7-Point Checklist and Skin Lesion Classification using Multi-Task Multi-Modal Neural Nets

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh, *Senior Member, IEEE*

***Abstract*—We propose a multi-task deep convolutional neural network, trained on multi-modal data (clinical and dermoscopic images, and patient meta-data), to classify the 7-point melanoma checklist criteria and perform skin lesion diagnosis. Our neural network is trained using several multi-task loss functions, where each loss considers different combinations of the input modalities, which allows our model to be robust to missing data at inference time. Our final model classifies the 7-point checklist and skin condition diagnosis, produces multi-modal feature vectors suitable for image retrieval, and localizes clinically discriminant regions. We benchmark our approach using 1011 lesion cases, and report comprehensive results over all 7-point criteria and diagnosis. We also make our dataset (images and metadata) publicly available online at http://derm.cs.sfu.ca.**

***Index Terms*—skin, dermatology, 7-point checklist, melanoma, classification, convolutional neural networks, deep learning**

I. Introduction

SKIN cancer is the most common malignancy in fair-skinned populations, and the incidences of melanoma and non-melanoma skin cancers are rising, resulting in high economic costs [1]. Early melanoma diagnosis appears to improve patient outcomes [2], and skin cancer detection can be improved through approaches such as screening patients with focused skin symptoms using physician-directed total body skin examinations [3].

Epiluminescence microscopy or dermoscopy, which is a noninvasive in-vivo imaging technique, uncovers detailed morphological and visual properties of pigmented lesions. Kittler et al. [4] reported that, for experienced dermatologists, the accuracy in diagnosing pigmented skin lesions improves when using dermoscopy compared to the unaided eye. However, accurate diagnosis is challenging for non-experts.

Pattern analysis, which subjectivity assesses multiple subtle lesion features, is commonly used by experienced dermatologists to distinguish between benign and malignant skin tumours. To simplify diagnoses, rule-based diagnostic algorithms such as the ABCD rule [5] and the 7-point checklist [6] have been proposed and are commonly accepted [7]. In this work we focus on the 7-point checklist, which requires identifying seven dermoscopic criteria (Table I) associated with melanoma, where each criteria is assigned a score. The lesion is diagnosed as melanoma when the sum of the scores exceeds a given threshold [6], [8]. Although some literature recommends pattern analysis over the 7-point checklist [9], some works report a trade-off between melanoma sensitivity and specificity. For example, among dermatology residents, the 7-point checklist was found to give higher sensitivity, but lower specificity than pattern analysis [9]. A similar result was found among experienced dermatologists using a lowered 7-point checklist threshold [8]. This indicates limitations with both approaches, and motivates additional study. Further, although the 7-point checklist and pattern analysis diagnostic procedures are different, the 7-point checklist criteria are based on the criteria used in the process of pattern analysis [10]. Detecting these criteria may aid with more interpretable diagnostic models regardless of the preferred diagnostic procedure (e.g., report the presence of dermoscopic features associated with malignancy, retrieve images with specific criteria).

Computer aided approaches to classifying dermoscopic images have attracted significant research attention, as automated analysis has the potential to empower patients with timely, reproducible diagnoses, especially in remote communities with limited clinical access. Furthermore, the increasing prevalence of mobile and relatively inexpensive dermatoscopes, suggests increased access to personal dermoscopy imaging devices.

A. Approaches to Detect the 7-point Checklist Criteria

Many previous works focus on detecting a single criterion from the 7-point checklist. For example, Mirzaalian et al. [11] detected absent, regular, and irregular streaks by enhancing streaks using Hessian based tubular filters. They tested on 99 dermoscopic images from the Argenziano et al. [12] dataset. Madooie et al. [13] detected the presence of blue white veils by mapping image regions to a discrete set of Munsell colors, using 223 images also selected from [12].

A few works detected the entire 7-point checklist. Fabbrocini et al. [14] detected all 7-point checklist criteria by designing separate pipelines that consider each criterion's unique characteristics. However, each pipeline adds complexity and requires careful tuning of hyper-parameters. For example, to detect irregular streaks, precise lesion border detection is required to compute an "irregularity" index, which considers how the lesion border differs from a straight line when the lesion is divided into segments. To detect irregular dots and globules, they applied statistical region merging [15] to find candidate dark segments, extracted morphological features, and applied thresholds (set experimentally) to detect rounded areas. Similar customized pipelines were set for all criteria. Wadhawan et al. [16] also proposed a system to detect all 7-point checklist criteria. Taking a machine learning approach, they extracted human engineered features (e.g., Haar wavelet,

J. Kawahara, S. Daneshvar, and G. Hamarneh are with Simon Fraser University.

G. Argenziano is with the Department of Dermatology, Second University of Naples, IT.

Manuscript submitted Oct 31, 2017.

local binary patterns, colour histograms) from a segmented region of interest. For each criterion, they selected a subset of features that correlated well with the criterion, and used these subsets to train a support vector machine. For evaluation, they considered 385 low difficulty images from the Argenziano et al. [12] dataset, out of which 347 could be segmented to create satisfactory lesion boundaries.

B. Approaches to Directly Classify Skin Conditions

Rather than detecting the 7-point checklist to infer a melanoma diagnosis, other works have explored directly classifying the disease from the image. For instance, the International Skin Imaging Collaboration's skin lesion classification challenge [17], asks participants to directly classify benign from malignant lesions. The top performing classification approach by Yu et al. [18] fine-tuned a Residual Neural Network [19] pretrained over ImageNet [20]. Over the DermoFit dataset [21], which is composed of 10 types of skin conditions in standardized clinical images, Kawahara et al. [22] demonstrated how using features from a neural network pretrained over ImageNet to classify skin diseases outperformed approaches that rely on handcrafted human engineered features [21], [23]. Over the Argenziano et al. [12] dataset, Menegola et al. [24] showed that fine-tuning a neural network pretrained only over ImageNet [20], performed better than training a neural network from scratch, or when pretrained over a dataset that included retinopathy images.

