$

Overall the per-pixel accuracy score is calculated, which is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Ablation study

To evaluate the proposed segmentation model, we conducted detailed ablation studies. As shown in Fig. 4, our studies explored how different backbones on a U-Net performed. The predictions from these models were aggregated to form an ensemble, which included the original U-Net architecture as well.

In our experiments, we employed an 80:10:10 data split for training, validation, and testing, respectively. This partitioning strategy ensures that the model is trained on a diverse set of examples, validated on a separate dataset to tune hyperparameters (as shown in Table 3), and finally tested on an independent set to evaluate its generalization performance. The rationale behind this split is to strike a balance between providing sufficient data for training and robustly assessing the model's performance on previously unseen instances. The model EfficientNet-B3 took 6 hours to train and during inference stage it takes only 30sec to generate results from a single slice.

Quantitative results

Table 4 shows the results of the studies in the form of a performance evaluation of the proposed models, both when independently and in the form of an ensemble. The results show how the proposed model outperforms the existing approach by a significant margin in cancerous classes, where an average increase of 6% is observed.

Based on the results, it is also important to mention that an ensemble of models did not improve the results significantly, and an individual model with an EfficientNet-B3 backbone would achieve the same results as those of an ensemble model approach. The results clearly highlight that proposed modified EfficientNet-B3 backbone

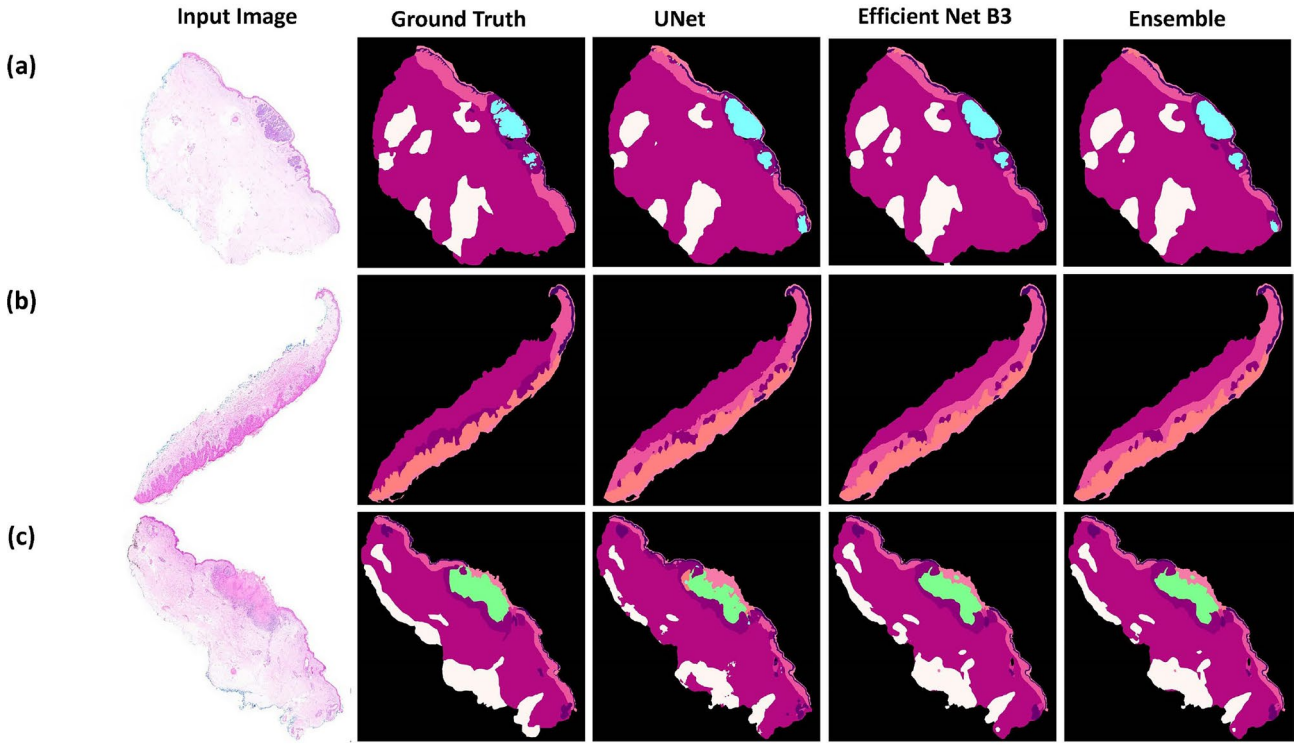


Figure 4. Visualizing the accuracy: generated masks of skin cancer types from the Queensland Dataset compared to Ground Truth: (**a**) Generated mask with predominant BCC cancer visually represented. (**b**) Generated mask with predominant IEC cancer displayed. (**c**) Generated mask illustrating predominant SCC cancer.

