# Paper III:
Divergence-based colour features for melanoma detection

# DIVERGENCE-BASED COLOUR FEATURES FOR MELANOMA DETECTION

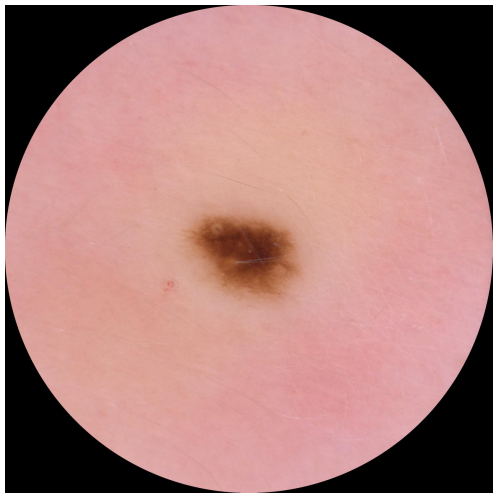
Kajsa Møllersen*, Jon Yngve Hardeberg† and Fred Godtliebsen‡
*Norwegian Centre for Integrated Care and Telemedicine
University Hospital of North Norway, 9038 Tromsø, Norway
Email: kajsa.mollersen@telemed.no
†Faculty of Computer Science and Media Technology
Gjøvik University College, 2815 Gjøvik, Norway
‡Department of Mathematics and Statistics
UiT The Arctic University of Norway, 9037 Tromsø, Norway

***Abstract*—Melanoma is a deadly form of skin cancer which is difficult to detect in its early stages. Several computer-aided diagnostic systems based on dermoscopic images of skin lesions intend to improve melanoma detection. Colour is an important factor in correctly classifying a skin lesion. Here, we introduce divergence-based colour features, using the Kullback-Leibler information as a preferred divergence function. These features are based on the divergence between the distribution of the pixel values of a lesion image, and that of the pixel values of either a benign or a malignant model. The features' sensitivities and specificities are reported, along with the contribution to an existing classifier for skin lesions. The features improve the performance of the existing classifier and are therefore relevant for melanoma detection.**

Keywords: *Melanoma, Colour feature, Divergence, Kullback-Leibler, Gaussian mixture distribution*

## I. Introduction

Melanoma is the deadliest form of all skin cancers [1]. Early detection is crucial, since the survival rate of the patient drops rapidly as the tumour evolves [1]. A dermoscope (magnifying lens, surrounding light and glass plate) can ease the detection of melanoma since it lets the light penetrate the uppermost skin layer and by doing that provides more information about the lesion [2]. Fig. 1 shows examples of dermoscopic images. Several rules can help doctors interpret what they see through the dermoscope, e.g. the ABCD rule of dermoscopy, the 7-point checklist and more [3]. These rules all have in common that colour is a major feature [3]. Many computer-aided diagnostic (CAD) systems for melanoma detection based on dermoscopic images exist [4]. These systems follow the same procedure of image pre-processing (noise reduction, downsampling, etc.), segmenting the lesion from the skin, feature value calculation, feature selection and classification. A number of feature algorithms has been introduced, many of them concerning colour [4]. The colour features can roughly be divided into two categories; specific colours (often light brown, dark brown, red, black, blue and white, which are the colours in the ABCD rule), and statistical value of the colours



(a) Benign lesion

(b) Malignant lesion (melanoma)

Fig. 1: Dermoscopic images of skin lesions.

2015 Colour and Visual Computing Symposium (CVCS) 978-1-4799-1765-5/15/$31.00 2015 IEEE

in a lesion, typically the estimated moments (mean, standard deviation, skewness) for each colour channel, as well as entropy and energy [4]. Among the more sophisticated colour features, Celebi and Zornberg [5] proposed a feature based on $k$-means clustering and symbolic regression. Seidenari *et al.* [6] calculated specific colours, estimating the number of colours based on a training set.

Here, we introduce a new type of statistical colour feature whose value reflects the divergence between the distribution of the pixel values of a skin lesion image and the distribution of a benign model or a malignant model. Gaussian mixture models (GMM), also referred to as Gaussian mixture distribution, are used to estimate the distributions. The Kullback-Leibler information in combination with importance sampling form the basis of the new colour features. The paper is organised as follows. Section II gives the necessary technical background. Section III introduces the new features. Section IV presents the data set and the specific method used. Section V gives the results. Section VI discusses the findings.