C. Contributions

This is the first work that predicts the entire 7-point criteria and the diagnosis (including the melanoma classification) in a single optimization, where predictions are derived from a multi-modal convolutional neural network that considers clinical, dermoscopic, and meta-data. Further, we show how our proposed deep architecture is used for three common tasks: classification, extracting feature vectors for image retrieval that consider clinical criteria, and localization of discriminate regions. We also publicly release the Argenziano et al. [12] dataset. While this dataset has been partly used in other publications (e.g., [13], [16], [24], [25]), it has not been readily available to the public. This dataset has been noted to have "excellent interobserver agreements" [26], and was used to teach dermatologists [9], [27], suggesting that it is a suitable source for training machine learning algorithms.

II. Methods

Given a dataset of skin lesions, we define each unique lesion as a *case*. The i-th case can have multiple types of information associated with it, such as a dermoscopy $x_d^{(i)}$ image, a clinical $x_c^{(i)}$ image, and patient meta-data $x_m^{(i)}$. Dermoscopic images x_d are captured with a dermatoscope and offer a standardized field of view and controlled acquisition (e.g., lighting and field of view). Clinical images x_c are less standardized, taken at various fields of view, and can contain image artefacts (e.g., a ruler to measure the lesion). Patient meta-data x_m includes other types of information, such as patient gender and lesion location.

Each case has a set of *categories* associated with it, assigned by a dermatologist. The categories are composed of labels for a diagnosis and the 7-point checklist criteria. The *diagnosis* $y^{(i)} \in \mathbb{Z}$ assigns an overall skin condition label to the image (e.g., melanoma, basal cell carcinoma). The *7-point checklist criteria* $z^{(i)} \in \mathbb{Z}^7$ consists of seven criteria that identify skin lesion properties that are indicative of melanoma [6], where the j-th criteria in the seven-point checklist $z_j^{(i)}$ has different labels associated with it. For example, "pigment network" is a 7-point checklist criteria with three labels: atypical, typical, and absent. We use the term *categories* to refer to both the diagnosis and the 7-point checklist criteria, while the term *labels* refers to items within a specific category. The full list of categories and labels is given in Table I.

A. Multi-Modal Multi-Task Loss Function

Rather than developing a separate model or pipeline for each individual category as is commonly done, we present a single model that predicts all labels within each category in a single optimization. We use a convolutional neural network (CNN), which consists of a designed architecture and a set of trainable parameters θ. Given the case data x (which represents different combinations of the input modalities), we define a *multi-task* loss function L for all eight (7-point checklist and a diagnosis) categories as,

$$L(x,y,z;\theta) = \ell(x,y;\theta) + \sum_{j=1}^{7} \ell(x,z_j;\theta) \tag{1}$$

where $\ell(\cdot)$ is the categorical cross-entropy loss defined as,

$$\ell(x,c;\theta) = -\frac{1}{b}\sum_{i=1}^{b}\sum_{j=1}^{J_c} w(c)_j \, c_j^{(i)} \log\left[\sigma(x^{(i)};\theta)_{c,j}\right] \tag{2}$$

where b is the number of cases in a mini-batch, $w(c)_j$ defined in Eq. 5 gives a higher weight to infrequent labels, J_c is the number of labels for the c-th category, and $\sigma(x^{(i)};\theta)_{c,j}$ is the probability predicted by the neural network parameterised by θ for the j-th label of the c-th category given input $x^{(i)}$. The multi-task loss (Eq. 1) is a function of the input modalities x; however, the available data x may vary by case (e.g., missing meta-data). In order to handle these cases, we define a *multi-modal multi-task* loss function that considers multiple combinations of the input modalities as,

$$\begin{aligned} L(x_d,x_c,x_m,y,z;\theta) &= L((x_d,x_c,x_m),y,z;\theta_{dcm}) \\ &+ L(x_d,y,z;\theta_d) + L((x_d,x_m),y,z;\theta_{dm}) \\ &+ L(x_c,y,z;\theta_c) + L((x_c,x_m),y,z;\theta_{cm}) \end{aligned} \tag{3}$$

where each multi-task loss $L(\cdot)$ (Eq. 1) is a function of different combinations of the input modalities. For example, the first term $L((x_d,x_c,x_m);\cdot)$ is a function of the dermoscopic, clinical, and meta-data. While the last term $L((x_c,x_m);\cdot)$ is a function of the clinical image and the meta-data (but not the dermoscopic image). We represent all parameters in the model as θ, and indicate those parameters that are updated based on the input type with the subscripts (e.g., θ_{dm} represents parameters updated based on x_d and x_m).