Parameter	Value
Batch size	4
Buffer size	40
Epochs	50
Learning rate	0.0001
Dimensions	256 × 256

Table 3. Hyperparameters for model training. The learning rate was dynamic and decayed as training progressed.

Layer	UNet	ENB3 backbone	Ensemble	Simon et al.[23]
BKG	**0.99**	**0.99**	**0.99**	0.95
BCC	0.91	0.90	**0.91**	0.86
SCC	0.70	**0.86**	0.83	0.85
IEC	0.82	0.82	**0.83**	0.70
EPI	0.70	0.78	0.78	**0.83**
GLD	0.81	**0.89**	**0.89**	0.87
INF	0.64	**0.70**	**0.70**	0.57
RET	**0.91**	**0.91**	**0.91**	0.70
FOL	0.65	0.66	**0.67**	0.61
PAP	0.68	0.73	0.72	**0.80**
HYP	0.85	0.89	0.89	**0.96**
KER	0.79	**0.85**	0.83	0.84

Table 4. Class wise recall. Bold text shows best result, Second best is underlined.

network show improved results specifically for cancerous regions i.e. BCC, SCC and IEC. Our model performed most poorly on the *FOL* class with a recall of 0.67. This was primarily due to its unbalance in the dataset while also depending on the depth of the biopsy in the case of shave biopsies for example, some hair follicles may be included in the specimen obtained. However, since shave biopsies only remove a superficial layer of skin, the hair follicles may not be fully intact.

Figure 5 shows the confusion matrix for the EfficientNet-B3 model. This confusion matrix visualizes the performance of our segmentation model across 12 classes, with a focus on the recall metric. Each row of the matrix corresponds to the actual class, while each column represents the predicted class. The main diagonal, normalized to highlight recall, shows the percentage of correct predictions for each class. Notably, the model exhibits lower performance in identifying the 'PAP' class, primarily due to its frequent confusion with the 'RET' class. This confusion arises from the similarity in layers and proximity of these classes in the skin. Additionally, the limited data available for training the model on the 'PAP' class exacerbated this issue. Another class where the model under performs is 'FOL', again largely due to the insufficient data for effective training. Nevertheless the results are still better than the existing ones due to a strong pre-trained backbone.

Qualitative analysis

The visual results of the experiment support the findings in the quantitative analysis. We also see how an EfficientNet-B3 backbone greatly improves the results of the experiment. The model predicts the *BCC* class much better than others, this is again evident from the skewness in the dataset. The model confuses between *IEC* and

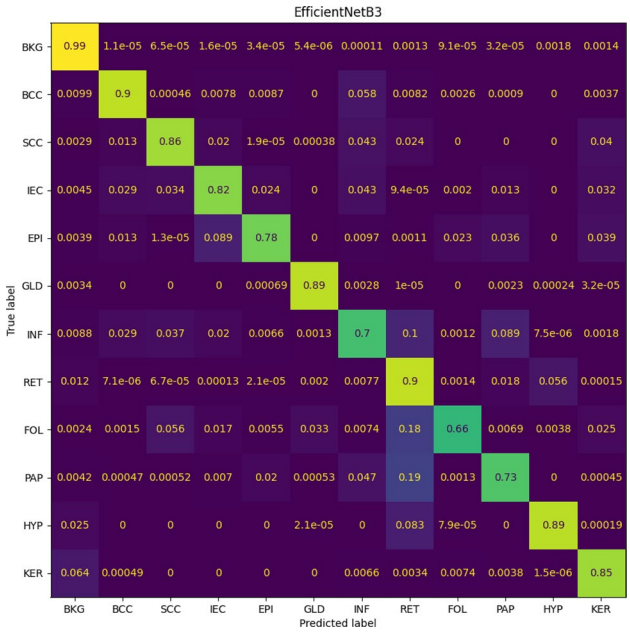


Figure 5. Confusion matrix showing recall values for EfficientNet-B3 model against each class.

