Classification of Ten Skin Lesion Classes: Hierarchical KNN *versus* Deep Net

Robert B. Fisher[1(✉)], Jonathan Rees[1], and Antoine Bertrand[2]

[1] University of Edinburgh, Edinburgh, Scotland
rbf@inf.ed.ac.uk
[2] INP Grenoble, Grenoble, France

Abstract. This paper investigates the visual classification of the 10 skin lesions most commonly encountered in a clinical setting (including melanoma (MEL) and melanocytic nevi (ML)), unlike the majority of previous research that focuses solely on melanoma *versus* melanocytic nevi classification. Two families of architectures are explored: (1) semi-learned hierarchical classifiers and (2) deep net classifiers. Although many applications have benefited by switching to a deep net architecture, here there is little accuracy benefit: hierarchical KNN classifier 78.1%, flat deep net 78.7% and refined hierarchical deep net 80.1% (all 5 fold cross-validated). The classifiers have comparable or higher accuracy than the five previous research results that have used the Edinburgh DERMOFIT 10 lesion class dataset. More importantly, from a clinical perspective, the proposed hierarchical KNN approach produces: (1) 99.5% separation of melanoma from melanocytic nevi (76 MEL & 331 ML samples), (2) 100% separation of melanoma from seborrheic keratosis (SK) (76 MEL & 256 SK samples), and (3) 90.6% separation of basal cell carcinoma (BCC) plus squamous cell carcinoma (SCC) from seborrheic keratosis (SK) (327 BCC/SCC & 256 SK samples). Moreover, combining classes BCC/SCC & ML/SK to give a modified 8 class hierarchical KNN classifier gives a considerably improved 87.1% accuracy. On the other hand, the deepnet binary cancer/non-cancer classifier had better performance (0.913) than the KNN classifier (0.874). In conclusion, there is not much difference between the two families of approaches, and that performance is approaching clinically useful rates.

Keywords: Skin cancer · Melanoma · RGB image analysis

1 Introduction

The incidence of most types of skin cancer is rising in fair skinned people. The causes of the increase are not certain, but it is hypothesized that increased ultraviolet exposure and increasing population ages are the main causes. Irrespective of the cause, skin cancer rates are increasing, as is awareness of skin cancer. This increased awareness has also led to increased reporting rates. In addition to the health risks associated with cancer, a second consequence is the increasing

© Springer Nature Switzerland AG 2020
Y. Zheng et al. (Eds.): MIUA 2019, CCIS 1065, pp. 86–98, 2020.
https://doi.org/10.1007/978-3-030-39343-4_8

medical cost: more people are visiting their primary medical care practitioner with suspicious lesions, and are then forwarded onto dermatology specialists. As many, perhaps a majority, of the referrals are for normal, but unusual looking, lesions, this leads to a considerable expense. Eliminating these unnecessary referrals is a good goal, along with improving outcomes.

A second issue is that there are different types of skin cancer, in part arising from different cell types in the skin. Most people are familiar with melanoma, a dangerous cancer, but it is considerably less common than, for example, basal cell carcinoma. Because of the rarity of melanoma, a primary care practitioner might only encounter one of these every 5–10 years, leading to the risk of over or under referring people onto a specialist. Hence, it is good to have tools that can help discriminate between different skin cancer types.

A third issue is the priority of referrals, which might be routine or urgent. Melanoma and squamous cell carcinoma metastasize and are capable of spreading quickly, and thus need to be treated urgently. Other types of skin cancer grow more slowly, and may not even need treatment. Moreover, there are many normal lesion types that may look unusual at times. This also motivates the need for discrimination between the types of cancer and other lesions.

This motivates the research presented here, which classifies the 10 lesion types most commonly encountered by a general practice doctor. Because of the healthcare costs arising from false positives, and the health and potentially life costs of incorrect decisions, even small improvements in performance can result in considerable cost reduction and health increases.

The research presented here uses standard RGB camera data. There is another research stream based on dermoscopy [6], which is a device typically using contact or polarized light. This device can give better results than RGB image data, but has been typically limited to melanoma *versus* melanocytic nevus (mole) discrimination. Here, we focus on 10 types of lesion instead of 2, so use RGB images.