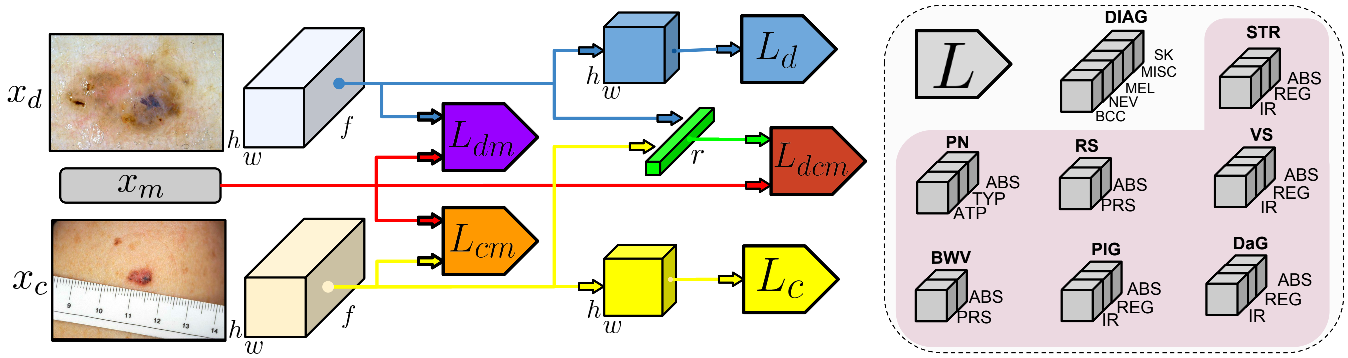

Fig. 1. The proposed architecture considers dermoscopic x_d, clinical x_c, and meta-data x_m when classifying all 7-point criteria and diagnosis. Each multi-task loss (L *block*) is trained on different combinations of the input modalities (e.g., L_{dm} is a function of x_m and x_d). As each L block gives predictions based on the data it was trained on, this single model is robust to missing data at inference time. The *blue* and *yellow* blocks immediately before the multi-task loss indicates the layer that is used to localize the discriminate regions. The *green bar* indicates the multi-modal feature vector used for image retrieval.

At inference time, as each multi-task loss function only depends on a subset of the input types, given a specific combination of the input modalities, we can use the predictions from the classification layer that matches the available input (e.g., if only a dermoscopic image is available, use the classification layer trained only on dermoscopic images). The architecture is further defined in Fig. 1 and in Section II-C.

During training, given a dataset of n cases, our goal is to learn the parameters * of the CNN that minimizes,

$$^* = \text{argmin} \underset{i}{X} L(x_c^{(i)}; x_d^{(i)}; x_m^{(i)}; z^{(i)}; y^{(i)}; \ ) + \ jj \ jj \quad (4)$$

where $L(\ )$ is defined as in Eq. 3, and $jj \ jj$ is the L2 norm regularization term weighted by $\ = 0{:}0005$ (experimenting with other $\ =$ [0.00005, 0.0001, 0.001] values yielded less than 1% differences in averaged accuracy and AUROC scores). In practice, Eq. 4 is often minimized using gradient descent with randomly sampled mini-batches of size b. However, in imbalanced datasets where the frequency of the labels greatly differs, training a model on randomly sampled mini-batches with imbalanced labels can lead to a model that is biased towards the majority class, as infrequent labels contribute little to the parameter updates.

B. Mini-Batches Sampled and Weighed by Label

To address the label imbalance problem, for each mini-batch with b cases, we ensure there exists at least k cases that belong to each unique label. To enforce this, each mini-batch is formed by randomly sampling with replacement k cases for each unique label. This causes the model weights to be updated based on all the unique labels in each gradient descent step. As we have 24 unique labels across all categories (Table I), this constrains our mini-batches to be of size $b = 24k$.

While sampling by labels improves class balance, labels within a mini-batch are still imbalanced since the category labels are not mutually exclusive, and including a case within one category, will also include its labels in all other categories. In order to further address class imbalance, we assign a higher weight to cases with labels that occur infrequently within a given mini-batch,

$$w(c)_j = \frac{\max(1c)}{(1c)_j} \quad (5)$$

where $c \ 2 \ Z^{b \times J_c}$ is a matrix representing b cases of 1-hot-encoded labels with J_c possible labels, $1 \ 2 \ Z^{1 \times b}$ is composed of all ones, $\max(1c)$ returns a scalar indicating the number of cases of the most frequent label, and $(1c)_j$ returns a scalar indicating the number of cases with the j-th label. Since each mini-batch has at least $k > 0$ labels, we avoid divide-by-zero errors and note how each computed weight is bound by $h1; \frac{b-(J_c-1)k}{k}i$. To derive the upper bound we note that the maximum value the numerator of Eq. 5 can take is $b \ \ (J_c \ \ 1)k$, where $(J_c \ \ 1)k$ is subtracted since there must be at least $(J_c \ \ 1)k$ cases with different labels in a single mini-batch (enforced in our sampling). The minimum value in the denominator of Eq. 5 is k (also enforced in the sampling). The lower bound is 1 since the value of the denominator in Eq. 5 cannot exceed the numerator.

C. Architecture to Classify, Localize, and Retrieve Images

In this section we describe the details of the layers used to form our model. We build upon a model pretrained over ImageNet [20], and remove the final output task-specific layer. We define this as our *base model*, which acts as a dense feature extractor, and outputs responses of size $h \ \ w \ \ f$ (height, width, and number of feature maps, respectively).

1) Classify and localize from a single modality: The following layers allows us to localize the discriminant regions in an input image, and to classify categories from a single image. For each category with l labels, we add a convolutional layer with filters of size $f \ \ 1 \ \ 1 \ \ l$, to the $h \ \ w \ \ f$ output of the base model. As in the work of Lin et al. [28], this layer is followed by a global spatial averaging pooling layer, where the categorical cross entropy loss (Eq. 1) is applied to the classification layers (Fig. 1 L_d, L_c). These pooled output responses classify using only a single image modality. In order to highlight the important image regions that contribute to the l-th label, we visualize (Fig. 4) the $h \ \ w$ responses (before spatial global pooling) at the l-th label (Fig. 1 top blue and bottom yellow blocks). Separate layers are created for the clinical and dermoscopic images with parameters $_c$, and $_d$.