SCC as both the classes have minor differences. Figure 6 shows how the model predictions compare to labelling done by pathologist.

For qualitative analysis in deep learning, uncertainty maps are commonly used to visualize the output of a model and identify regions where the model needs improvement or additional data. It provides valuable insights about how confident the models is about its predictions. For generating the maps as shown in Fig. 7, the predictions from the softmax function were transformed using the output of the last convolutional layer into a probability distribution over the 12 different classes. The resulting probability map can be visualized as a heat map where the intensity of each pixel represents the confidence of the models prediction for that pixel. The formula for the softmax function is:

$$\text{softmax}(x) = \frac{e^x}{\sum(e^x)} \tag{5}$$

where the x is the input to the softmax function, the output of the last convolutional layer of the model. The softmax function normalizes the output of the last convolutional layer into a probability distribution over the classes. After receiving the probability distribution we subtract it from 1 and with each pixel's maximum probability it is selected and then transformed based on the nipy-spectral color map to show uncertainty in each class.

$$P(x) = 1 - max(\text{softmax}(x)) \tag{6}$$

The heat maps show that the model segments out the cancerous area with high precision and confidence, while the border layers do show less confidence due to merging of the overlapping patches.

In the context of surgical applications, this model is specifically tailored to enhance the evaluation of surgical margins in skin cancer surgeries, a critical aspect that significantly impacts patient outcomes. Surgical margin clearance is a vital consideration in oncology surgery. It refers to the distance between the tumor boundary and the nearest edge of excised tissue. Ensuring adequate margin clearance is paramount to achieving complete tumor resection and minimizing the risk of recurrence.. Historically, the assessment of these margins has been a manual and subjective process, often leading to variability in interpretations and surgical outcomes.

Figure 8 comprises three parts: the original whole slide skin image, its segmented mask as generated by the AI model, and a visual representation of the margin clearance. The segmented mask clearly delineates the cancerous tissues, while the margin clearance visualization aids in understanding the extent of the spread of the cancer and the necessary boundary for excision.

Furthermore, the integration of our AI model into clinical practice promises to standardize the evaluation of surgical margins. By providing objective and quantifiable measurements, it reduces the reliance on subjective assessments, thereby potentially decreasing the variance in surgical outcomes. Moreover, this technology can serve as a valuable educational tool for pathologists and surgeons, offering insights into the complex patterns of tumor spreading skin cancer.

Comparison with literature

The proposed approach for segmentation of WSI skin images using deep learning was compared to the approach presented in Thomas et al.[23] as shown in Table 4 and Table 5. The results showed that the proposed approach

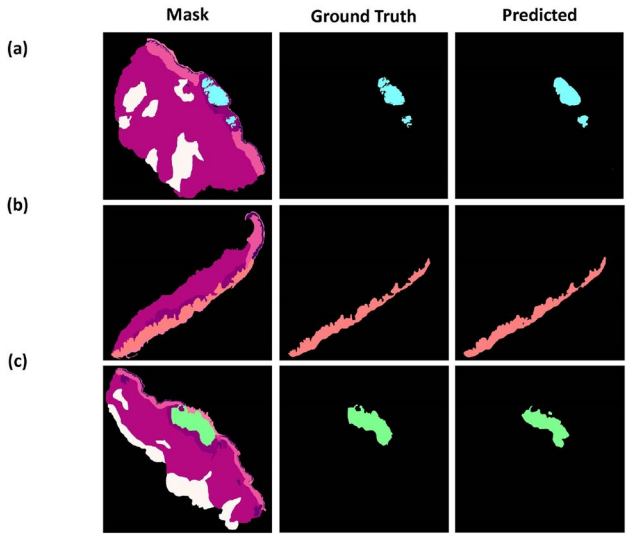


Figure 6. Isolating BCC, SCC, and IEC Masks in Different Models. (**a**) Compares the Basal cell carcinoma labeling provided by a doctor, versus that of the trained model. (**b**) Compares the Intra-epidermal cell carcinoma labeling provided by a doctor, versus that of the trained model. The IEC was difficult to detect as there was a lot of overlapping in it and the Epidermal class. (**c**) Compares the Squamous cell carcinoma labeled by a doctor to the trained model label.

