

Automatic Scoring of Erythema and Scaling Severity in Psoriasis Diagnosis

Juan Lu^{1,*}, Ed Kazmierczak¹, Jonathan H. Manton², and Rodney Sinclair³

¹ Department of Computing and Information Systems, University of Melbourne

² Department of Electrical and Electronic Engineering, University of Melbourne

³ Department of Medicine(Dermatology), University of Melbourne,
and St. Vincent's Hospital Melbourne

jualu@student.unimelb.edu.au,

{edmundak, jmanton, rodneyds}@unimelb.edu.au

Abstract. Psoriasis is a common skin disease with no known cure. It is both subjective and time consuming to evaluate the severity of psoriasis lesions using manual methods. More objective automated methods are in great demand in both psoriasis research and in clinical practice. This paper presents an algorithm for scoring the severity of psoriasis lesions from 2D digital skin images. The algorithm uses the redness of the inflamed skin, or erythema, and the relative area and roughness of the flaky scaled skin, or scaling, in lesions to score lesion severity. The algorithm is validated by comparing the severity scores given by the algorithm against those given by dermatologists and against other automated severity scoring techniques.

Keywords: Skin image analysis, Severity modelling, Classification, Computer-aided diagnosis, Psoriasis.

1 Introduction

Psoriasis is a chronic skin disease with no known cure. It manifests as lesions consisting of red inflamed and itchy skin (*erythema*) and scaly flaky skin (*scaling*). There are an estimated 125 million people worldwide suffering this disease. In Australia, the percentage of affected people is between 2%-5% [1]. As there is no known cure, a great deal of effort has been expended on finding good treatments for psoriasis symptoms. However, no treatment is generally accepted and different physicians will treat the same symptoms differently. There is a need to compare various treatment methods objectively to determine which treatment is more effective in both psoriasis research and in clinical practice. Psoriasis severity scores are commonly used for comparing treatments [2].

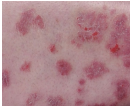
The severity score is a number that is used to classify the severity of psoriasis. A number of severity scores have been proposed in recent decades [3-5]. A widely used severity score is the *Psoriasis Area and Severity Index*, or PASI

* Corresponding author.

score [6]. The PASI score is calculated by dividing the body into a number of regions and grading the severity of the erythema (the red inflamed skin), and the severity of the scaling (the scaly, flaky skin typically found inside a lesion) within a region [6]. The severity of erythema and scaling in PASI scores are estimated visually often leading to significant inter- and intra- individual variation in scores. Further, PASI scores require that the symptoms of several lesions are estimated which greatly increases the workload of dermatologists. An objective and automatic scoring method will greatly help reduce variation in scoring and help to improve treatment research and clinical outcomes for patients as well as reducing the workload for clinicians.

Table 1 gives the different classes of PASI scores and some examples of lesions that would have the corresponding classification.

Table 1. PASI Erythema and Scaling Severity Intensity Scoring

Scores	Grade	Erythema scoring	Erythema images	Scaling scoring	Scaling im- ages
1	Mild	Light red		Fine scaling covering part of the lesion	
2	Moderate	Red, but not dark red		Fine to rough scaling covering a large part of the lesion	
3	Severe	Dark red		Rough, thick scaling covering a large part of the lesion	
4	Very severe	Very dark red		Very rough, very thick scaling totally covering the lesion	

The aim of this paper is to give a procedure for automatically estimating the severity of erythema and scaling using 2D digital skin images and the PASI severity scale for erythema and scaling.

Computer aided methods for psoriasis severity scoring have been under investigation for a number of decades [7]. In [8] the severity scores for erythema are correlated with the hue (H) value and saturation (S) value in the HSV colour model. The colour differences between psoriasis lesions and normal skin were investigated in [9] which concludes that the distribution of erythema severity is correlated with the difference in hue value.

