Delineation of Skin Strata in Reflectance Confocal Microscopy Images With Recurrent Convolutional Networks

Alican Bozkurt*
Northeastern University
alican@ece.neu.edu

Trevor Gale*
Northeastern University
tgale@ece.neu.edu

Kivanc Kose
Memorial Sloan Kettering Cancer Center
kosek@mskcc.org

Christi Alessi-Fox
Caliber I.D., Inc.
cmalessi@caliberid.com

Dana H. Brooks
Northeastern University
brooks@ece.neu.edu

Milind Rajadhyaksha
Memorial Sloan Kettering Cancer Center
rajadhym@mskcc.org

Jennifer Dy
Northeastern University
jdy@ece.neu.edu

Abstract

Reflectance confocal microscopy (RCM) is an effective, non-invasive pre-screening tool for cancer diagnosis. However, acquiring and reading RCM images requires extensive training and experience, and novice clinicians exhibit high variance in diagnostic accuracy. Consequently, there is a compelling need for quantitative tools to standardize image acquisition and analysis. In this study, we use deep recurrent convolutional neural networks to delineate skin strata in stacks of RCM images collected at consecutive depths. To perform diagnostic analysis, clinicians collect RCM images at 4-5 specific layers in the tissue. Our model automates this process by discriminating between RCM images of different layers. Testing our model on an expert labeled dataset of 504 RCM stacks, we achieve 87.97% *classification accuracy, and 9-fold reduction in the number of anatomically impossible errors compared to the previous state-of-the-art.*

1. Introduction

Reflectance confocal microscopy (RCM) is a non-invasive, optical imaging technology that enables users to examine $1.5\,\mu m$ thick layers of skin at $0.5\,\mu m$ pixel lateral resolution. Imaging can go as deep as $200\,\mu m$, which is sufficient for diagnosing several skin conditions, and typically covers the whole epidermis and papillary dermis. Recent studies have demonstrated that RCM imaging is highly sensitive ($90-100\,\%$) and specific ($70-90\,\%$) for detecting skin cancers [18]. Moreover, the combination of RCM and dermoscopy has been shown to reduce the rate of biopsy of benign lesions per detected malignancy by $\sim 2\times$, leading to better patient care [2, 17].

When applying RCM imaging, the field of view is usually limited to $1\,mm^2$ in order to maintain high pixel resolution in the images. However, clinicians typically require images over a much larger area, including the lesion and its periphery, to perform reliable diagnostic analysis. To cover the necessary area, multiple RCM images are collected in a non-overlapping grid (up to $6\,mm$-by-$6\,mm$). These images are then stitched together to form a larger, high-resolution image referred to as a *mosaic*.

While individual RCM images can be collected in under one second, constructing a mosaic over a large area can be very slow because the microscope must readjust its location for every image. Thus, exhaustively capturing mosaics at different depths in the skin is an expensive process. On top of the time required for data acquisition, analyzing numerous large mosaics can be very time consuming for clinicians. To avoid these costs, practitioners typically limit their data collection to five mosaics at the stratum granulosum, upper dermal-epidermal junction (DEJ), middle DEJ, lower DEJ, and papillary dermis. The depth of each of these strata vary from patient to patient and across different locations on the body, and thus must be identified prior to the collection of mosaics.

To identify the correct depths for gathering mosaics, clinicians take single RCM images at $1-5\,\mu m$ at consecutive depths from the epidermis to the dermis. This set of images is referred to as a RCM *stack* and each RCM image in the stack is called a RCM *slice*. After obtaining the stack, the clinician classifies each image as either epidermis, DEJ,

*Authors contributed equally

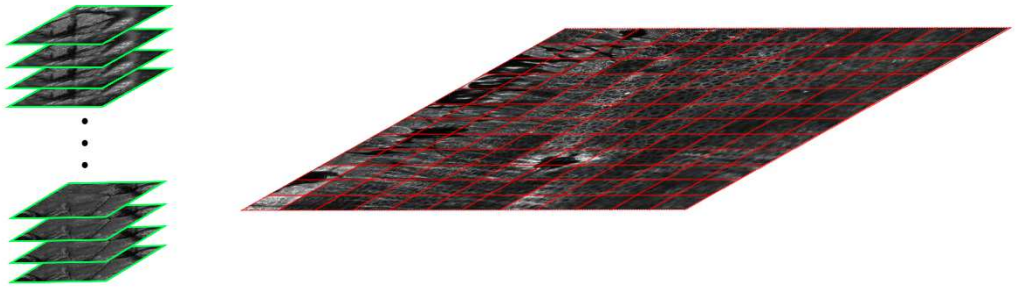

Figure 1: **RCM Data Structure.** (left) A 3D stack of RCM images used to determine the depth of different skin strata. (right) A mosaic of RCM images used for diagnosis. Colored borders represent single RCM images of the same dimensions.

or dermis. The clinician then uses the stack as a reference to collect mosaics at appropriate depths for diagnostic analysis.