2) Classify using image and meta-data: As the meta-data (gender, lesion location, and lesion elevation) is categorical, we one-hot encode the meta-data to produce a meta-data vector. In order to classify based on image and meta-data, we apply a global spatial averaging pooling layer to the $h \quad w$ output of the base model, apply batch normalization [29], and then concatenate the $1 \quad 1$ normalized visual responses with the one-hot encoded meta-data vector. We add a convolutional layer of size $f \quad 1 \quad 1 \quad l$ for each category, to form a classification layer (Fig. 1 $L_{dm}; L_{cm}$) used with the multi-task loss (Eq. 1). This is repeated for both the clinical and dermoscopic modalities to update $_{cm}$, and $_{dm}$.

3) Multi-modal feature vectors to retrieve and classify: We combine information from both clinical and dermoscopic images by adding a convolutional layer of size $f \quad 1 \quad 1 \quad r$ that takes as input the global average pooled visual responses of the clinical and dermoscopic specific models. This gives us an r-dimensional feature vector (Fig. 1 green bar) that is a function of both the clinical and dermoscopic images, which is used for multi-modal image retrieval (see Results). We concatenate the one-hot encoded meta-data to these visual pooled features, add a convolutional layer for each category, and apply the multi-task loss (Eq. 1) to form a final classification layer (Fig. 1 L_{dcm}) where parameters $_{dcm}$ are updated based on the clinical and dermoscopic images, and the meta-data.

D. Inferring a Melanoma Diagnosis

As our model both directly classifies the disease diagnosis, and classifies each of the 7-point checklist criteria, there are two ways to infer a melanoma diagnosis. The first is to *directly classify melanoma* from the diagnosis category, and the second is to *infer melanoma based on the 7-point criteria* [6]. To infer melanoma based on the 7-point criteria, given predictions $\hat{z}_j^{(i)}$ for the j-th 7-point criteria of the i-th case, we compute a melanoma score $S^{(i)}$, which, if exceeds a threshold t, indicates a prediction of melanoma,

$$\hat{y}^{(i)}_{7pt} = \begin{cases} melanoma; & \text{if } S^{(i)} \quad t \\ not\ melanoma; & \text{otherwise} \end{cases} \tag{6}$$

where, $S^{(i)} = \sum_{j=1}^{7} score(\hat{z}_j^{(i)})$

using a $score(\hat{z}_j)$ function that looks up the *7pt-score* from Table I that corresponds to the predicted 7-point label $\hat{z}_j^{(i)}$. The original threshold was $t = 3$ [6], which was later revised to $t = 1$ [8] in order to improve sensitivity [30]. We report results for both directly classifying melanoma, as well as inferring melanoma based on the 7-point checklist under varying thresholds in the Results section.

TABLE I
DETAILS OF THE DATASET. SECTION HEADERS INDICATE THE CATEGORIES. THE *abbrev* COLUMN INDICATES THE ABBREVIATION FOR THE LABEL; *name* REPRESENTS THE FULL NAME OF THE LABEL; *7pt-score* INDICATES THE CONTRIBUTION TO THE 7-POINT MELANOMA SCORE BY THE LABEL (WHERE "-" INDICATES NO CONTRIBUTION); AND, *# imgs* INDICATES HOW MANY IMAGES EXIST WITH THE PARTICULAR LABEL. WITHIN A CATEGORY, THE LABELS THAT ARE GROUPED TOGETHER IN OUR EXPERIMENTS ARE ASSIGNED THE SAME ABBREVIATION.

abbrev.	name	7pt-score	# imgs
	DIAGNOSIS (DIAG)		
BCC	basal cell carcinoma	-	42
NEV	blue nevus	-	28
NEV	clark nevus	-	399
NEV	combined nevus	-	13
NEV	congenital nevus	-	17
NEV	dermal nevus	-	33
NEV	recurrent nevus	-	6
NEV	reed or spitz nevus	-	79
MEL	melanoma	-	1
MEL	melanoma (in situ)	-	64
MEL	melanoma (less than 0.76 mm)	-	102
MEL	melanoma (0.76 to 1.5 mm)	-	53
MEL	melanoma (more than 1.5 mm)	-	28
MEL	melanoma metastasis	-	4
MISC	dermatofibroma	-	20
MISC	lentigo	-	24
MISC	melanosis	-	16
MISC	miscellaneous	-	8
MISC	vascular lesion	-	29
SK	seborrheic keratosis	-	45
	SEVEN POINT CRITERIA		
	1. Pigment Network (PN)		
ABS	absent	0	400
TYP	typical	0	381
ATP	atypical	2	230
	2. Blue Whitish Veil (BWV)		
ABS	absent	0	816
PRS	present	2	195
	3. Vascular Structures (VS)		
ABS	absent	0	823
REG	arborizing	0	31
REG	comma	0	23
REG	hairpin	0	15
REG	within regression	0	46
REG	wreath	0	2
IR	dotted	2	53
IR	linear irregular	2	18
	4. Pigmentation (PIG)		
ABS	absent	0	588
REG	diffuse regular	0	115
REG	localized regular	0	3
IR	diffuse irregular	1	265
IR	localized irregular	1	40
	5. Streaks (STR)		
ABS	absent	0	653
REG	regular	0	107
IR	irregular	1	251
	6. Dots and Globules (DaG)		
ABS	absent	0	229
REG	regular	0	334
IR	irregular	1	448
	7. Regression Structures (RS)		
ABS	absent	0	758
PRS	blue areas	1	116
PRS	white areas	1	38
PRS	combinations	1	99

III. Results

Our full dataset as described in Section II contains 1011 cases. We use 413 cases to train the model (Eq. 4), 203 cases to validate design decisions (i.e., set hyper-parameters), and 395 cases to test and report results. Subsets were chosen to ensure a similar distribution of categories. Four cases were missing clinical images, and were replaced with dermoscopic image. All images were resized to $512 \quad 512 \quad 3$.