However, in [8, 9] the colour value is sampled randomly ignoring the variation in lesion and skin colour to assess psoriasis severity. The colours of pigments in

lesions are used by [10] to derive mean colour values in RGB colours space and then these are used to grade the severity of lesions using K-Nearest Neighbours.

To the best of our knowledge the only work that attempts to grade the severity of scaling is given in [10], where the severity of scaling is derived by building a decision tree using the area of scaling inside the lesion to determine the lesion severity.

The algorithm presented in this paper differs by deriving a series of features that relate to the redness of erythema, the relative area of scaling and *roughness* of scaling, to evaluate the severity of erythema and scaling in lesions. The features used in this paper are based on a haemoglobin and melanin colour space [11] and the relative scaling area and texture [12]. Haemoglobin is related to the red colour in skin and melanin is related to the yellow and brown colour in skin. The algorithm and the severity features proposed are validated using multivariable analysis and classification evaluation.

2 A Method for Scoring the Severity of Symptoms: Erythema and Scaling in Psoriasis Lesions

2.1 Segmenting Erythema and Scaling in 2D Digital Skin Images

The severity scoring algorithm depends on first being able to identify psoriasis lesions. A typical psoriasis lesion is composed of an area of red inflamed skin (erythema) that surrounds an area of raised, rough scaly skin. Psoriasis lesions can also manifest as just erythema or as patches of scaly skin that are only partially surrounded by erythema. In our previous work an algorithm for segmenting erythema from normal skin is given in [11] and for segmenting scaling from erythema and normal skin is given in [12].

In [11] there are two key steps in the segmentation of erythema: (1) the decomposition of skin colour into **melanin** and **haemoglobin** components; and (2) erythema classification. Firstly, two independent components are extracted in the log RGB colour space using Independent Component Analysis, which correlates well with the haemoglobin and melanin.

Assuming that the variation in skin colour in the image are caused by melanin and haemoglobin and that melanin and haemoglobin are mutually independent then skin colour can be expressed as a linear combination of the melanin and haemoglobin components. A simple model of skin colour in the RGB colour space in terms of melanin and haemoglobin components is given by:

$$L_{x,y} = c^m q_{x,y}^m + c^h q_{x,y}^h + \Delta \quad (1)$$

where c^m and c^h are the melanin and haemoglobin basis vectors in the RGB colour space, $q_{x,y}^m$ and $q_{x,y}^h$ are the quantities of each pigment for the skin colour $L_{x,y}$ at coordinate (x, y) , and Δ is a constant vector that depends on other skin pigments and skin structure. Secondly, a Support Vector Machine (SVM) is then used to separate erythema pixels from normal skin using the erythema features $q_{x,y}^m$ and $q_{x,y}^h$.

Scaling is segmented from normal skin and erythema using a classification algorithm composed of the following steps:

Step 1. A **scaling colour contrast filter** is constructed using the L^* component and the a^* component in the $L^*a^*b^*$ colour space. The filter heightens the contrast between the whiter scaling pixels and the surrounding red erythema pixels.

The filter is a good differentiator of scaling from erythema but does not always give enough contrast to differentiate scaling from normal skin, especially under bright lights or if the skin is pale like scaly skin in a lesion.

Step 2. Scaling, which is rough, can be differentiated from normal skin, which is smooth, based on texture. Gabor filters have proved a useful tool for estimating the degree of 'roughness' and so are used here.

The second step is to derive a set of Gabor responses for each pixel in the image using a bank of Gabor filters tuned for different directions and spatial frequencies. The resulting **Gabor feature** in the image can be displayed using a gray-scale value that captures the degree of 'roughness' at the pixel.

Step 3. The Gabor features together with the colour contrast features obtained from the scaling colour contrast filter are then used to segment scaling from normal skin using a SVM smoothed by a Markov Random Field (MRF), which properly classifies any pixels that are misclassified by the SVM.

Figure 1 shows the segmentation of erythema and scaling in a psoriasis lesion.