Following this general technique, the process of applying RCM imaging to perform diagnosis can be broken down into two steps:

1. Collecting stacks and identifying the depth of different skin strata at the location of interest on the patients skin
2. Collecting mosaics of RCM images at a number of diagnostically relevant depths and analyzing morphological and cellular features of the skin tissue

One of the largest barriers to wider clinical adoption of RCM imaging is the difficulty of interpreting the images. RCM images have comparable resolution to histological images, but they lack the tissue specific contrast provided by the exogenous dyes used in histology. RCM images are gray-scale, and contrast is provided by the variance in the refractive index of different skin components. Thus, clinicians must analyze the texture of cellular and morphological structures in a local area to make diagnostic predictions. Early adopters with years of experience are capable of interpreting the images with high diagnostic sensitivity and specificity, but novice users often exhibit highly variable diagnostic accuracy. Due to the lack of proper training programs, gaining the necessary experience to effectively leverage this technology can be slow and inefficient.

Thus, there is a compelling need for quantitative tools for RCM image analysis in the clinical practice. In this work, we focus on automating the segmentation of RCM stacks to identify the depth of different skin strata. We first train a deep convolutional neural network (CNN) to classify individual RCM images as epidermis, DEJ, and dermis and then exploit the sequential structure of the data by augmenting the CNN with recurrent neural network (RNN) layers. Automation of this task will reduce time requirements for applying RCM imaging, and help to reduce the variance in diagnostic accuracy across clinicians by enabling consistently accurate collection of mosaics at the target layers of the skin. Moreover, our system can be used as a tool to train novice practitioners to interpret patterns in RCM images of skin tissue.

The contributions of this work are as follows. We conduct a thorough investigation of popular deep neural network architectures applied to RCM image classification and demonstrate significant improvements over the previous state-of-the-art results on this task. In addition to increased classification accuracy, our models also achieve a large reduction in the number of anatomically impossible errors compared to previous state-of-the-art methods, demonstrating a deeper understanding of the structure of RCM data. We evaluate our models on the largest dataset available for this task, which is composed of 21412 expert labeled RCM images from 504 different stacks. This dataset is also notable for containing disease suspicious (*e.g.* benign nevus and melanoma) samples. To the extent of our knowledge, all other datasets used for this task contain only healthy skin. However, the primary need for this system is for suspicious, lesional skin.

2. Related Work

Delineating skin strata in RCM stacks has been a topic of interest for many researchers [14, 15, 7, 21, 8, 12]. While a variety of algorithms have been applied to tackle this task, the different approaches can be broken into two main groups.

The first and more complex group aims to find a continuous 3D boundary between the layers of skin. The depth of different skin strata varies significantly between different points on the skin, with the boundary between the epidermis and dermis forming an undulated, 3D surface similar to an egg carton. Modeling this 3D boundary can provide clinicians with a detailed understanding of how the skin varies beneath the surface, but at the cost of significantly increased difficulty when compared to other methods. Locating the entire boundary is also unnecessary when estimating depths for mosaic acquisition.

The second set of methods approach the problem from

an image classification perspective. These techniques classify individual images from an RCM stack, and then take the approximate start and end points of each layer of skin as the points where the classifications transition between layers. Our method falls into this second category, as we perform image-wise classification of RCM slices to learn the location of the different layers of skin. For the sake of completeness, we discuss work from both categories in this section.