This paper presents two approaches to recognizing the 10 classes. The first more traditional approach is based on a combination of generic and hand-crafted features, feature selection from a large pool of potential features, and a hierarchical decision tree using a K-nearest neighbor classifier. A second deepnet approach is also presented for comparison. The cross-validated hierarchical 10 class accuracy (78.1%) and the cross-validated deepnet with refinement accuracy (80.1%) are comparable to the best previous performances. **The key contributions of the paper are: (1) a hierarchical decision tree structure and associated features best suited for discrimination at each level, (2) a deepnet architecture with BCC/SCC and ML/SK refinement that has 2% better performance, (3) improved classification accuracy (87.1% for 8 merged classes as compared to 78.1% for 10 classes), (4) a malignant melanoma *versus* benign nevi classification accuracy of 99.5%, (5) a malignant melanoma *versus* seborrheic keratosis classification accuracy of 100%, and (6) a clinically relevant BCC/SCC *versus* seborrheic keratosis classification accuracy of 90.6%.**

2 Background

There is a long history of research into automated diagnosis of skin cancer, in part because skin cancer is the most common cancer [1]. A second factor is the lesions appear on the skin, thus making them amenable to visual analysis. The most commonly investigated issues are (1) the discrimination between melanoma (the most serious form of skin cancer) and melanocytic nevi (the most-commonly confused benign lesion) and (2) the segmentation of the boundary between normal and lesion skin (typically because the boundary shape is one factor commonly used for human and machine diagnosis). The most commonly used imaging modalities are color and dermoscopy (a contact sensor) images. A general review of this research can be found in [11,13,14].

A recent breakthrough is the Stanford deep neural network [5], trained using over 129K clinical images (including those used here), and covering over 2000 skin diseases. Their experiments considered four situations: (1) classification of a lesion into one of 3 classes (malignant, benign and non-neoplastic (which is not considered in the work presented here)), (2) refined classification into 9 classes (5 of which correspond to the classes considered here), (3) keratinocyte carcinomas (classes BCC and SCC here) *versus* benign seborrheic keratoses (class SK here) and (4) malignant melanoma (class (MEL) *versus* melanocytic nevi (class ML). In case 1, their deep net achieved 0.721 accuracy as compared to the dermatologist accuracy of approximately 0.658. In case 2, the deep net achieved 0.554 accuracy as compared to the dermatologist accuracy of 0.542. In cases 3 and 4, accuracy values are not explicitly presented, but from the sensitivity/specificity curves, one can estimate approximately 0.92 accuracy for the deepnet approach, with the dermatologists performing somewhat worse.

One important issue raised in [11] is the absence of quality public benchmark datasets, especially covering more than the classification of melanoma *versus* melanocytic nevi. From 2013, the Edinburgh Dermofit Image Library [2] (1300 lesions in 10 classes, validated by two clinical dermatologists and one pathologist - more details in Sect. 3) has been available. It was part of the training data for the research of Esteva *et al.* described above, and is the core dataset for the research results described below.

The first result was by Ballerini *et al.* [2] which investigated the automated classification of 5 lesion classes (AK, BCC, ML, SCC, SK - see Sect. 2 for labels), and achieved 3-fold cross-validated accuracy of 0.939 on malignant *versus* benign, and 0.743 over 960 lesions from the 5 classes. A 2 level hierarchical classifier was used. Following that work, Di Leo [4] extended the lesion analysis to cover all 10 lesion classes in the dataset of 1300 lesions. That research resulted in an accuracy of 0.885 on malignant *versus* benign and 0.67 over all lesions from the 10 classes.