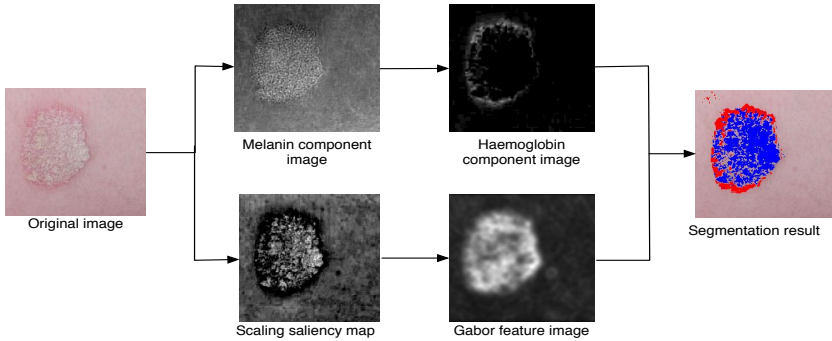


Fig. 1. Segmentation of erythema and scaling: segmented erythema is marked in red, and segmented scaling is marked in blue

2.2 Determining the Severity of the Psoriasis Lesions Using Erythema and Scalingness

In this paper we view erythema and scaling scoring as a multiclass classification problem. The erythema and scaling are can be simply referred as symptom type I and symptom type II respectively.

The input is a 2D digital image of a lesion described by a set F with a series of features depending on the symptom type that is being scored. Ignoring the

zero score due to its non-ambiguity in the clinical scoring, the output will be the set $C = \{1 := \text{Mild}, 2 := \text{Moderate}, 3 := \text{Severe}, 4 := \text{Very severe}\}$. The classification function P is built, such that $P : F \rightarrow C$.

The rule of PASI guidelines for scoring the symptom type I erythema is based on the colour of the lesion (as shown in Table 1). The deeper the red colour, the higher the severity score.

The haemoglobin and melanin components are the features used for scoring erythema. Note that dermatologists score erythema by comparing with the surrounding normal skin and so we employ the relative haemoglobin and relative melanin features, $\nabla \bar{q}^h$ and $\nabla \bar{q}^m$, in our classification algorithm. The erythema severity feature set is given by $F = \{\nabla \bar{q}^h, \nabla \bar{q}^m\}$, where $\nabla \bar{q}^h$ is the mean difference of haemoglobin values between erythema and normal skin, and $\nabla \bar{q}^m$ is the mean difference of melanin values between erythema and the normal skin in the lesion.

Scoring of erythema is done by a KNN for its simplicity, where $K = 5$ has been empirically determined. The severity of erythema is decided by the majority training samples in the K nearest neighbourhoods.

The PASI scoring of symptom type II: scaling is also shown in Table 1. The severity scoring rule is based on the roughness of the scaling and the area of scaling relative to the whole lesion.

The features used to differentiate scaling from erythema and normal skin are thus $F = \{\bar{g}, r\}$, where $\bar{g}$ is the mean value of Gabor features describing the degree of roughness and r is the ratio of scaling area to the whole lesion area.

The performance of the C4.5 decision tree used in our algorithm is better than a range of other classification methods which was determined empirically for the scaling severity classification problem. The decision tree is split based on a normalised information gain criterion, followed by post-pruning to avoid the over-fitting [13].

3 Validating the Method

3.1 The Experimental Design

Samples of psoriasis skin images are collected from the Skin & Cancer Foundation Victoria, where the imaging environment is carefully set to ensure controlled illumination. The images used in the comparison were those that were given the same PASI scores by two dermatologists.

The proposed algorithm is validated based linear correlation and classification performance. The linear correlation between the features used in our method with the severity scores given by dermatologists is analysed first. Then the classification accuracy of our scoring method is compared with other scoring systems.

In the linear correlation stage, a t-test is used to measure the correlation based on the hypothesis that the observation of one class is at least as extreme as another class. The resulting p-values and F-values are shown. Moreover, correlation between severity scores and severity features is analysed using correlation coefficients. The analysis shows a linear relationship using a normalised covariance.