2.1. 3D DEJ Boundary Delineation

In [14, 15], Kurugol *et al*. present different methods for delineating the DEJ in RCM stacks from darkly and lightly pigmented skin. Melanin pigmentation is the main source of color in human skin, and is also the primary source of contrast in RCM images. In skin tissue, melanin caps sit on top of basal cells above the DEJ and protect them from UV exposure by reflecting sun light. Thus, in darkly pigmented skin, the basal cell layer is more easily distinguished from other less contrastive structures in neighboring layers of skin. For dark skin samples, the authors use a peak detection method to identify the intensity contrast provided by the melanin capped basal cells, and then mark the end of the basal layer as the DEJ boundary. For fair skin, the authors extract log-Gabor [6] and wavelet [19] features from RCM slices and use these to distinguish the epidermis from the dermis based on their textural appearance. Because the textural differences between the lower epidermis and papillary dermis are subtle around the DEJ, the authors delineate the DEJ as a thick transition band. For both types of skin, they make predictions on small regions of RCM images (referred to as a *tiles*) so that they can estimate a 3D surface. Testing on a dataset of 15 stacks of dark skin and 15 stacks of fair skin, they locate the depth of the DEJ with an average error of $7.9 \pm 6.4\,\mu m$ for dark skin and $8.3 \pm 5.8\,\mu m$ for fair skin.

Building off this work, in [7], Ghanta *et al*. aimed to increase the accuracy of delineation by incorporating a mathematical shape (micro-anatomy) model for the DEJ into the texture based appearance models. The DEJ is an undulated boundary, where hills are formed by projections of dermis into epidermis and valleys are formed by projection of epidermis into dermis. The authors fit 3D ellipses to these hills and valleys to model the 3D shape of the boundary. The parameters of the ellipses are modelled using a probabilistic framework and are inferred through a Gibbs sampling method. Testing on a dataset of 15 fair and 9 dark skin stacks, their algorithm achieved a mean accuracy of $12.1 \pm 7.0\,\mu m$ and $5.41 \pm 3.94\,\mu m$ respectively.

2.2. RCM Image Classification

In [21], Somoza *et al*. use Leung-Malik (LM) filter bank [16] based texton features to model the textural appearance of individual RCM slices. For each image in their training set, they extract texton features, find a bag of words representation, and finally describe each image as a histogram of its texton features. New samples are classified using a k-nearest neighbor classifier. Testing on image-wise labeled RCM stacks, they report correlation coefficients of 0.84 to 0.95 between their predictions and the ground truth.

Hames *et al*. [8] take a similar approach, but find a texton representation of random 7-by-7 patches extracted from a set of training images instead of using predefined texton filters. The authors describe RCM images by finding a bag of words representation of their texton filters followed by a histogram binning method. They then train a logistic regression classifier on 235 RCM stacks of healthy skin. Their model achieves 85.6% classification accuracy on a test set of 100 RCM stacks.

Kaur *et al*. [12] leverage the same texton extraction as Somoza *et al*., taking texton filters with a support of 5-by-5 pixels. They then construct a texton dictionary by clustering the filter outputs of randomly selected patches into 50 clusters using k-means, and use the cluster centers to form a bag of features representation. They assign each pixel to 8 of its closest textons with a weight inversely proportional to the distance between the texton. The individual assignments for each pixel in an RCM image are binned into a histogram that they use to describe the image. The authors then train a 3-layer neural network using these histograms. On their dataset of 15 stacks, they report 81.73% accuracy classifying images from the exterior epidermis, stratum corneum, stratum granulosum, stratum spinosum, stratum basale, and the papillary dermis.

3. Methodology

Given the sequential structure of RCM stack data, recurrent neural networks naturally lend themselves to the problem of skin strata identification. Human skin maintains a strict ordering of different strata; the transition between the layers must be smooth (layers are contiguous, *e.g*. epidermis→dermis→epidermis transition is not possible) and unidirectional (dermis→DEJ or DEJ→epidermis transitions are not possible). These constraints provide powerful cues that are exploited by human experts for more accurate classification. Given these requirements, recurrent networks are the obvious choice for this problem, as they are able to take the sequential dependencies between different images in a stack into account.

Within each RCM image, there is also a significant amount of spatial information present in the varying texture of the tissue. In past work, convolutional neural networks have demonstrated the ability to learn high-level features from images, and have been applied with great success to numerous image classification tasks [13, 23, 10].

Given these characteristics of our data, we adopt a hybrid neural network architecture with both convolutional

and recurrent layers similar to that proposed in [5]. We first trained a deep convolutional network to learn important spatial features for the classification of individual RCM images, and then augmented the network with recurrent layers so that the classifier could take the features of other RCM slices in the stack into account. Following the convention used in [5], we refer to models with this structure as recurrent convolutional networks (RCNs). For our deep CNN architecture, we used a modified Inception-V3 model [24], where we added an additional fully connected layer with 256 neurons before the last layer.

After training the model to classify individual RCM images as epidermis, DEJ, or dermis, we removed the 3-class classifier layer and the non-linearity on the penultimate fully connected layer. We then fixed the weights of the trained network and appended recurrent layers. The two different techniques that we experimented with for training the recurrent layers are explained in Section 3.1 and Section 3.2.