More recently, Kawahara *et al.* [9] developed a deep net to classify the 10 lesion classes over the same 1300 images. The algorithm used a logistic regression classifier applied to a set of features extracted from the final convolutional layers of a pre-trained deep network. Multiple sizes of image were input to enhance scale invariance, and each image was normalized relative to its mean RGB value to enhance invariance to skin tone. As well as substantially improving the 10

lesion accuracy, the proposed method did not require segmented lesions (which was required by the approach proposed here). Follow-on research by Kawahara and Hamarneh [10] developed a dual tract deep network with the image at 2 resolutions through the two paths, which were then combined at the end. Using auxiliary loss functions and data augmentation, the resulting 10-class performance was 0.795, which was stated as an improvement on the methods of [9], although the new methodology used less training data and so the initial baseline was lower. A key benefit of the multi-scale approach was the ability to exploit image properties that appear at both scales. As with the original research, the proposed method did not require segmented lesions.

3 Edinburgh DERMOFIT Dataset

The dataset used in the experiments presented in this paper was the Edinburgh DERMOFIT Dataset[1] [2]. The images were acquired using a Canon EOS 350D SLR camera. Lighting was controlled using a ring flash and all images were captured at the same distance (approximately 50 cm) with a pixel resolution of about 0.03 mm. Image sizes are typically 400 × 400 centered on the cropped lesions plus about an equal width and height of normal skin. The images used here are all RGB, although the dataset also contains some registered depth images. The ground truth used for the experiments is based on agreed classifications by two dermatologists and a pathologist. This dataset has been used by other groups [4,5,9,10], as discussed above.

The dataset contains 1300 lesions from the 10 classes of lesions most commonly presented to consultants. The first 5 classes are cancerous or precancerous: actinic keratosis (AK): 45 examples, basal cell carcinoma (BCC): 239, squamous cell carcinoma (SCC): 88, intraepithelial carcinoma (IEC): 78, and melanoma (MEL): 76. The other five classes are benign, but commonly encountered: melanocytic nevus/mole (ML): 331 examples, seborrheic keratosis (SK): 257, pyogenic granuloma (PYO): 24, haemangioma (VASC): 96, and dermatofibroma (DF): 65.

4 Hierarchical Classifier Methodology

The process uses RGB images, from which a set of 2500+ features are extracted. The key steps in the feature extraction are: (1) specular highlight removal, (2) lesion segmentation, (3) feature extraction, (4) feature selection, (5) hierarchical decision tree classification. Because some lesions had specular regions, the combination of the ring-flash and camera locations results in specular highlights. These were identified ([2], Section 5.2) using thresholds on the saturation and intensity. Highlight pixels were not used in the feature calculations.

Lesion segmentation used a binary region-based active contour approach, using statistics based on the lesion and normal skin regions. Morphological opening was applied afterwards to slightly improve the boundaries. Details can be found in [12]. This produced a segmentation mask which covers the lesion.

[1] homepages.inf.ed.ac.uk/rbf/DERMOFIT/datasets.htm.

4.1 Feature Calculation

The majority of the 2551 features are calculated from generalized co-occurrence texture matrices. More details of the features are given in Section 4.2 of [2]. The texture features are described by "XY FUNC DIST QUANT", where X,Y ∈ { R, G, B }, { L, a, b } or { H, S, V } gives the co-occurring color channels from the lesion, DIST is the co-occurring pixel separation, (5, 10, ... 30), QUANT is the number of gray levels ∈ {64, 128, 256}. The co-occurrence matrices are computed at 4 orientations and then averaged. From each matrix, 12 summary scalar features are extracted, including FUNC ∈ { Contrast, ClusterShade, Correlation, Energy, MaxProbability, Variance}, as described in [7]. As well as using these features directly, the difference (l-s: lesion-normal skin) and ratio (l/s: lesion/normal skin) features were computed.

After these features were calculated, a z-normalization process was applied, where the mean and standard deviation were estimated from the inner 90% of the values. Features above or below the 95^{th} or 5^{th} percentile were truncated. Some of the top features selected (see next section) for use by the decision tree are listed in Table 1.