In the classification performance evaluation stage, classification accuracy is calculated as the percentage of correctly classified samples. In our experiment, the classification accuracy for each severity degree is evaluated by a 10-fold cross validation. Assuming the accuracy difference between two classifiers is a t-distribution, by setting the confidence level to be 0.95, the confidence interval of true accuracy difference is analysed. If zero is not in the range of the confidence interval, the performance of two classifiers is significantly different.

3.2 Symptom Type I: Erythema Severity Scoring

There are 88 images of erythema lesions spanning a number of different skin types and ethnicities, where 10 samples score 1, 31 samples score 2, 37 samples score 3, and 10 samples score 4.

Linear Correlation Analysis. The result of linear correlation between erythema severity and colour features is shown in Table 2. Table 2 also shows relative hue component $\nabla \overline{H^{ab}}$ used in [9], and relative spectral components $\nabla \overline{R}$, $\nabla \overline{G}$, $\nabla \overline{B}$ from the RGB colour space proposed in [10]. The relative colour component features are defined to be the difference of mean values between erythema and normal skin. Specifically, the hue component in [9] is defined as the arctangent of ratio of a^* component to b^* component in the $L^*a^*b^*$ colour space.

The distribution of the melanin and haemoglobin components with changes to erythema severity intensity is presented with box plots in Figure 2. Notice that though mean values of the melanin component, haemoglobin component and their summation are linearly related with the severity intensity in general, it is hard to distinguish erythema severity by setting thresholding values alone.

Scoring. The result of our erythema scoring method is compared with the results obtained by the Minimum Centre Distance (MCD), where a test sample is grouped into a class in which the mean value of the training samples is closest to the test sample in a feature space [14].

Table 2. Analysis Of Linear Correlation Between Erythema Severity And Colour Features

Features	p-value	F-value	Correlation coefficient
$\nabla \overline{H^{ab}}$	0.299	1.24	0.048
$\nabla \overline{R}$	4.52E-06	10.79	-0.448
$\nabla \overline{G}$	2.18E-10	21.4	-0.63
$\nabla \overline{B}$	6.19E-07	12.72	-0.538
$\nabla \overline{R} + \nabla \overline{G} + \nabla \overline{B}$	2.10E-08	16.22	-0.569
$\nabla \overline{q^m}$	0.048	2.74	0.194
$\nabla \overline{q^h}$	1.44E-09	19.19	0.616
$\nabla \overline{q^h} + \nabla \overline{q^m}$	6.58E-06	10.439	0.456

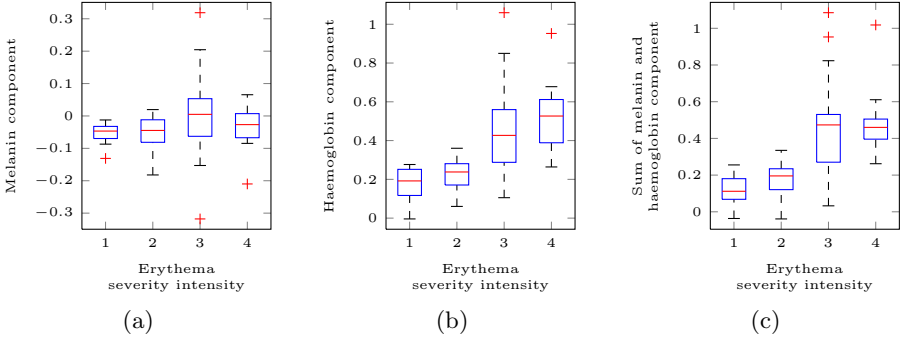


Fig. 2. Box plots of haemoglobin component and melanin component distributions with changes of erythema severity. (a) Box plot of melanin component distribution with changes of erythema severity intensity. (b) Box plot of haemoglobin component distribution with changes of erythema severity intensity. (c) Box plot of summation of haemoglobin component and melanin component distribution with changes of erythema severity intensity.