For both training schemes, we experimented with bidirectional RNN layers [20] as well as various different recurrent units (standard RNN cells, gated recurrent units (GRUs) [3], and long short-term memory units (LSTMs) [11]). Bidirectional RNN architectures have achieved state-of-the-art results on speech-to-text tasks, but run into issues in real-time applications because they require the whole input sequence to be available before inference can be performed [9, 1]. However, this is not an issue for our task. The process of inspecting the RCM stacks in both directions is also a very logical step that an expert might take while performing classification.

3.1. Partial Sequence Training

The first approach we took for training the RCN model was training on sequences containing a local neighborhood of N images around the subject RCM image (see Figure 2. For every sample in our dataset, we constructed a sequence of N RCM slices centered around the target sample. We then trained our network on batches of these sequences. This training procedure imitates the technique of examining neighboring RCM slices that dermatologists apply while classifying RCM images. All partial sequence models presented in this study were trained using image sequences of length 3 ($N = 3$), *i.e.* while classifying a slice of interest, the model only has its immediate neighbours in the stack as additional information. We experimented with larger N, but saw no increase in accuracy.

3.2. Sequence-to-Sequence Training

The second approach we took for training the RCN model was full sequence-to-sequence training. As illustrated in Figure 3, the model processes whole RCM stacks and outputs predictions for each image in the stack. This approach is potentially more flexible, as we provide the model with the complete RCM stack and allow it to learn what information is useful for slice-wise classification.

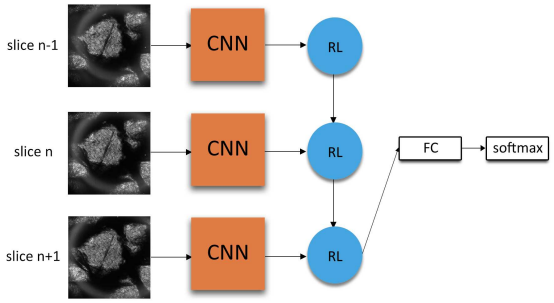

Figure 2: **Partial Sequence Training Scheme.** The output of the recurrent layer for the last element in the partial sequence is used for classification.

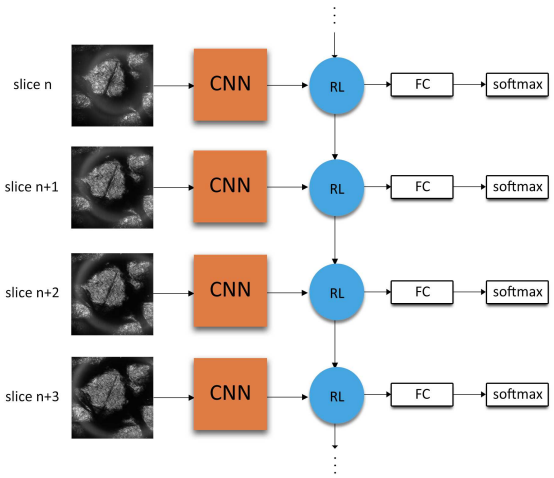

Figure 3: **Sequence-to-Sequence Training Scheme.** The output of the recurrent layer at each step is used for classification of that sample.

4. Dataset

The dataset used in this work is composed of 504 RCM stacks that were gathered from 2 different studies. 196 stacks come from a multi-center, National Cancer Institute (NCI) study that was conducted at the Memorial Sloan Kettering Cancer Center (New York, New York), the University of Rochester Medical Center (Rochester, New York), and the Skin Cancer Associates Center (Plantation, Florida). Each individual image in this set of stacks is labeled by at least 2 experts as one of 3 classes: epidermis, DEJ, or dermis. The other 308 stacks were taken from a previous study conducted by the Dermatology Research Centre at the University of Queensland (Brisbane, Australia) [8]. These

stacks are labeled into one of 4 classes: stratum corneum, epidermis, DEJ, or dermis. For this study, we merged the stratum corneum, which is the top layer of the epidermis, and the epidermis classes together. The overall dataset consists of 21412 RCM images.

The dataset also contains normal, benign melanocytic, and diseased skin samples. This is very important, as clinicians typically image suspicious lesional skin where the appearance of the tissue is very different from healthy and/or non-lesional skin. Thus, for our system to be applicable in practice, it must be robust to the variance in appearance of tissue across healthy and unhealthy images.