## II. Model fitting, divergences, importance sampling and cross-validation

### A. Model fitting

Any continuous distribution, $f(\mathbf{x})$, can be approximated with arbitrary accuracy by a GMM [7]:

$$f(\mathbf{x}) \approx f_K(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k), \tag{1}$$

where $\pi_k > 0$, $\sum_{k=1}^{K} \pi_k = 1$. When the distribution $f(\mathbf{x})$ is unknown, the model $f_K(\mathbf{x})$ is fitted to observed values, and the accuracy then depends on the quality of the observations. Fitting is commonly done by the Expectation-Maximization (EM) algorithm [8], which requires a pre-set number of components, $K$. The estimation of $K$ is usually done by fitting several models with $K_{min}, \ldots, K_{max}$ many components, and then using a criterion that balances good fit and complexity to select the best model. The Bayesian information criterion (BIC) [9] is well known and widely used for this purpose [10].

### B. Divergence between distributions

The divergence of two distributions, $p_s(\mathbf{x})$ and $p_t(\mathbf{x})$, is a measure of how different the two distributions are. A number of divergence functions exists, and the choice of divergence function must be made according to criteria that are relevant for the problem at hand. Many well-known divergence functions are symmetric, e.g. Variational distance, Hellinger/Kolmogorov distance, Kullback-Leibler/Jeffrey divergence, Chernoff distance, Bhattacharyya distance, Matusita distance [11]–[14], and Shannon-entropy-based divergences [15], [16]. An example is the Jensen-Shannon divergence [17]

$$\begin{aligned} JS(p_s, p_t) =& \pi_s \int_{\mathcal{X}} p_s(\mathbf{x}) \log \frac{p_s(\mathbf{x})}{\pi_s p_s(\mathbf{x}) + \pi_t p_t(\mathbf{x})} d\mathbf{x} \\ &+ \pi_t \int_{\mathcal{X}} p_t(\mathbf{x}) \log \frac{p_t(\mathbf{x})}{\pi_s p_s(\mathbf{x}) + \pi_t p_t(\mathbf{x})} d\mathbf{x}, \end{aligned} \tag{2}$$

where $\pi_s, \pi_t \geq 0$, $\pi_s + \pi_t = 1$ are the weights of $p_s$ and $p_t$. A well-known non-symmetric divergence is the Kullback-Leibler information [18]

$$I(p_s, p_t) = \int_{\mathcal{X}} p_s(\mathbf{x}) \log \frac{p_s(\mathbf{x})}{p_t(\mathbf{x})} d\mathbf{x}. \tag{3}$$

### C. Importance sampling

The integrals in Eq. 2 and Eq. 3 do not have analytical solutions when $p_s(\mathbf{x})$ or $p_t(\mathbf{x})$ are GMMs, and a numerical approximation is needed. Using points on a regular grid is time consuming and inefficient for higher dimensions and/or when the integrand has low value for a large subspace. In Monte Carlo integration [19, p. 83], the points are randomly sampled. In importance sampling [19, p. 90], the points are sampled from a distribution, preferably with high density for subspaces with large contributions to the integral, and then weighted by the density of the distribution in that point:

$$\int_{\mathcal{X}} h(\mathbf{x}) d\mathbf{x} \approx \frac{1}{N} \sum_{i=1}^{N} h(\mathbf{x}_i)/g(\mathbf{x}_i), \tag{4}$$

where $N$ is the number of samples and $g(\mathbf{x})$ is the probability density function (pdf) from which the samples are drawn.

### D. Cross-validation

Ideally, observed data are divided into two separate sets: a training set and a test set. The training set is used to construct a model, which is tested on an independent test set, i.e. a data set that has not been used in any part of the model construction, including feature design, feature selection and classifier training. If the training set is too small compared to the model complexity, the model will be unstable in the sense that replacing a small fraction of the data in the training set with other data from the same distribution will lead to a different model. If the test set is too small, the observed performance of the model is unreliable. If training and test sets of sufficient size cannot be provided, cross-validation can be used. The data set is partitioned into independent training and test sets. In stratified $K$-fold cross-validation, each class in the data set is divided into $K$ folds of equal size. Fold number $1, \ldots, K$ is set aside as test set, while the rest of the data is used to construct the model. The procedure is repeated $K$ times, until all the data have been used in the test set. If parameter adjustment, feature selection or model selection is done, these steps have to be repeated for every new training set, see e.g. [20, p.245], [21], [22]. The choice of $K$ is not trivial, since it affects both the bias and the variance of the model's performance [23], [24]. Low $K$ gives negative bias, and both high and low $K$ give high variance. 5- and 10-fold cross-validation are commonly used.