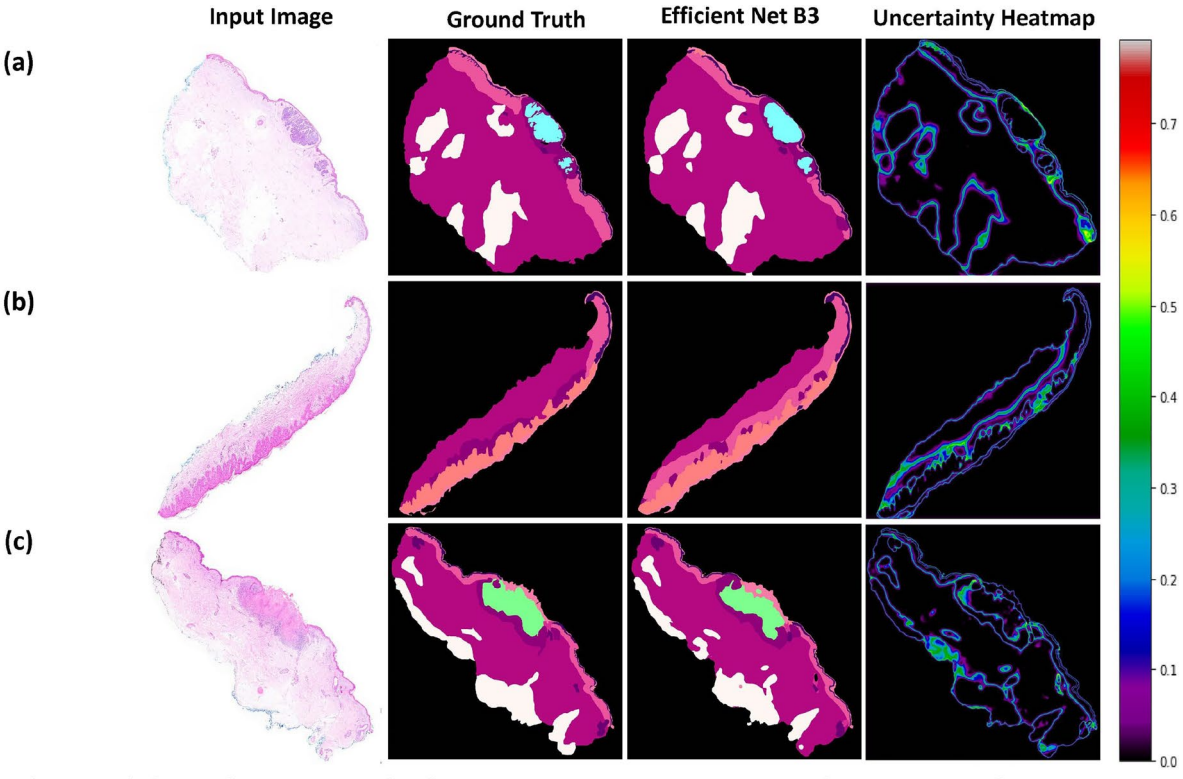


Figure 7. The model predictions with their respective uncertainty heatmaps.The Uncertainty heatmaps provided an interpretable way of diagnosis, whereby a model's confidence indicates the need for a physician to judge the model's results and where they need to pitch in for improved diagnoses. Generated mask and uncertainty heatmap of (**a**) BCC cancer (**b**) IEC cancer (**c**) and SCC cancer.

performed better in most cases, with an overall increase in accuracy and an improvement in average class accuracy. These results demonstrate the effectiveness of the proposed approach and its superiority over the existing approach in the literature. There is 5% to 10% increase in comparison to[23] for mean class accuracy and overall accuracy respectively. The improvement in accuracy can be attributed to the use of an EfficientNet-B3 backbone, which manages to extract and learn better features than the existent approach, and addresses class imbalance through data augmentation. These findings contribute to the ongoing research in the field and provide valuable insights for future studies.