Figure 3 illustrates the classification accuracy of KNN and MCD clustering samples using features of relative colour component $\nabla\overline{R}, \nabla\overline{G}$ and $\nabla\overline{B}$ in the RGB colour space and the features of relative haemoglobin and melanin component, ∇q^h and ∇q^m . KNN clustering method using the relative haemoglobin and melanin components shows the highest accuracy compared with other methods. The mean accuracy in the 10-fold cross validation is 78.85%. Moreover, the performance of KNN is better than MCD in general. For erythema, the accuracy of severity 1 and severity 4 are better than the accuracy for severity 2 and severity 3. This is due to the ambiguity in scoring severity 2 and severity 3 lesions.

Assuming that the accuracy differences in the k-fold cross validation is a t-distribution, the statistical significance of the accuracy difference between the top classifier using KNN with the features used in our algorithm and other severity classifiers are shown in Table 3. The indications are that performance of the top classifier is significantly better than the rest.

Table 3. Accuracy difference between the classifier using KNN in space spanned by haemoglobin component and melanin component and other implemented classifiers

Compared classifier	Colour space	Accuracy difference
MCD	RGB colour space	[-0.097, -0.092]
MCD	Haemoglobin and melanin component	[-0.1, -0.096]
KNN	RGB colour space	[-0.052, -0.05]

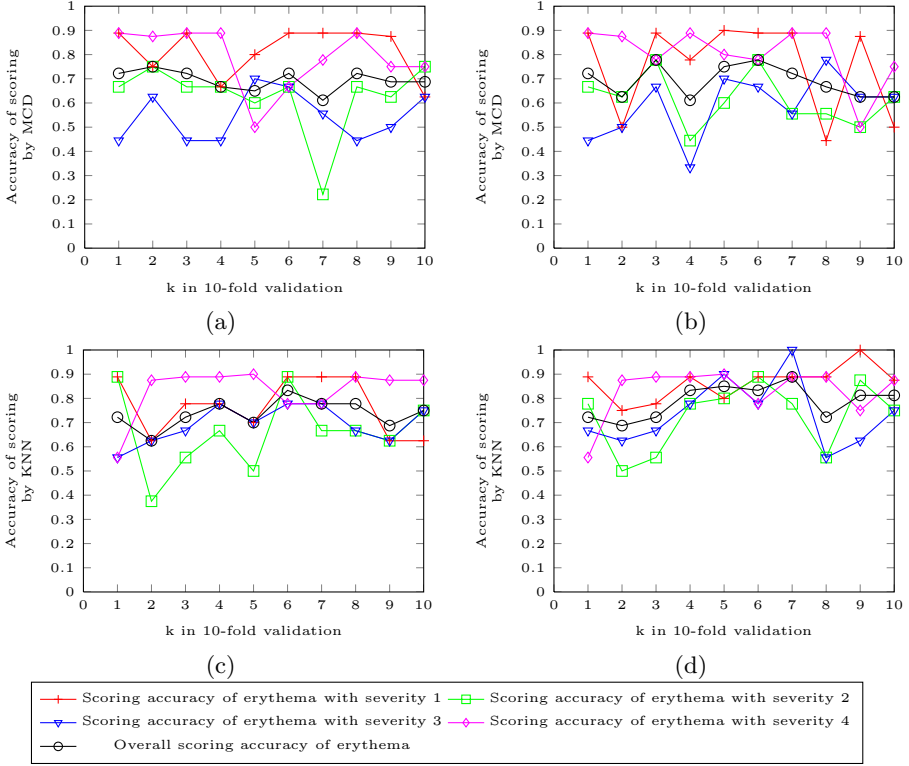


Fig. 3. Accuracy of erythema scoring with 10-fold cross validation for a series of severities and the corresponding classifier. (a) Accuracy of erythema scoring in the RGB colour space using MCD. (b) Accuracy of erythema scoring in the space spanned by the haemoglobin component and melanin component using MCD. (c) Accuracy of erythema scoring in the RGB colour space using KNN. (d) Accuracy of erythema scoring in the space spanned by the haemoglobin component and melanin component using KNN.