Because there is much correlation between the color channels, and the different feature scales and quantizations, a feature reduction method was applied. The feature calculation process described above resulted in 17079 features. These features were cross-correlated over the 1300 lesions. The features were then sequentially examined. Any feature whose absolute correlation was greater than 0.99 with a previously selected feature was removed. This reduced the potential feature set to 2489. Some additional lesion-specific features (for each of R, G, B) were added to give 2551 features (also normalized):

- Ratio of mean lesion periphery to lesion center color
- Ratio of mean lesion color to non-lesion color
- Std dev of lesion color
- Ratio of lesion color std dev to non-lesion color std dev
- Ratio of mean lesion color to lesion color std dev
- Six gray-level moment invariants
- Given a unit sum normalized histogram of lesion pixel intensities H_l and normal skin pixel intensities H_n, use features $mean(H_l. - H_n)$, $std(H_l. - H_n)$, $mean(H_l./H_n)$, $std(H_l./H_n)$, $H_l. - H_n$ and $H_l./H_n$. The latter 2 features are histograms and the Bhattacharyya distance is used.

4.2 Feature and Parameter Selection

From the 2551 initial features, greedy Forward Sequential Feature Selection (`en.wikipedia... .org/wiki/Feature_selection`) is used to select an effective subset of features for each of the 9 tests shown in Fig. 1. This stage results in 2–7 features selected for each test. Table 2 lists the number of features selected and the top 2 for each of the tests. The full set of features used can be

Table 1. Overview of top selected features. See text body for details.

68	GB Correlation d10 L64	222	HH ClusterShade d5 L128
245	HH Energy d5 L256	267	HH Correlation d30 L256
272	aa Contrast d5 L64	523	ns SS Correlation d15 L64
622	ns bb Homogeneity d10 L64	630	ns bb Correlation d5 L128
832	l-s SS Dissimilarity d5 L64	834	l-s HH Variance d5 L64
950	l-s aa Autocorrelation d5 L64	959	l-s La Variance d5 L64
1467	l/s HH ClusterShade d25 L256	1536	l/s La Variance d5 L64
1556	l/s aa Contrast d15 L64	1824	so sigma Im G s3 n4
2303	M-m lo sigma Re R s1 G s1	2422	sigma Im G s1
2433	l-s mu Re G s1	2441	l/s mu Re G s1
2503	G mean(l)/std(l)	2526	R std(hist(l)./ hist(s))

seen at: `homepages.inf.ed.ac.uk/rbf/... DERMOFIT/SCusedfeatures.htm`. We tersely describe in Table 1 the top features of the tests. The tests use a K-Nearest Neighbor classifier with a Euclidean distance measure $\sum_r (x_r - n_r)^2$ where x_r is the r^{th} property of the test sample and n_r is from a neighbor sample.

Also included in the parameter optimization stage was the selection of the optimal value K value to use in the K-Nearest Neighbor algorithm. The K values reported for the different tests reported in Table 2 were found by considering odd values of K from 3 to 19. The best performing K was selected over multiple cross-validated trials, but generally there was only 1–3% variation in the results for K in 7–19. Performance evaluation, and parameter and feature selection used 5-fold cross validation (using the Matlab `cvpartition` function), with 1 of the 5 subsets as an independent test set. The splits kept the lesion classes balanced.

4.3 Hierarchical Decision Tree

The lesion classification uses a hierarchical decision tree, where a different K-NN is trained for each decision node. Other classifiers (*e.g.* Random Forest or multi-class SVM) or decision node algorithms (*e.g.* SVM) could be investigated, but we chose a K-NN because of the intuition that there was much in-class variety and so data-driven classification might perform better. Several varieties of deepnet based classification are presented in the next section.

The choice of branching in the tree is motivated partly by clinical needs (*i.e.* cancer *versus* non-cancer) and partly by classifier performance (*i.e.* the lower levels are chosen based on experimental exploration of performance). Figure 1 shows the selected decision tree. Exploration (based on experience rather than a full systematic search) of different cancer subtree structures showed that the PYO/VASC branch was most effectively isolated first, with then a two way split between IEC, MEL and DF *versus* the rest. These two initial decisions could be performed almost perfectly, thus reducing the decision task to smaller decisions (and without propogating errors).

While a general classification is valuable, from a clinical perspective several more focussed binary decisions are important, namely melanoma (MEL) *versus* melanocytic nevi (ML), melanoma (MEL) *versus* seborrheic keratosis (SK), and basal cell carcinoma (BCC) plus squamous cell carcinoma (SCC) *versus* seborrheic keratosis (SK). We implemented these binary decisions each with a single test, after again doing feature selection, also as reported in Table 2.