Conclusion and future work

Early evaluation of skin cancer is important and reliable detection of skin diseases is necessary in dermatology. Skin segmentation tasks provide important data insights relevant to patient's disease management. Recently, dermatological researchers and regulatory agencies have been attentive on self-acting skin image recognition means to lessen the time, cost and need for human evaluation. In many cases, the outcome of the entire scan is highly dependent on the segmented approach, as evaluating the healing process and other treatment steps depend on the segmented areas. From traditional image processing methods to deep learning algorithms based on computer vision technology, skin segmentation is becoming more and more convenient and efficient. However, the process still receives backlash from the skin community over the use of automated technologies and mechanisms that cannot be easily explained. The proposed technique is able to apply semantic segmentation to extract 12 categories from skin tissue samples. The experiments have shown that U-net model with EfficentNet-B3 back bone gave promising results when compared with existing literature. Despite significant advances in recognition algorithms, a major limitation and challenge is to develop an up-to-date and comprehensive algorithm that excels in accuracy and flexibility across diverse datasets. The variability in staining protocols among labs, coupled with variations in color and contrast of tissue samples, poses hurdles for achieving precise skin segmentation. Additionally, the large resolution of Whole Slide images, often in Gigapixels, contributes to slower processing speeds. For future work in the field, it can be interesting to utilize latest transformers based deep learning models like Segformer and Swin-Unet and see how good they are in this semantic segmentation problem in comparison to convolutional models. Another requirement in this field is to have dataset agnostic algorithms. So experimenting with knowledge distillation along with above mentioned models is required.