For our experiments, we partitioned the dataset into training, validation, and testing sets of 245, 61, and 198 stacks respectively. Because one patient may have multiple stacks in the dataset, we partition the data patient-wise (*i.e.* all stacks from a particular patient can either be in the training, validation, or testing set).

5. Experiments

All RCN models were implemented using the Keras [4] deep learning library and trained using the Theano [25] backend on a single NVIDIA Tesla K40 GPU.

The original inception-V3 network is designed for RGB images. Since RCM data is grayscale, we triple the number of filters in first layer to keep number of parameters same. We trained this modified CNN on a dataset augmented with randomly sheared, zoomed, rotated, stretched, horizontally and vertically flipped versions of training images.

We were not able to train full RCN models end-to-end, as the batch normalization layers in inception-V3 model are not designed to be trained in a time distributed setting. Removing these layers allowed us to train the complete RCN, but the model performed significantly worse. To overcome this problem, we first trained a CNN using the RCM images in the training set. We then removed the last layer of the trained CNN and used the pruned network as a feature extractor to obtain feature representations for each slice in the dataset. The recurrent layers are then trained on sequences of extracted features. While this approach makes experimentation with different CNNs more difficult, it allows us to use CNNs with batch normalization, and avoids significant redundant computation while training different RNNs. It also enabled us to train sequence-to-sequence models on a single GPU, as the full CNN + RNN model was too large to fit into GPU memory.

All RCNs in Table 1 have a single recurrent layer of 64 units, followed by a fully connected layer and a softmax for classification. For training partial sequence models, we concatenated extracted features of each image with features of preceding and following images, and used these concatenated features as inputs to the RNN. The partial sequence models were trained for 100 epochs with a batch size of 128 sequences, and a learning rate of 0.01. To avoid overfitting, we used dropout on the recurrent connections with a coefficient of 0.5.

For training sequence to sequence models, we concatenated extracted features of all images for each stack. Because our dataset contains stacks of various lengths, we fixed the maximum sequence length to 71 slices[1] and zero-padded shorter sequences. The padding was then masked out during training and testing. The RNN layers were trained for 200 epochs with a batch size of 4 sequences, and a learning rate of 0.001. To avoid overfitting, we used dropout on the recurrent connections and an L1 regularization penalty on the recurrent weights with coefficients of 0.1 and 0.05 respectively.

For each RCN model, the model snapshot with the best validation accuracy was selected, and its performance on the test set is reported in Table 1 [2].

For comparison, we used the publicly available python code from [8]. We also implemented the method presented in [12] by following the instructions in their paper. We report the results of training and testing on our dataset using these methods in Table 1.

6. Results & Analysis

We report test set performance for all of our models in Table 1. The table is broken into three row blocks, where all models in the same block share the same type of inputs. The first block is made up of models that take in whole RCM stacks, and output predictions for each image in the stack. The second block contains models that take in partial sequences to make predictions for individual RCM images. This includes our partial sequence RCN models, as well as an Inception-V3 model that we trained using the partial sequence data (IV3-Context in Table 1) for comparison. The last block contains models that classify single input RCM images, including Inception-V3 and ResNet-50 models that we trained for comparison, and the models proposed in [8] and [12]. Each block is sorted in descending order of test accuracy. In addition to test accuracy, we also report sensitivity and specificity for each of the three classes. The highest accuracy, specificity, and sensitivity for each row block of models are marked in bold.

There are a number of interesting observations that we can make from the results in Table 1. The best model overall was the sequence-to-sequence model with bidirectional GRU, which achieved 87.97% accuracy on the test set. This model outperformed the best previously published method by 3.49%, representing a 22.49% reduction in classification error rate. Comparing the partial sequence and sequence-

[1] The single RCM stack in our dataset of length greater than 71 was clipped from 101 to 71 slices for the seq-to-seq models

[2] Sequence-to-sequence LSTMs gave extremely unstable results and are not reported