Table 2. Top 2 features for each of the key K-NN decisions in the decision tree and cross-validated performance on ground-truthed data.

Test	K	Num of features used	Feat 1	Feat 2	Accuracy
PYO/VASC vs rest	11	5	2503	834	0.974
MEL/IEC/DF vs rest	13	2	2433	2441	1.000
MEL vs IEC/DF	19	3	959	832	0.831
AK/BCC/SCC vs ML/SK	17	9	222	1536	0.916
AK vs BCC/SCC	17	4	2422	272	0.876
PYO vs VASC	15	3	523	622	0.852
IEC vs DF	15	7	630	1467	0.888
BCC vs SCC	15	7	245	1556	0.814
ML vs SK	11	7	950	267	0.850
MEL vs ML	9	2	2303	1824	0.995
MEL vs SK	3	1	2433		1.000
BCC/SCC vs SK	11	5	68	2526	0.906

5 Decision Tree Experiment Results

Evaluation of the classification performance using the decision tree presented above used leave-one out cross-validation. The confusion matrix in Table 4 summarizes the performance for the detailed classification results over the 10 classes. The mean accuracy over all lesions (micro-average - averaging over all lesions) was 0.781 and the accuracy over all classes (macro-average - averaging over the performance for each class) was 0.705. Mean sensitivity is 0.705 and mean specificity is 0.972 (when averaging the sensitivities and specificities of each class over all 10 classes). A comparison of the results with previous researchers is seen in Table 3.

The new 10 class results are comparable those of [9,10] and considerably better than the others. Combining classes BCC/SCC & ML/SK to give an 8 class decision produces considerably better results. "Kawahara and Hamarneh [10] repeated the experiments from Kawahara *et al.* [9], but changed the experimental setup to use half the training images, and omitted data augmentation in order to focus on the effect of including multi-resolution images." (private communication from authors).

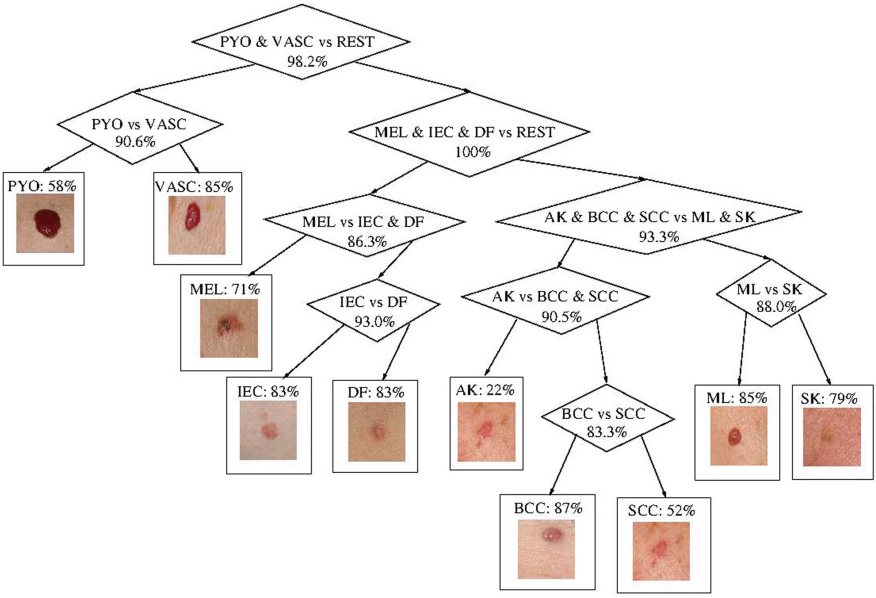

Fig. 1. Decision tree for lesion classification, also giving example images of the 10 lesion types. The numbers given at the decision boxes are the test data accuracies over the relevant classes (*i.e.* ignoring other classes that also go down that tree path). The numbers at the leaves of the tree are the final test accuracies over the whole dataset.