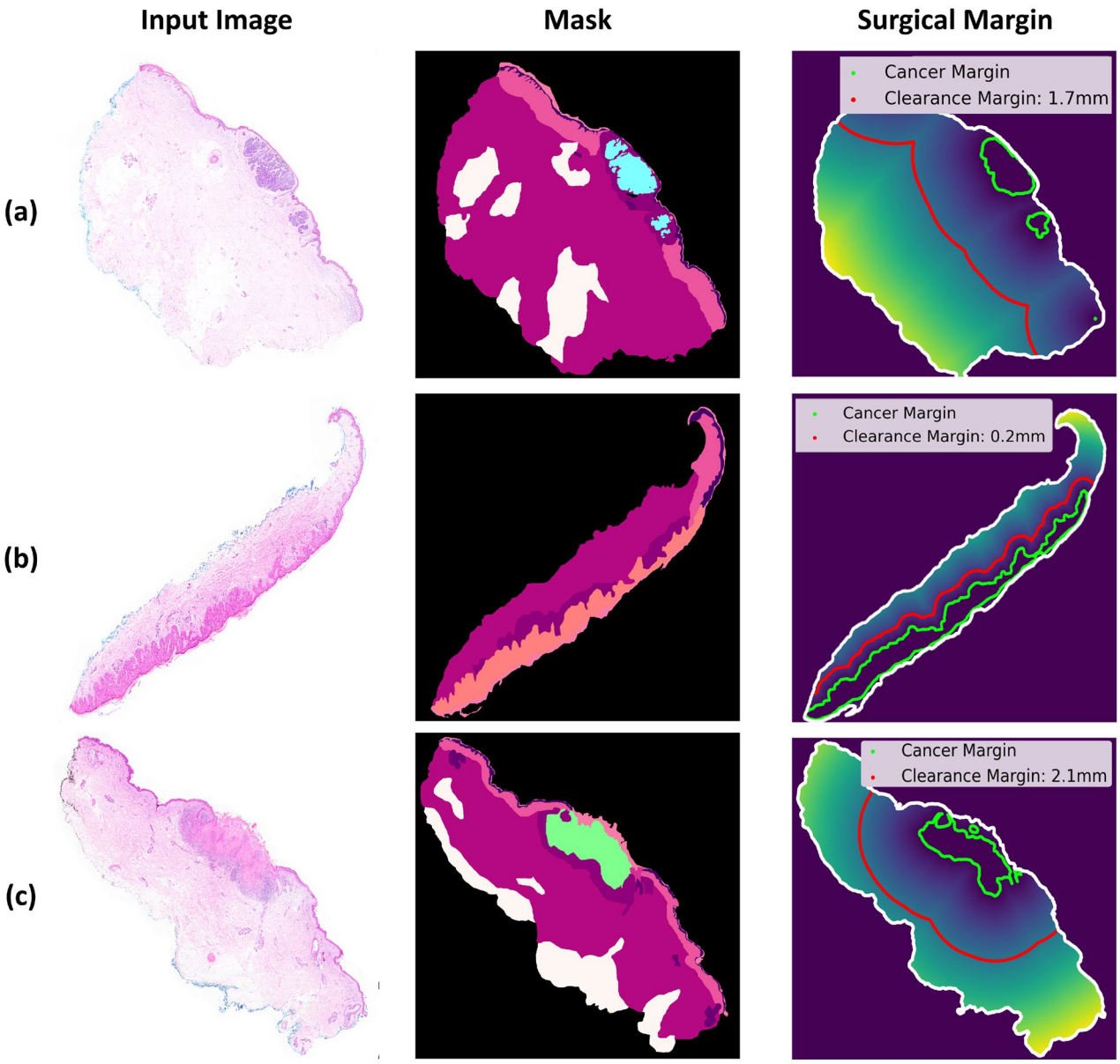


Figure 8. The generated surgical margin clearance. The green line highlights the location of the cancerous region. The red line indicates the clearance margin, i.e., the region where the cancer might have spread to and where the doctor should most likely perform a cut. Input Image, Generated Mask and Surgical Margin Clearance of (**a**) BCC cancer (**b**) IEC cancer (**c**) and SCC cancer.

Sr#	Model	F-1 score	Accuracy	Mean class accuracy
1	Simon et al.[23]	–	0.85	0.79
2	U-Net	0.92	0.91	0.78
3	EfficientNet-B3	0.93	0.95	0.83
4	Ensemble	0.94	0.95	0.84

Table 5. Comparison of models using different performance parameters.

Data Availibility

The authors have used publicly available dataset that can be accessed at https://espace.library.uq.edu.au/view/UQ:8be4bd0[24].

Received: 18 July 2023; Accepted: 23 August 2024
Published online: 08 October 2024