3.3 Symptom Type II: Scaling Severity Scoring

In the scaling severity scoring part, 52 images of psoriatic lesion are collected. They are composed of 10 images with scaling severity 1, 17 images with scaling severity 2, 18 images with scaling severity 3, and 7 images with scaling severity 4.

Linear Correlation Analysis. Linear correlation analysis is shown in Table 4. We examined the feature scaling area s , which is considered as a major element related with scaling severity in [10], and the feature set proposed in this paper.

Observe that even though scaling area has a higher correlation coefficient than relative scaling area, it does not possess priority in p-values and F-values, especially when compared to the roughness degree.

Figure 4 shows the distribution of relative scaling area, scaling roughness degree and their summation with changes of scaling severity intensity. We can see

Table 4. Analysis of Linear Correlation Between Scaling Severity And Its Features

Features	p-value	F-value	Correlation coefficient
s	0.023	3.44	0.316
r	0.019	3.61	0.18
$\bar{g}$	8.67E-09	19.81	0.641
$r+\bar{g}$	4.59E-07	14.55	0.58

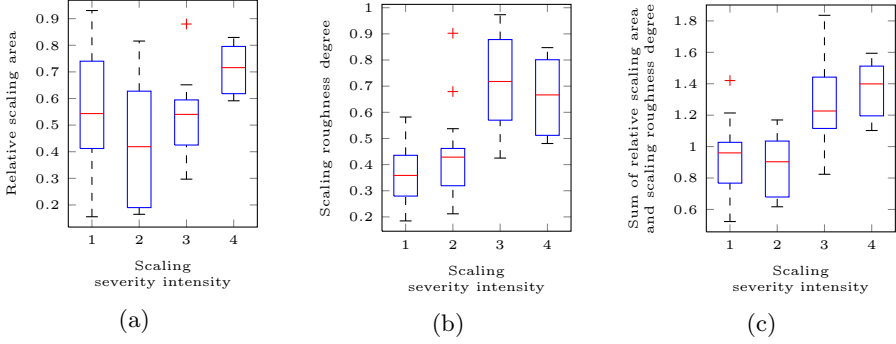


Fig. 4. Box plots of relative scaling area and scaling roughness degree distribution with changes of scaling severity intensity. (a) Box plot of relative scaling area distribution with changes of scaling severity intensity. (b) Box plot of scaling roughness degree distribution with changes of scaling severity intensity. (c) Box plot of summation of relative scaling area and scaling roughness degree distribution with changes of scaling severity intensity.

that ranges of the two features and their summation overlap for different scaling severity intensities. No clear linear relationship exists between the features and the severity intensities.

Scoring. The performance of our scaling scoring method is compared with KNN and the method in [10], where a decision tree is employed to score the severity degree with the scaling area. The result is shown in Figure 5. For KNN and the decision tree, scoring accuracy for severity 1 and severity 4 is generally better than scoring accuracy for severity 2 and severity 3 as with the symptom type I: erythema. Moreover, using the feature set of our method the accuracy is much better than features proposed in [10]. Using features in [10], KNN and the decision tree achieve accuracy of 79.58% and 78.58% respectively, while using our proposed features, an accuracy of 86.75% for KNN and 88.67% for the decision tree are achieved on the samples.

In Table 5, the statistical significance of the accuracy difference between the classifier using the decision tree with the relative scaling area and the scaling roughness degree and other implemented scaling scoring classifiers is shown. All of the difference ranges do not include zero. Thus the accuracy of the decision tree we proposed is significantly different when compared with the other classifiers.