The confusion matrix for the final classification is shown in Table 4. There are three significant observations: (1) The AK lesion class, which has the worst performance, is mainly confused with the BCC and SCC classes. Visual inspection of the AK lesions shows that many of the lesions look a little like BCC and SCC lesions. (2) Many of the misclassifications are between the ML and SK classes. Neither of these are cancerous, so confusion between the classes has no real consequences. (3) Many of the other misclassifications are between the BCC and SCC classes. Both are cancers, but SCC needs more urgent treatment.

Merging BCC/SCC and ML/SK into 2 classes (and using the same tree) improves the classification rate to 0.871 (*i.e.* over an 8 class decision problem.).

To compare with the third experiment of Esteva *et al.* [5] (keratinocyte carcinomas (classes BCC and SCC here) *versus* benign seborrheic keratoses (class SK here)), we explored a KNN classifier with features selected from the same pool of features. Best accuracy of 0.906 was achieved with K = 9 and top features 168 and 186 over 5-fold cross-validation using the set of 327 BCC + SCC and 257 SK lesions. This is comparable to the accuracy (0.92) estimated from Esteva *et al.*'s sensitivity/specificity curves. To compare with the fourth experiment of Esteva *et al.* [5] (malignant melanoma (class (MEL) *versus* benign nevi (class ML)), we again explored a KNN classifier with features selected from the same pool of features. Accuracy of 0.995 (compared to their 0.92) was achieved with K = 9 and

Table 3. 10 class performance comparison with previous research.

Paper	Lesion (Micro) Accuracy	Class (Macro) Accuracy
Ballerini *et al.* [2]	0.743^{*}	0.592
Di Leo *et al.* [4]	0.67	-
Esteva *et al.* [5]	$>0.554^{+}$	-
Kawahara *et al.* [9]	0.818^{x}	-
Kawahara *et al.* [10]	0.795	-
New algorithm	0.781	0.705
New algorithm (BCC/SCC & ML/SK merged)	**0.871**	**0.736**

∗: The results of Ballerini *et al.* consider only 5 classes (AK, BCC, ML, SCC, SK). +: The results of Esteva *et al.* cover 9 classes, 5 of which roughly correspond to those considered in this paper. x: 0.818 was reported in [9] but 0.795 was reported and compared to in the later publication [10].

Table 4. Confusion Matrix: Row label is true, Column label is classification. Micro-average = 0.781. Macro-average accuracy = 0.705.

	AK	BC	ML	SC	SK	ME	DF	VA	PY	IE	TOT	ACC
AK	10	20	1	11	3	0	0	0	0	0	45	0.22
BCC	2	208	7	17	5	0	0	0	0	0	239	0.87
ML	2	10	280	0	39	0	0	0	0	0	331	0.85
SCC	0	34	0	46	8	0	0	0	0	0	88	0.52
SK	2	21	27	5	202	0	0	0	0	0	257	0.79
MEL	0	0	0	0	0	54	7	2	0	13	76	0.71
DF	0	0	0	0	0	4	54	3	0	4	65	0.83
VASC	0	0	0	0	0	6	1	82	3	5	97	0.85
PYO	0	0	0	0	0	2	0	7	14	1	24	0.58
IEC	0	0	0	0	0	5	5	1	2	65	78	0.83

only 2 features (2303 and 1824) over 5-fold cross-validation using the set of 331 ML and 76 MEL lesions. This compares to an accuracy of 0.92 estimated from Esteva *et al.*'s sensitivy/specificity curves [5]. Similarly, the clinically important discrimination between malignant melanoma (MEL) *versus* seborrheic keratosis (SK) (256 SK + 76 MEL samples) in this dataset gave perfect 5-fold cross-validated classification using K = 3 and 1 feature (2433). This high performance suggests that the dataset should be enlarged; however, melanoma is actually a rather uncommon cancer compared to, for example, BCC.