Method	Recurrent Unit	Bidirec.	Accuracy	Sensitivity			Specificity		
				Epidermis	DEJ	Dermis	Epidermis	DEJ	Dermis
Seq2seq	GRU	Yes	**87.97**	93.95	**83.22**	84.16	95.82	**90.54**	95.51
Seq2seq	GRU	No	87.10	93.91	82.35	82.16	94.43	90.31	**95.60**
Seq2seq	RNN	No	86.88	**94.72**	80.69	82.87	93.78	91.20	94.99
Seq2seq	RNN	Yes	86.70	93.91	81.50	81.87	94.15	90.40	95.21
Partial seq.	GRU	Yes	**87.52**	**94.14**	82.54	**83.33**	94.78	**90.83**	95.44
Partial seq.	RNN	No	87.45	93.53	83.40	82.62	94.96	90.26	95.75
Partial seq.	LSTM	Yes	87.44	93.50	83.83	81.90	95.31	89.93	95.78
Partial seq.	LSTM	No	87.40	93.27	**84.29**	81.44	95.51	89.58	95.93
Partial seq.	RNN	Yes	87.32	93.64	83.28	82.09	94.97	90.02	95.80
IV3-Context	-	-	86.95	92.00	83.52	82.64	95.61	89.24	95.56
Partial seq.	GRU	No	86.53	90.35	85.54	80.88	**96.39**	87.54	**96.18**
Inception-V3 [24]	-	-	**84.87**	88.83	**84.66**	78.18	**95.84**	85.73	**96.23**
ResNet50 [10]	-	-	81.36	**89.87**	73.08	79.31	91.47	87.01	93.04
Hames *et al*. [8]	-	-	84.48	88.87	80.93	**81.85**	93.81	**87.81**	94.78
Kaur *et al*. [12]	-	-	64.33	73.99	51.14	68.27	86.22	74.85	84.89

Table 1: **Classification Results.** This table shows accuracy, sensitivity, and specificity results on the test set for each method.

to-sequence models, we observe that unidirectional GRU and both standard RNN models trained using the sequence-to-sequence scheme were outperformed by nearly all partial sequence methods, the one exception being the partial sequence unidirectional GRU model. In the sequence-to-sequence training scheme, we process the entire RCM stack at once and rely on the networks ability to identify the important information for classifying each image. Whereas in the partial sequence scheme, we effectively predetermine that the neighboring images contain the relevant information necessary to classify an RCM slice. We hypothesize that the simpler unidirectional and standard RNN architectures had a more difficult time learning on the full sequence data. Following this logic, it is reasonable to conclude that RCM images beyond the immediate area of the target slice contain some important information for classification of that slice, and that the more complex bidirectional GRU network was able to leverage this information to increase classification accuracy.

It is interesting to note that RCNs with GRUs outperformed those with LSTM cells, which is contrary to what we expected given the recent success of LSTM networks [22, 5]. To understand this discrepancy, we evaluated our LSTM and GRU partial sequence models on the validation set and found that the LSTM cells outperformed GRUs by 0.5% and 0.08% with normal and bidirectional models respectively. Despite the higher performance on the validation set, the LSTM models performed slightly worse on the test set, which indicates slight overfitting to the training/validation data.

Sensitivity and specificity were very similar across our experiments and appeared to vary proportionally with test accuracy. However, it is interesting to note that all of our models were significantly less sensitive to the DEJ and dermis. This is consistent with other results in the literature [15] and our expectations, as a typical stack of RCM images will contain more epidermis samples than DEJ and dermis because reflectance confocal microscopes can only image down to the papillary dermis (44% of samples in our dataset are epidermis, compared to 34% DEJ and 22% dermis). Moreover, due to optical aberrations around the rete ridges (valleys of the undulated DEJ boundary), the DEJ-to-dermis boundary can appear fuzzy, making it harder to detect. Thus, the level of DEJ-to-dermis boundary in a given stack is partially subjective, even for expert readers. This uncertainty helps to explain the lower sensitivity to DEJ and dermis compared to epidermis.

While the hybrid neural network model used in [12] performed poorly when trained on our dataset, it is worth noting that the logistic regression model presented by Hames *et al*. achieves performance comparable with the Inception-V3 network, and even outperforms the ResNet-50 model. While our recurrent models provide significant improvements in testing accuracy, they lack the interpretability of the regression approach which is a potential drawback in medical applications.

6.1. Types of Errors

The main goal of incorporating the sequential nature of RCM stack data into our model was to build a classifier that can understand the structure (unidirectionality and smoothness) of the data, and leverage this structure to increase classification accuracy. In this section, we analyze the types of errors made by each model to understand how well they learned these constraints.

We can categorize classification errors into two different types. Errors of the first type occur when a boundary starts shallower or deeper than expected in a stack. As explained in Section 3, the transition between labels in RCM stacks are smooth and unidirectional. These errors are tolerable

because they do not violate these constraints. Moreover, this type of error may arise due to inter-expert subjectivity. It is often difficult to assign a single label to a whole RCM image, as images acquired near boundaries can contain features from both classes. Thus, image-wise ground truth labels are partially subjective, and can vary from expert to expert. Given this inherent uncertainty, it is reasonable for our classifier to make these types of mistakes as long as they do not extend beyond the boundary region.