References

1. Asgari Taghanaki, S., Abhishek, K., Cohen, J. P., Cohen-Adad, J. & Hamarneh, G. Deep semantic segmentation of natural and medical images: A review. *Artif. Intell. Rev.* **54**(1), 137–178 (2021).
2. Rezaei, M., Harmuth, K., Gierke, W., Kellermeier, T., Fischer, M., Yang, H., & Meinel, C.: A conditional adversarial network for semantic segmentation of brain tumor. In: *International MICCAI Brainlesion Workshop* 241–252 (Springer, 2017).
3. Schreiner, T. G., Turcan, I., Olariu, M. A., Ciobanu, R. C. & Adam, M. Liquid biopsy and dielectrophoretic analysis-complementary methods in skin cancer monitoring. *Appl. Sci.* **12**(7), 3366 (2022).
4. Sarker, M. M. K. *et al.* Slsnet: Skin lesion segmentation using a lightweight generative adversarial network. *Expert Syst. Appl.* **183**, 115433 (2021).
5. Moradi, N. & Mahdavi-Amiri, N. Multi-class segmentation of skin lesions via joint dictionary learning. *Biomed. Signal Process. Control* **68**, 102787 (2021).
6. Stankovic, Z., Allen, B. D., Garcia, J., Jarvis, K. B. & Markl, M. 4d flow imaging with mri. *Cardiovasc. Diagnos. Ther.* **4**(2), 173 (2014).
7. Zheng, H., Yang, L., Chen, J., Han, J., Zhang, Y., Liang, P., Zhao, Z., Wang, C., & Chen, D.Z. Biomedical image segmentation via representative annotation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33 5901–5908 (2019)
8. Bagheri, F., Tarokh, M. J. & Ziaratban, M. Skin lesion segmentation from dermoscopic images by using mask r-cnn, retina-deeplab, and graph-based methods. *Biomed. Signal Process. Control* **67**, 102533 (2021).
9. Wibowo, A., Purnama, S. R., Wirawan, P. W. & Rasyidi, H. Lightweight encoder-decoder model for automatic skin lesion segmentation. *Inform. Med. Unlocked* **25**, 100640 (2021).
10. Ronneberger, O., Fischer, P., & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* 234–241 (Springer, 2015).
11. Haider, A., Arsalan, M., Nam, S.H., Hong, J.S., Sultan, H., & Park, K.R. Multi-scale feature retention and aggregation for colorectal cancer diagnosis using gastrointestinal images. *Eng. Appl. Artif. Intell.* **125**, 106749 (2023). https://doi.org/10.1016/j.engappai.2023.106749.
12. Tan, M., Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning* 6105–6114 (PMLR, 2019).
13. Kosgiker, G. M. & Deshpande, A. A novel segcap algorithm based enhanced segmentation of dermoscopic images of interest. *Mater. Today Proc.* **51**, 779–787 (2022).
14. Gu, R., Wang, L. & Zhang, L. De-net: A deep edge network with boundary information for automatic skin lesion segmentation. *Neurocomputing* **468**, 71–84 (2022).
15. Anthimopoulos, M. *et al.* Semantic segmentation of pathological lung tissue with dilated fully convolutional networks. *IEEE J. Biomed. Health Inform.* **23**(2), 714–722 (2018).
16. Li, D. *et al.* Deep learning for virtual histological staining of bright-field microscopic images of unlabeled carotid artery tissue. *Mol. imaging Biol.* **22**(5), 1301–1309 (2020).
17. Zhang, Y. *et al.* Msmanet: A multi-scale mesh aggregation network for brain tumor segmentation. *Appl. Soft Comput.* **110**, 107733 (2021).
18. Soulami, K. B., Kaabouch, N., Saidi, M. N. & Tamtaoui, A. Breast cancer: One-stage automated detection, segmentation, and classification of digital mammograms using unet model based-semantic segmentation. *Biomed. Signal Process. Control* **66**, 102481 (2021).
19. Wu, H. *et al.* Scs-net: A scale and context sensitive network for retinal vessel segmentation. *Med. Image Anal.* **70**, 102025 (2021).
20. Kleczek, P., Jaworek-Korjakowska, J. & Gorgon, M. A novel method for tissue segmentation in high-resolution h &e-stained histopathological whole-slide images. *Comput. Med. Imaging Graph.* **79**, 101686 (2020).
21. Oskal, K. R., Risdal, M., Janssen, E. A., Undersrud, E. S. & Gulsrud, T. O. A u-net based approach to epidermal tissue segmentation in whole slide histopathological images. *SN Appl. Sci.* **1**, 1–12 (2019).
22. Nofallah, S. *et al.* Segmenting skin biopsy images with coarse and sparse annotations using u-net. *J. Digit. Imaging* **35**(5), 1238–1249 (2022).
23. Thomas, S. M., Lefevre, J. G., Baxter, G. & Hamilton, N. A. Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer. *Med. Image Anal.* **68**, 101915 (2021).
24. Thomas, S., & Hamilton, N. *Histopathology Non-melanoma Skin Cancer Segmentation Dataset* (2021).
25. Kriegsmann, K. *et al.* Corrigendum: Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Front. Oncol.* **13**, 1201237 (2023).
26. Vooban. *Smoothly-Blend-Image-Patches* (GitHub, 2017)
27. Haider, A., Arsalan, M., Park, C., Sultan, H. & Park, K. R. Exploring deep feature-blending capabilities to assist glaucoma screening. *Appl. Soft Comput.* **133**, 109918 (2023).
28. Real, E., Aggarwal, A., Huang, Y., Le, Q.V.: Regularized evolution for image classifier architecture search. In: Proceedings Of The Aaai Conference On Artificial Intelligence, vol. 33 4780–4789 (2019)
29. Huang, G., Sun, Y., Liu, Z. &, Sedra, D., Weinberger, K.Q. Deep networks with stochastic depth. In: *European Conference on Computer Vision* 646–661 (Springer, 2016).
30. Zagoruyko, S., Komodakis, N.: *Wide Residual Networks*. arXiv preprint arXiv:1605.07146 (2016)
31. Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., Lee, H., Ngiam, J., Le, Q.V., Wu, Y., et al.: Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Adv. Neural Inf. Process. Syst.* **32**, 66 (2019).
32. Minhas, K. *et al.* Accurate pixel-wise skin segmentation using shallow fully convolutional neural network. *IEEE Access* **8**, 156314–156327 (2020).

Author contributions

M.Z.A. and H.R. performed the experimentation and initial writing. M.U.A. and A.S. looked after the whole research work and did teh complete review of the manuscript. T.H. and T.R. prepared all visualization and also performed data preparations.

Competing interests

Authors declare no competing interest.

Additional information

Correspondence and requests for materials should be addressed to M.U.A.