6 Deep Net Classifier Methodology

Given the general success of deepnets in classification image tasks, we investigated [3] three variations of a classifier based on the Resnet-50 architecture [8] pretrained on the ImageNet dataset and then tuned on the skin lesion samples in the same 5-fold cross-validation manner. The variations were:

1. A standard deepnet with 10 output classes, where the class with the highest activation level is selected.
2. A hierarchy of classifiers with the same structure as the decision tree presented above, except where each classification node is replaced with a Resnet-50 classifier.
3. A standard deepnet with 10 output classes, with a refinement stage. If the top two activation levels for a lesion from the standard deepnet were either {BCC, SCC} or {ML, SK}, then the lesion went to an additional binary Resnet-50 trained to discriminate the two classes.

Preprocessing of the segmented images: (1) produced standard 224 * 224 images, and (2) rescaled the RGB values of the whole image to give the background normal skin a standard value (computed by mean across the whole dataset). Data augmentation was by flipping, translation and rotation as there was no preferred orientation in the lesion images. Further augmentations such as color transformations, cropping and affine deformations as proposed by [15] could be added, which improved their melanoma classification accuracy by about 1% on the ISIC Challenge 2017 dataset. Training performance optimization used a grid search over the network hyperparameters. As deepnets are known to train differently even with the same data, the main result is an average over multiple (7) trainings. This is configuration (1) below.

Several other configurations were investigated using the same general deepnet: (2) The decision tree structure from Sect. 5, except each decision node is replaced by a 2 class deep net. (2x) For comparison, we list the performance of the decision tree from Sect. 5. (3) If the deepnet selected one of BCC, SCC, ML, or SK and the activation level was less than 0.88, then the lesion was re-classified using a 2 class BCC/SCC or ML/SK deep net. (4) A classifier for 8 classes, where the cancerous classes SCC/BCC were merged, and the benign classes ML/SK were merged. (4x) For comparison, the performance of the 8 class decision tree from Sect. 5. (5) A deepnet producing only a two-class cancer/not cancer decision (5x) A two-class cancer/not cancer decision using the same KNN (K = 17, 10 features) methods used in the decision tree. The resulting performances are shown in Table 5.

The results show that there is not much difference in cross-validated performance between the basic decision tree (0.781) and basic deepnet (0.787). Given the variability of deepnet training and cross-validation, there is probably no statistical significance between these. Interestingly, the reproduction of the decision tree with deepnets replacing the KNN classifiers produced distinctly worse performance (0.742 *vs* 0.781). It is unclear why. It is clear that applying the refinement based on easily confused classes gives better 10-class results

Table 5. Summary of deepnet results and comparisons with KNN classifier.

Case	Algorithm	Accuracy
1	Flat Resnet-50	0.787 ± 1.0
2	Decision tree with Resnet-50 nodes	0.742
2x	10 class decision tree from Sect. 5	0.781
3	Flat Resnet-50 with BCC/SCC and ML/SK refinement	0.801
4	8 class deep net with BCC/SCC and ML/SK merged	0.855 ± 0.4
4x	8 class decision tree from Sect. 5	0.871
5	Resnet-50 2 class cancer vs non-cancer	0.913 ± 0.71
5x	KNN based 2 class cancer vs non-cancer	0.874 ± 0.01

(0.801 *vs* 0.787), but still not as good as combining the BCC/SCC and ML/SK (0.855 *vs* 0.801). The 8-class deepnet performed worse than the 8-class decision tree (0.855 *vs* 0.871). Possibly the binary cancer/non-cancer classifier performed better when using the deepnet (0.913 *vs* 0.874).

7 Discussion

The approaches presented in this paper have achieved good performance on the 10 lesion type classification task and even better performance on the modified 8 class and binary problems. These lesions are those most commonly encountered in a clinical context, so there is a clear potential for medical use. Of interest is the fact that this was achieved using traditional features (and is thus more 'explainable') as well as using a deep learning algorithm.

Although the demonstrated performance is good, and generally better than dermatologist performance [5], there are still limitations. In particular, the BCC *versus* SCC, and ML *versus* SK portions of the classification tree have the lowest performance. Another poorly performing decision is AK *versus* other cancers (although the performance rate looks good, the large imbalance between the classes masks the poor AK identification - a class with few samples). We hypothesize that part of the difficulties arise because of the small dataset size, particularly for classes AK, DF, IEC, and PYO. When the K-NN classifier is used, it is necessary to have enough samples to have a good set of neighbors. This can also affect the other classes, because some lesion types may have difference appearance subclasses (*e.g.* BCC).