The second type of errors, which we refer to as *inconsistent* errors, are transitions that violate the sequential constraints of the data. These transitions are epidermis→dermis, DEJ→epidermis, dermis→epidermis, dermis→DEJ. To quantify the consistency of a model, we output predictions for each RCM stack in the dataset and count the number of inconsistent transitions between classes. We report these numbers for the best sequence-to-sequence and partial sequence RCNs, the Inception-V3 model, and the models presented in [8], and [12] in Table 2. Note that these results do not directly correspond to accuracy; it is possible for a set of labels to be consistent, but not accurate.

Method	**Error Types**				**Total**
	Epidermis → Dermis	DEJ→ Epidermis	Dermis→ Epidermis	Dermis → DEJ	
Seq2seq (Bidir. GRU)	0	4	0	3	7
Partial seq. (Bidir. GRU)	3	10	5	5	23
Inception-V3	3	25	8	32	68
Hames *et al.*	14	59	11	56	140
Kaur *et al.*	32	255	16	99	402

Table 2: **Inconsistent Errors.** This table shows the number of anatomically impossible predictions made by each model.

As expected, methods that do not take full stack information into account produce more inconsistencies. The partial sequence RCN produces $\sim 3\times$ fewer inconsistencies than the best single-image model while only considering neighborhoods of three images. The sequence-to-sequence RCN performs significantly better than the rest, with $\sim 3\times$ fewer inconsistencies than the partial sequence RCN and $\sim 9\times$ fewer inconsistencies than the best non-RCN model.

As the ultimate goal of our model is to delineate the epidermis-DEJ and DEJ-dermis borders, we also quantify the performance of our RCNs by looking at the distribution of the distance between the predicted boundaries and the ground truth boundaries. To obtain consistent transition boundaries, we use a two-step post-processing heuristic. The first step applies a median filter to the predictions using a kernel size of 3 to remove singular decisions. The second step then applies a causal max filter, which replaces each prediction with the maximum value in the sequence of predictions before it. We also apply this heuristic to the predictions of the Inception-V3 model, and the models presented in [8], and [12] for the sake of comparison. Figure 4 shows boxplots of the absolute deviation between the predicted and ground truth boundaries for each of these models. Red lines indicate the median, blue boxes contain the interquartile range, whiskers contain the 1.5 interquartile range, and red crosses are the outliers.

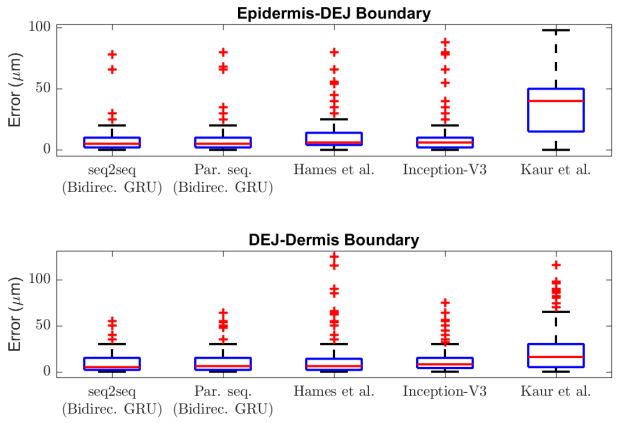

Figure 4: **Error Distributions.** This table shows boxplots of error distances for the predictions of the epidermis-DEJ and DEJ-dermis boundaries for each model. Red lines indicate the median, blue boxes contain the interquartile range, whiskers contain the 1.5 interquartile range, and red crosses are the outliers.

Analyzing the distributions in Figure 4, we see the sequence-to-sequence RCN, and partial sequence RCN achieve the lowest median error ($5\,\mu m$). The method proposed in [8] and the Inception-V3 model achieve the second lowest median error ($6\,\mu m$), but this difference is not statistically significant. Additionally, our methods have lower number of outliers than other methods. Note that these results are produced after the post-processing heuristic is applied. In order to make a fairer comparison, Figure 4 should be analyzed with Table 2, which shows the number of errors the heuristic algorithm needs to correct. A similar result follows for the DEJ-dermis boundary.