TABLE II
THE ACCURACY OF EACH OF THE 7 POINT CRITERIA AND DIAGNOSIS. THE COLUMN *avg.* AVERAGES THE ACCURACY OVER EACH ROW.

Experiment	BWV	DaG	PIG	PN	RS	STR	VS	DIAG	avg.
frequent	81.0	44.8	56.5	39.5	73.2	65.1	79.2	55.4	61.8
x-unbalanced	87.6	56.7	65.6	68.1	78.2	75.9	81.3	68.4	72.7
x-balanced	87.3	60.3	64.8	68.9	78.2	75.7	81.5	70.9	73.4
x_c	79.2	52.7	56.5	57.0	71.6	60.3	75.2	60.0	64.1
x_c+x_m	77.7	51.9	59.2	59.5	72.9	62.8	76.7	61.5	65.3
x_d	85.8	60.8	62.8	69.4	77.5	71.4	80.3	71.9	72.5
x_d+x_m	85.1	59.7	63.3	69.4	76.7	74.2	81.5	73.4	72.9
x-combine	87.1	60.0	66.1	70.9	77.2	74.2	79.7	74.2	**73.7**
x_d+x_c-retrieve	86.8	56.7	62.8	65.3	78.0	73.4	81.0	71.1	71.9
Ngiam [31]	83.0	59.2	61.3	65.6	73.9	69.4	75.7	70.6	69.8
x_c	77.5	50.6	52.9	56.5	67.8	59.7	75.9	58.2	62.4
x_d	82.5	60.5	63.3	67.8	69.6	71.1	72.7	66.8	69.3

The original dataset contains labels at the most granular level (Table I). As some labels occur infrequently (e.g., two wreath vascular structure cases) and many labels have a similar clinical interpretation (e.g., types of benign nevi), we group infrequent labels with similar clinical interpretations into a single label. For example, in the diagnosis category, the NEV label groups all the nevi labels (e.g., blue nevus, clark nevus, etc) into a single label. We follow the same approach for the 7-point criteria where infrequent labels with similar clinical meaning and melanoma score contributions (i.e., a value in the *7pt-score* column in Table I) are grouped. For example, within the category vascular structures, we group linear irregular and dotted labels into a single irregular label IR as the presence of either is indicative of melanoma. The final label grouping is shown in the *abbrev* column in Table I.

To quantify the prediction performance of our method, for each category, we compute the prediction accuracy to indicate each category's overall performance (Table II). Accuracy, however, summarizes the performance over all labels, and may hide the performance of infrequent labels. Thus we also report detailed metrics for each label (Table III, IV, V).

We first report results using the most frequent training set labels as the test predictions in order to compute baseline results in the context of an imbalanced dataset. Table II (experiment *frequent*) shows that this simple approach yields an average accuracy of 61.8%, and thus model performance should be considered relative to this baseline.

Model Training: Our experiments use Inception V3 [32], pretrained over ImageNet [20] as our *base model*. We replace the class-specific layer with a trainable layer for each loss function as described in Section II-C and illustrated in Fig. 1. We augment the training images in real-time with flips, rotations, zooms, and height and width shifts. To train, we freeze all pretrained parameters, and train with a learning rate of 0.001 for 50 epochs, then reduce the learning rate to 0.0001 for 25 epochs, unfreeze the deepest frozen "inception block", and repeat for 25 epochs until all layers are unfrozen up to the second "inception block". Finally, we train for 25 epochs on un-augmented data, for a total of 300 epochs. We use Keras [33] with TensorFlow [34] to create and optimize our models. We optimized using stochastic gradient descent, with a decay of 1e-6, and momentum of 0.9. We observed that even though our model was trained with multiple loss functions (Eq. 3), it consistently reduced the loss over the training data.

Unbalanced vs. Balanced Training: We compare the performance of a model trained on balanced data by first training a model using random mini-batches with uniform class weights, and report results under the experiment name *x-unbalanced*. Unless otherwise stated, results are computed using the predictions that are a function of the entire input (i.e., Fig. 1 L_{dcm}). We compare *x-unbalanced*, to the same model trained on mini-batches sampled and weighted by label (described in Sec II-B) using the experiment name *x-balanced*. When training with balanced mini-batches, we observe a small increase in overall accuracy; however, as noted earlier, accuracy does not well highlight improvements made to classifying infrequent labels. The averaged metrics across all labels increase for the 7-point checklist (Table III) and diagnosis (Table IV), when compared to the model trained without balancing the classes (*x-unbalanced*). Notably, *x-balanced* improves the sensitivity and precision of detecting irregular vascular structure (VS IR in Table III) from 0% in *un-balanced* for both, to 10% and 60%, respectively. A similar performance increase is seen in the sensitivity (5.3% to 21.1%) and precision (33.3% to 50%) of detecting seborrheic keratosis (SK in Table IV). To compare the performance of the imbalanced experiment (i.e., *x-unbalanced*) with the balanced experiment (i.e., *x-balanced*), we apply a Friedman test [35] using AUROC scores for each category, where the AUROC scores are averaged within each category, and obtained a statistically significant difference between the two models ($p = 0.0047$).