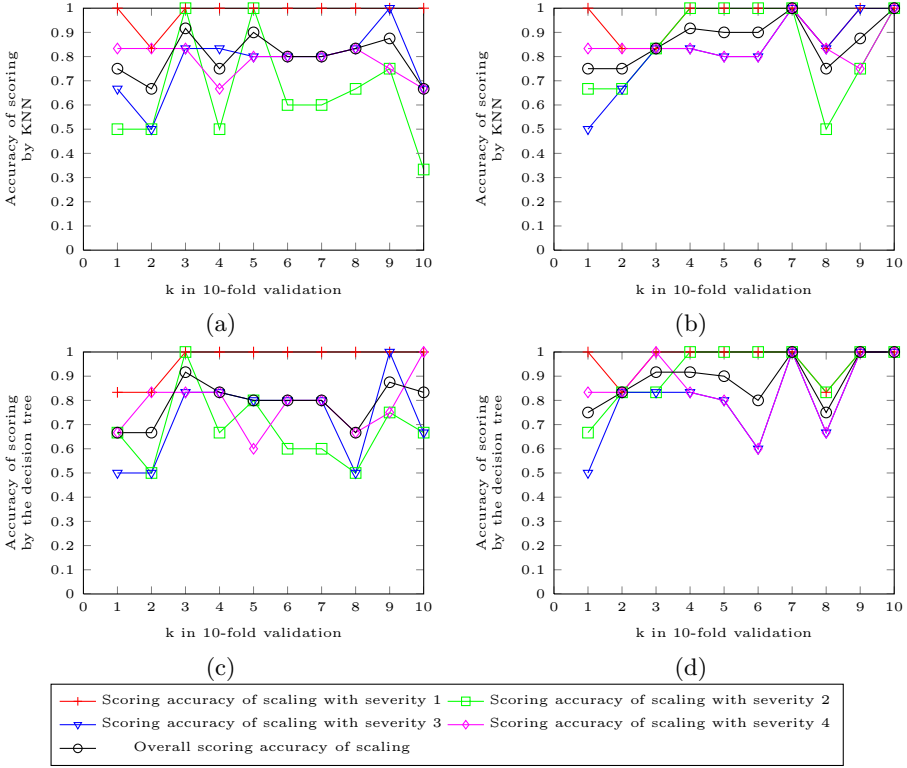


Fig. 5. Accuracy of scaling scoring with 10-fold cross validation for a series of severities and the corresponding classifier. (a) Accuracy of scaling scoring using KNN with the feature proposed in [10]. (b) Accuracy of scaling scoring using KNN with the features of relative scaling area and scaling roughness degree. (c) Accuracy of scaling scoring using the decision tree with the feature proposed in [10]. (d) Accuracy of scaling scoring using the decision tree with the features of relative scaling area and scaling roughness degree.

Table 5. Accuracy difference between the classifier using the decision tree with relative scaling area and scaling roughness degree and other implemented scaling scoring classifiers

Compared classifier	Scaling features	Accuracy difference
KNN	Features in [10]	[0.087, 0.095]
KNN	Relative scaling area and scaling roughness degree	[0.081, 0.02]
Decision tree	Features in [10]	[0.1, 0.102]

4 Discussion

Comparing the erythema scoring method proposed here with [9] shows that our method performs better on the sample. The statistics in Table 2 indicates that our method performs better on the sample, while Figure 3 indicates that scoring performance using the melanin and haemoglobin components is better on the sample than the RGB colours, as suggested in [10]. This can be due to the fact that melanin is distributed in normal skin and is factoring in erythema intensity.

Our scaling severity scoring method outperformed the method in [10], where only scaling area is used. Table 4 indicates that the roughness degree in our model has significantly better statistical correlation values.

Both erythema scoring and scaling scoring give good classification accuracy for the lowest and highest severity, while the classification accuracy for medium severity is a little lower. The severity features derived from segmentation have the obvious advantage in distinguishing symptoms from normal skin, but to further classify the symptom severity, more features closely related to the severity intensities will be explored in future.