7. Conclusion

In this study, we presented a method based on deep convolutional and recurrent neural networks for classifying skin strata in RCM stacks. We evaluated our method on the largest and most comprehensive dataset for this task, and demonstrated a significant increase in the accuracy of skin strata delineation in RCM stacks. The test scenario used in this study is more realistic compared to those used by most previous methods, in the sense that clinicians are typically interested in imaging disease suspicious lesions rather than normal skin. In addition to increased classification accuracy, our best RCN achieved a $\sim 9\times$ reduction in the num-

ber of physically impossible transitions between layers of skin when compared to the previous state-of-the-art methods. Our experiments show that our method outperforms techniques designed for smaller datasets that comprise only healthy skin, and other deep learning based methods which do not incorporate full stack information. Overall, our results show that combining knowledge of the intrinsic properties of a dataset with the strengths of deep neural networks can yield a powerful tool for solving medical imaging problems, and can help to guide clinicians in their clinical practice.

8. Acknowledgements

This work was supported in part by the National Cancer Institute (NCI) under grants R01CA156773 and R01CA199673, the National Institute of Biomedical Imaging and Bioengineerings (NIBIB) Image Guided Interventions Program under grant R01EB012466, the NCI Core Center grant P30CA008748, and NIBIB grant R01EB020029.

References

[1] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 173–182, 2016. 4

[2] S. Borsari, R. Pampena, A. Lallas, A. Kyrgidis, E. Moscarella, E. Benati, M. Raucci, G. Pellacani, I. Zalaudek, G. Argenziano, et al. Clinical indications for use of reflectance confocal microscopy for skin cancer diagnosis. *JAMA dermatology*, 152(10):1093–1098, 2016. 1

[3] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 4

[4] F. Chollet. Keras. https://github.com/fchollet/keras, 2017. 5

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 4, 6

[6] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of Optical Society of America*, 4(12):2379–2394, 1987. 3

[7] S. Ghanta, M. I. Jordan, K. Kose, D. H. Brooks, M. Rajadhyaksha, and J. G. Dy. A marked poisson process driven latent shape model for 3d segmentation of reflectance confocal microscopy image stacks of human skin. *IEEE Transactions on Image Processing*, 26(1):172–184, Jan 2017. 2, 3

[8] S. C. Hames, M. Ardigò, H. P. Soyer, A. P. Bradley, and T. W. Prow. Automated segmentation of skin strata in reflectance confocal microscopy depth stacks. *PloS one*, 11(4):1–12, 2016. 2, 3, 4, 5, 6, 7

[9] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567, 2014. 4

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3, 6

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997. 4

[12] P. Kaur, K. J. Dana, G. O. Cula, and C. Mack. Hybrid deep learning for reflectance confocal microscopy skin images. In *2016 23rd International Conference on Pattern Recognition*, Dec 2016. 2, 3, 5, 6, 7

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 3

[14] S. Kurugol, J. G. Dy, D. H. Brooks, and M. Rajadhyaksha. Pilot study of semiautomated localization of the dermal/epidermal junction in reflectance confocal microscopy images of skin. *Journal of Biomedical Optics*, 16(3):036005, 2011. 2, 3

[15] S. Kurugol, K. Kose, B. Park, J. G. Dy, D. H. Brooks, and M. Rajadhyaksha. Automated delineation of dermalepidermal junction in reflectance confocal microscopy image stacks of human skin. *Journal of Investigative Dermatology*, 135(3):710 – 717, 2015. 2, 3, 6

[16] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vision*, 43(1):29–44, June 2001. 3

[17] G. Pellacani, A. Witkowski, A. M. Cesinaro, A. Losi, G. L. Colombo, A. Campagna, C. Longo, S. Piana, N. De Carvalho, F. Giusti, et al. Cost–benefit of reflectance confocal microscopy in the diagnostic performance of melanoma. *Journal of the European Academy of Dermatology and Venereology*, 30(3):413–419, 2016. 1

[18] M. Rajadhyaksha, A. Marghoob, A. Rossi, A. C. Halpern, and K. S. Nehal. Reflectance confocal microscopy of skin in vivo: From bench to bedside. *Lasers in Surgery and Medicine*, 2016. 1

[19] T. Randen and J. H. Husoy. Filtering for texture classification: a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:291–300, 1999. 3

[20] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, Nov 1997. 4

[21] E. Somoza, G. O. Cula, C. Correa, J. B. Hirsch, Editors:, A. e. Campilho, and M. Kamel. *Automatic Localization of Skin Layers in Reflectance Confocal Microscopy*, pages 141–150. Springer International Publishing, Cham, 2014. 2, 3

[22] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 6

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 4, 6

[25] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, 2016. 5