Performance Based on Input: To examine the classification performance as a function of the input modalities, we report the average accuracy using the predicted responses from different classification layers (L blocks in Fig. 1). We see the average classification accuracy using the clinical images and meta-data (experiment x_c and x_c+x_m in Table II) is much lower than when using the dermoscopic images and meta-data (experiment x_d and x_d+x_m). Dermoscopic images likely yield higher classification accuracy since the 7-point checklist was designed to detect features visible under dermoscopy. The classification layer that uses clinical, dermoscopic, and meta-data together yields the highest average accuracy. However, we note including clinical images gives relatively small improvements over using dermoscopic images alone, and that this improvement may be partly due to the additional layer that joins the two modalities. Our observations differ from those reported by Ge et al. [36], which showed larger accuracy improvements when incorporating clinical and dermoscopic images into a single model, as well as similar diagnosis performance for each modality. We note Ge et al. [36] report results over a larger dataset, which may in part explain our different observations. Our results, separated by input modalities, illustrate that our approach degrades gracefully with missing data, making it applicable to scenarios where only partial patient data is available.

Other Multi-Modal Approaches: Our approach of using multiple losses that are a function of different input modalities (Eq. 3), differs from other multi-modal approaches such as the work by Ngiam et al. [31], which randomly sets some input modalities to zero during training. We perform an additional

TABLE III
THE SEVEN POINT CRITERIA RESULTS. COLUMNS INDICATE THE SEVEN-POINT CRITERIA, SEPARATED BY THE LABELS THAT BELONG WITHIN EACH CRITERIA. THE FINAL *avg.* COLUMN IS THE RESULT AVERAGED OVER THE ENTIRE ROW. EACH ROW REPRESENTS AN EXPERIMENT, DIVIDED INTO RESULTS FOR SENSITIVITY (*sens.*), SPECIFICITY (*spec.*), PRECISION (*prec.*), AND AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (*auroc.*). LABEL ABBREVIATIONS ARE DEFINED IN TABLE I.

7pt criteria		BWV		DaG			PIG			PN			RS		STR			VS			
Experiment.	met.	ABS	PRS	ABS	REG	IR	ABS	REG	IR	ABS	TYP	ATP	ABS	PRS	ABS	REG	IR	ABS	REG	IR	Avg.
x-unbalanced	sens.	96.6	49.3	34.0	59.3	67.8	83.0	6.2	57.3	78.8	77.4	35.5	95.5	31.1	98.1	36.4	34.0	98.7	23.1	0.0	55.9
	spec.	49.3	96.6	92.2	72.2	67.4	53.5	99.4	80.1	80.8	75.5	93.7	31.1	95.5	47.8	98.6	94.0	22.0	97.1	100.0	76.1
	prec.	89.0	77.1	59.6	47.6	62.8	69.8	60.0	56.8	72.8	64.9	63.5	79.1	71.7	77.8	76.2	64.0	82.8	54.5	0.0	64.7
	auroc	87.0	87.0	72.3	72.6	76.4	77.4	67.2	78.1	87.8	83.6	78.6	79.9	79.9	84.2	87.8	78.3	82.1	81.8	73.4	79.8
x-balanced	sens.	92.5	65.3	43.0	66.1	66.1	73.5	16.7	67.7	78.2	76.0	41.9	84.1	62.3	90.7	43.2	50.0	96.8	30.8	10.0	60.8
	spec.	65.3	92.5	89.8	75.1	73.4	64.5	98.3	73.4	81.6	77.9	92.1	62.3	84.1	63.8	97.4	87.7	31.7	95.6	99.5	79.3
	prec.	91.9	67.1	58.9	53.1	66.9	72.9	57.1	53.8	73.5	66.9	61.9	85.9	58.9	82.3	67.9	56.0	84.4	51.6	60.0	**66.9**
	auroc	87.5	87.5	73.0	76.5	78.0	78.8	75.2	79.4	88.6	83.6	78.9	83.5	83.5	84.9	87.1	78.7	85.0	84.0	76.1	81.6
x-combine	sens.	89.4	77.3	47.0	67.8	62.1	77.6	29.2	59.7	77.6	78.1	48.4	81.3	66.0	86.0	54.5	51.1	92.3	42.3	13.3	**63.2**
	spec.	77.3	89.4	87.8	72.6	78.9	65.1	94.2	80.1	85.8	78.7	90.7	66.0	81.3	67.4	96.0	85.7	45.1	92.4	97.5	**80.6**
	prec.	94.4	63.0	56.6	51.3	70.5	74.2	41.2	57.8	78.1	68.3	61.6	86.7	56.5	83.1	63.2	52.7	86.5	45.8	30.8	64.3
	auroc	89.2	89.2	74.1	76.5	79.9	79.0	74.9	79.0	89.9	84.2	79.9	82.9	82.9	86.1	87.0	78.9	86.2	85.5	76.1	**82.2**
x_d+x_c retrieve	sens.	91.9	65.3	36.0	63.6	63.8	77.1	18.8	54.0	73.1	71.9	41.9	88.9	48.1	85.6	52.3	50.0	92.7	40.4	30.0	60.3
	spec.	65.3	91.9	91.5	70.0	71.1	59.3	96.0	76.8	79.9	77.5	89.1	48.1	88.9	66.7	95.2	86.0	45.1	93.9	97.5	78.4
	prec.	91.9	65.3	59.0	47.5	64.2	71.1	39.1	51.5	70.4	65.2	54.2	82.4	61.4	82.7	57.5	52.8	86.6	50.0	50.0	63.3

TABLE IV
THE RESULTS FOR THE DIAGNOSIS CATEGORY, AND FOR MELANOMA PREDICTION BASED ON THE PREDICTED SEVEN POINT SCORES. THE FINAL COLUMNS *Mel7* SHOWS THE RESULTS USING THE SCORES FROM THE PREDICTED SEVEN-POINT CHECKLIST TO PREDICT MELANOMA USING TWO COMMON THRESHOLDS, $t = 1$ AND $t = 3$.