Another complication related to the dataset size is the correctness of the ground-truth for the lesion classes. Although 2 clinical dermatologists and a pathologist concur on the lesion diagnosis for the 1300 lesions in the dataset, generating that consistent dataset also showed up many differences of opinion about the diagnoses. It is probable that there are some lesions that were incorrectly labeled by all three professionals (the real world is not tidy), although, since one of the three was a pathologist, the dataset is probably reasonably correctly labeled. Again, having additional correctly labeled samples would help overcome outlying mis-labeled samples, given the nearest neighbor classifier structure.

Another limitation arises from the fact that the images in this dataset are all acquired under carefully controlled lighting and capture conditions. By contrast, the images used by Esteva *et al.* [5] came from many sources, which is probably one of the reasons their performance is so much lower. In order to make the solution found here be practically usable, it must work with different cameras and under different lighting conditions. This is a direction for future research.

Acknowledgments. The research was part of the DERMOFIT project (homepages.inf.ed.ac.uk/rbf/DERMOFIT) funded originally by the Wellcome Trust (Grant No: 083928/Z/07/Z). Much of the ground-truthing, feature extraction and image analysis was done by colleagues B. Aldridge, L. Ballerini, C. Di Leo, and X. Li as reported previously.

References

1. American Cancer Society. Cancer Facts & Figures (2016)
2. Ballerini, L., Fisher, R.B., Aldridge, R.B., Rees, J.: A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Celebi, M.E., Schaefer, G. (eds.) Color Medical Image Analysis. Lecture Notes in Computer Vision and Biomechanics, vol. 6. Springer, Dordrecht (2013). https://doi.org/10.1007/978-94-007-5389-1_4
3. Bertrand, A.: Classification of skin lesions images using deep nets. Intern report, INP Grenoble (2018)
4. Di Leo, C., Bevilacqua, V., Ballerini, L., Fisher, R., Aldridge, B., Rees, J.: Hierarchical classification of ten skin lesion classes. In: Proceedings of Medical Image Analysis Workshop, Dundee (2015)
5. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115–118 (2017)
6. Ferris, L.K., et al.: Computer-aided classification of melanocytic lesions using dermoscopic images. J. Am. Acad. Dermatol. **73**(5), 769–776 (2015)
7. Haralick, R.M., Shanmungam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. B Cybern. **3**(6), 610–621 (1973)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of CVPR (2016)
9. Kawahara, J., BenTaieb, A., Hamarneh, G.: Deep features to classify skin lesions. In: Proceedings of IEEE 13th International Symposium on Biomedical Imaging (ISBI) (2016)
10. Kawahara, J., Hamarneh, G.: Multi-resolution-tract CNN with hybrid pretrained and skin-lesion trained layers. In: Wang, L., Adeli, E., Wang, Q., Shi, Y., Suk, H.-I. (eds.) MLMI 2016. LNCS, vol. 10019, pp. 164–171. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47157-0_20
11. Korotkov, K., Garcia, R.: Computerized analysis of pigmented skin lesions: a review. Artif. Intell. Med. **56**(2), 69–90 (2012)
12. Li, X., Aldridge, B., Ballerini, L., Fisher, R., Rees, J.: Depth data improves skin lesion segmentation. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 1100–1107. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04271-3_133

13. Maglogiannis, I., Doukas, C.N.: Overview of advanced computer vision systems for skin lesions characterization. IEEE Trans. Inf. Technol. Biomed. **13**(5), 721–733 (2009)
14. Masood, A., Al-Jumaily, A.A.: Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. Int. J. Biomed. Imaging **2013**, 22 (2013)
15. Perez, F., Vasconcelos, C., Avila, S., Valle, E.: Data augmentation for skin lesion analysis. In: Stoyanov, D., et al. (eds.) CARE/CLIP/OR 2.0/ISIC -2018. LNCS, vol. 11041, pp. 303–311. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01201-4_33