Severity 1 erythema may be improperly segmented if the training sets are not chosen carefully. Selecting training sets from individual images will avoid this problem. When it comes to the scaling segmentation, mis-segmentation happened to scaling with severity 1 as well. Shallow redness and shadows are major disturbances in this situation. Determining severity intensity scores in the presence of such disturbance is a topic for future work.

5 Conclusion

In this paper, we proposed automatic scoring methods for erythema and scaling in 2D psoriasis images. Our algorithm is shown to perform better when compared with previous severity models based on the concurrent scores from two dermatologists for a range of different images. In the future, we will continue to focus on the analysis of psoriasis images with the view to determining a wider range of severity scores. More comprehensive models and learning methods will be proposed to achieve this.

Acknowledgments. The authors would like thank to Skin & Cancer Foundation Victoria, Australia and St. Vincent's Hospital Melbourne for support of the research.

References

1. Jenner, N., Campbell, J., Plunkett, A., Marks, R.: Cost of psoriasis: A study on the morbidity and financial effects of having psoriasis in australia. *Australasian Journal of Dermatology* 43(4), 255–261 (2002)
2. Kreft, S., Kreft, M., Resman, A., Marko, P., Kreft, K.Z.: Computer-aided measurement of psoriatic lesion area in a multicenter clinical trial—comparison to physician's estimations. *Journal of Dermatological Science* 44(1), 21–27 (2006)

3. Department of Health and Aging Therapeutic Goods Administration. Australian Government: Guideline on clinical investigation of medicinal products indicated for the treatment of psoriasis (2004)
4. Langley, R.G., Ellis, C.N.: Evaluating psoriasis with psoriasis area and severity index, psoriasis global assessment, and lattice system physician's global assessment. *Journal of the American Academy of Dermatology* 51(4), 563–569 (2004)
5. Puzenat, E., Bronsard, V., Prey, S., Gourraud, P., Aractingi, S., Bagot, M., Cribier, B., Joly, P., Jullien, D., Le Maitre, M., Paul, C., Richard-Lallemant, M., Ortonne, J., Aubin, F.: What are the best outcome measures for assessing plaque psoriasis severity? a systematic review of the literature. *Journal of the European Academy of Dermatology and Venereology* 2, 10–16 (2010)
6. Fredriksson, T., Pettersson, U.: Severe psoriasis-oral therapy with a new retinoid. *Dermatologica* 157(4), 238–244 (1978)
7. Savolainen, L., Kontinen, J., Röning, J., Oikarinen, A.: Application of machine vision to assess involved surface in patients with psoriasis. *British Journal of Dermatology* 137(3), 395–400 (1997)
8. Ahmad Fadzil, M.H., Ihtatho, D.: Modeling psoriasis lesion colour for PASI erythema scoring. In: *International Symposium on Information Technology*, vol. 2, pp. 1–6 (2008)
9. Ahmad Fadzil, M.H., Ihtatho, D., Affandi, A.M., Hussein, S.: Objective assessment of psoriasis erythema for PASI scoring. In: *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4070–4073 (2008)
10. Delgado, D., Ersboll, B., Carstensen, J.M.: An image based system to automatically and objectively score the degree of redness and scaling in psoriasis lesions. In: *13th Danish Conference on Image Analysis and Pattern Recognition*, pp. 130–137 (2004)
11. Lu, J., Manton, J., Kazmierczak, E., Sinclair, R.: Erythema detection in digital skin images. In: *17th IEEE International Conference on Image Processing*, pp. 2545–2548 (2010)
12. Lu, J., Kazmierczak, E., Manton, J., Sinclair, R.: Unsupervised segmentation of scaling in 2d psoriasis skin images (2012), <http://repository.unimelb.edu.au/10187/15715>
13. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)
14. Masmoudi, H., Hewitt, S., Petrick, N., Myers, K., Gavrielides, M.: Automated quantitative assessment of her-2/neu immunohistochemical expression in breast cancer. *IEEE Transactions on Medical Imaging* 28(6), 916–925 (2009)