		DIAG						Mel7	
Experiment	met.	BCC	NEV	MEL	MISC	SK	Avg.	t=1	t=3
x unbalanced	sens.	25.0	94.1	44.6	35.0	5.3	40.8	90.1	47.5
	spec.	98.4	50.6	92.2	98.0	99.5	87.7	40.1	87.4
	prec.	40.0	70.3	66.2	66.7	33.3	55.3	34.1	56.5
	auroc	92.2	87.7	83.2	86.3	88.4	87.6	76.8	
x-balanced	sens.	25.0	91.3	55.4	42.5	15.8	46.0	96.0	69.3
	spec.	98.9	62.5	88.4	97.2	99.7	89.4	33.0	78.9
	prec.	50.0	75.2	62.2	63.0	75.0	65.1	33.0	53.0
	auroc	89.2	88.1	84.2	86.8	90.4	87.7	81.7	
x-combine	sens.	62.5	88.6	61.4	47.5	42.1	**60.4**	96.0	69.3
	spec.	97.9	71.6	88.8	97.5	99.5	**91.0**	36.1	77.6
	prec.	55.6	79.5	65.3	67.9	80.0	**69.6**	34.0	51.5
	auroc	92.9	89.7	86.3	88.3	91.0	**89.6**	81.6	
x_d+x_c retrieve	sens.	37.5	87.2	59.4	42.5	36.8	52.7	94.1	73.3
	spec.	97.9	69.9	87.8	97.2	98.1	90.2	36.1	78.6
	prec.	42.9	78.3	62.5	63.0	50.0	59.3	33.6	54.0

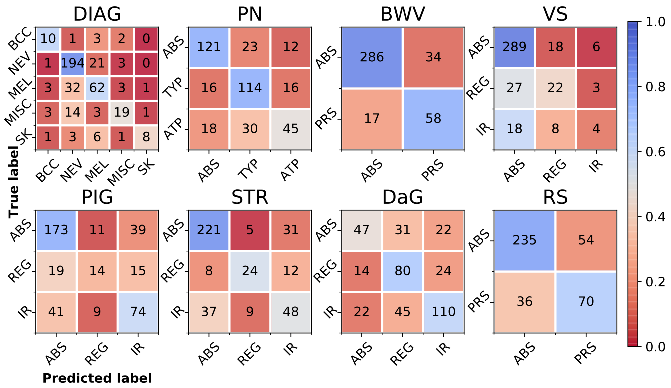

Fig. 2. Confusion matrices for each category using the test set predictions from our proposed model. The y-axis indicates the ground truth labels. The x-axis indicates the model's predicted labels. Numbers in each entry represent the number of cases classified as such. Colors indicate the percentage of each label in each entry, normalized by the total number of true labels.

experiment based on Ngiam et al.'s work [31], where on average we set a single input modality to zero in 75% of the samples within a mini-batch. The other 25% includes all three modalities. We remove all loss functions except for L_{dcm} (Fig. 1), and repeat the *x-balanced* experiment. We report test results in Table II for predictions based on all three modalities, only clinical images, and only dermoscopic images. We obtain consistently higher averaged accuracy in our proposed approach for each type of input. One possible reason for our improvement is that Ngiam et al. [31] learn a model that is robust to missing data, which may compete with learning disease patterns specific to a single modality. Whereas our approach may learn patterns specific to each modality, as the loss functions are trained on each individual modality.

Combining Classification Layers' Predictions: We also report results from averaging the predicted probabilities of the three classification layers that are a function of dermoscopic images (Fig. 1 $L_d; L_{dm}; L_{dcm}$) into a final prediction (experiment name *x-combine*). While this results in a minor decrease to the average precision over the 7-point checklist when compared to *x-balanced*, average sensitivity increases, and all metrics are increased in the diagnosis category (Table IV). This approach of combining multiple classification layers is analogous to averaging the predictions from multiple independent neural networks; however, our model shares most layers. We use the *x-combine* predictions to form the confusion matrices in Fig. 2 and the ROC curves in Fig. 3 (*left*).

Inferring Melanoma: As noted in the Methods section, we can infer a melanoma diagnosis by either direct classification, or via the 7-point checklist (Eq. 6). Fig. 3 (*right*) shows the ROC curve from directly diagnosing melanoma, and from thresholds based on the 7-point score (i.e., t in Eq. 6). We see that directly classifying melanoma yields a higher AUROC score than the predicted 7-point scores. However, at high sensitivity levels, the performance of both approaches are similar. In addition, the direct classification AUROC score comes at the cost of a less interpretable model, as this ROC curve is based on thresholding probabilities for a binary decision

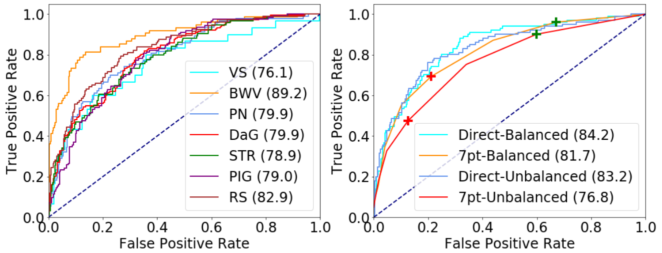

Fig. 3. *(Left)* One-vs-all ROC curves for each label in the 7-point criteria that contribute to melanoma. *(Right)* Melanoma ROC curves comparing direct melanoma classification with inference via the 7-point checklist, using unbalanced and balanced training procedures. The green and red cross indicates the threshold of 1 and 3, respectively, used in Eq. 6.

TABLE V
RELATED WORKS SEPARATED BY CATEGORY AND LABELS. WE REPORT THE AGGREGATED METRICS USED IN THE ORIGINAL WORKS. *rep* INDICATES IF WE COULD REPLICATE THE SAME TRAINING/TEST IMAGES AND REPORT A DIRECT COMPARISONS. (*METRIC AVERAGED BY WEIGHTED SAMPLE. OTHER METRICS ARE UNWEIGHTED AVERAGES, EXCEPT FOR THE BINARY CASES OF *sens*, *spec*, AND *prec*.)

	rep.	category (labels)	acc.	sens.	spec.	prec.	auroc
Sadeghi [25]	✗	STR	76.1	76.0*	-	74.2*	85.0*
ours		(ABS, REG, IR)	74.2	74.2*	74.9*	73.6*	84.5*
Wadhawan [16]	✗	BWV	-	79.5	79.2	-	-
ours		(ABS, PRS)	87.1	77.3	89.4	63.0	89.2
Wadhawan [16]	✗	RS	-	64.2	67.9	-	-
ours		(ABS, PRS)	77.2	66.0	81.3	71.6	82.9
Menegola [24]	✓	DIAG (BCC,	-	-	-	-	84.5
ours		MEL, Other)	80.8	64.9	84.8	74.6	88.5

(melanoma vs all), which is less clinically interpretable than the 7-point scores. We highlight that our approach outputs both results, and either diagnoses approach can be used. Finally, Fig. 3 (*right*) also shows that our *x-balanced* training improves melanoma detection for both approaches.

Works Using the Same Data: Comparing with other approaches is challenging as often different subsets of the data are used from various sources, with multi-class labels grouped to form binary problems. We compared with works that used the same dataset, and that reported the same class labels as our work. This is reported in Table V, along with a checkmark indicating if the exact subsets of the data used in this work was publicly available, allowing for a direct comparison. Sadeghi et al. [25] classified absent, regular, and irregular streaks using 945 images, of which 745 are from the same dataset as our work. Wadhawan et al. [16] used 347 "low difficulty" images from the same dataset as our work, and we compare with the two categories that we both report binary labels on. Our results do not exclude challenging images and do not rely on lesion segmentations. Menegola et al. [24] make the image names and cross-validation folds publicly available. We run new experiments using the same image names, perform 5 rounds of 2-fold cross validation based on their provided folds, and modify our diagnosis loss function to follow their 3-class experiment (melanoma vs. basal cell carcinoma vs other benign lesions). We follow the same training and inference procedure as *x-combine* and compare with their top performing approach. Table V suggests our model achieves results comparable to state-of-the-art among differing categories.

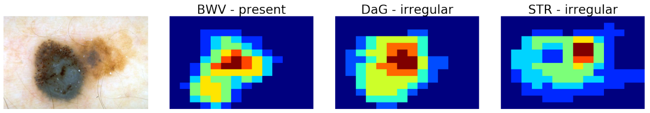

Fig. 4. Learned responses localize the image areas that contribute to the specific class label for a given input image.

Localization: With the goal of providing a model whose classification may be interpretable by humans, we visualize the areas of the image that contribute to the predicted label by viewing the learned h w responses that correspond to the l-th label. Given an image, we re-size the h w responses (e.g., x_d responses are represented by the top blue box in Fig. 1) to match the size of the original image, and show the response for select labels in Fig. 4. By visualizing those areas that influence the classification, users can check for the presence of these features in the detected areas and adjust their confidence in the machine's prediction accordingly. Ge et al. [36] used a similar approach based on class activation maps [37], to visualize the diagnosis category, while here we show localized results for clinical criteria.

Image Retrieval: We demonstrate our approach is able to retrieve clinically similar images with respect to the 7-point criteria and diagnosis (Fig. 5). For each image, we extract the r-dimensional responses (Fig. 1 green rectangle) that are a function of both the clinical and dermoscopic images (Sec. II-C). For each test case, we find the training case feature vector with the lowest cosine distance, and use the known training labels as our predictions (experiment x_c+x_d-*retrieve*). Kawahara et al. [38] used a similar approach to retrieve a path of visually similar images; whereas, this works learns a new compact multi-modal feature vector. This image retrieval approach achieves comparable averaged results (Table. II, III, IV) with the classification based approach. However, image retrieval has the additional advantage of allowing users to infer labels from expertly labeled images, rather than relying on a black box classification system, and may prove more interpretable than classification or localization approaches. We note how our multi-modal r-dimensional feature vector (Fig. 1 green bar) retrieves multiple modalities with a *single* vector, and that our loss function (Eq. 3) learns compact feature vectors that considers several clinically relevant criteria.

IV. CONCLUSION

We propose a neural network designed for multi-modal images and meta-data, that classifies all seven-point checklist criteria and skin lesion diagnosis within a single optimization (multi-task). Our architecture uses multiple loss functions to handle combinations of the input modalities, and at inference time is capable of making predictions with missing data. Further, our architecture is capable of localizing discriminate information, and produces feature vectors useful for image retrieval of clinically similar images. We observe that, for some criteria, our model is unable to distinguish among the labels (e.g., model almost always predicts absent vascular structures). We see these as active areas for improvement and hope for