## III. Divergence-based colour features

We propose a new type of feature for melanoma detection in dermoscopic images. The divergence-based colour features are defined as the divergence between the colour distribution in a lesion and that of a benign or a malignant model:

$$D(p_l, p_b) \qquad \text{and} \qquad D(p_l, p_m), \tag{5}$$

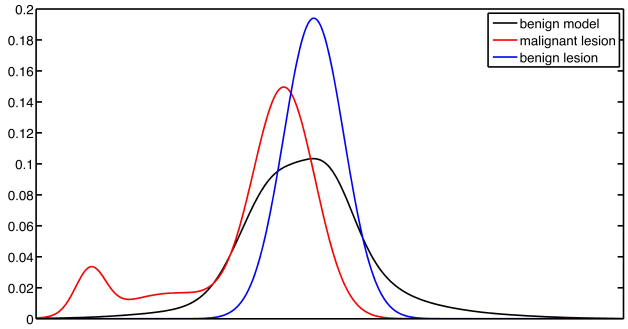



Fig. 2: The benign model distribution, $p_b(\mathbf{x})$ (black), has large dispersion. The lesion distribution $p_l(\mathbf{x}|\boldsymbol{\theta}_{bj})$ (blue) is enveloped by $p_b(\mathbf{x})$. The lesion distribution $p_l(\mathbf{x}|\boldsymbol{\theta}_{mj'})$ (red) only partially overlaps with $p_b(\mathbf{x})$. For the purpose of illustration, we have used only one dimension.

where $D$ is a divergence function, $p_l(\mathbf{x})$ is the pdf of the pixel values from a lesion image, and $p_b(\mathbf{x})$ and $p_m(\mathbf{x})$ are the pdfs of a benign and a malignant model, respectively.

The choice of divergence function has great impact on the features' values and should be made according to pre-defined assumptions. Assume that the pixel values of a lesion image are observations from the underlying distributions $p_l(\mathbf{x}|\boldsymbol{\theta}_{bj})$ and $p_l(\mathbf{x}|\boldsymbol{\theta}_{mj'})$, where $\boldsymbol{\theta}_{bj}$ is the parameter vector of benign lesion $j$ and $\boldsymbol{\theta}_{mj'}$ is the parameter vector of malignant lesion $j'$. Assume that they are continuous, and therefore can be approximated by GMMs. Thus, the parameter vectors $\boldsymbol{\theta}_{bj}$ and $\boldsymbol{\theta}_{mj'}$ are not estimated. Define the benign model distribution $p_b(\mathbf{x}) = \sum_{j=1}^{|ben|} w_j p_l(\mathbf{x}|\boldsymbol{\theta}_{bj})$, where $|ben|$ is the number of possible different benign lesion images, $w_j$ is the weight, $w_j > 0, \sum w_j = 1$. Assume that $p_l(\mathbf{x}|\boldsymbol{\theta}_{bj}) \to 0$ faster than $p_b(\mathbf{x}) \to 0$ for $\mathbf{x} \to \pm\infty$. This is illustrated in Fig. 2, where $p_l(\mathbf{x}|\boldsymbol{\theta}_{bj})$ (blue) drops to zero faster than $p_b(\mathbf{x})$ (black) for both $x \to -\infty$ and $x \to \infty$. In addition, assume that $p_l(\mathbf{x}|\boldsymbol{\theta}_{mj'}) \to 0$ slower than $p_b(\mathbf{x}) \to 0$ for some $\mathbf{x} \to \pm\infty$, as illustrated in Fig. 2, where $p_l(\mathbf{x}|\boldsymbol{\theta}_{mj'})$ (red) drops to zero slower than $p_b(\mathbf{x})$ (black) for $x \to -\infty$.

In other words, $p_b(\mathbf{x})$ envelops $p_l(\mathbf{x}|\boldsymbol{\theta}_{bj})$ due to $p_b(\mathbf{x})$'s large dispersion, whereas it only partly overlaps with $p_l(\mathbf{x}|\boldsymbol{\theta}_{mj'})$ due to the latter's shift in location. We define the distribution $p_m(\mathbf{x}) = \sum_{j'=1}^{|mal|} w_{j'} p_l(\mathbf{x}|\boldsymbol{\theta}_{mj'})$ and make equivalent assumptions as for $p_b(\mathbf{x})$. For ease of notation, we denote the lesion distributions $p_l(\mathbf{x})$, as we, in general, do not know if the lesion is benign or malignant.

A divergence function with high contribution for low values of $p_b(\mathbf{x})$ and high values of $p_l(\mathbf{x})/p_b(\mathbf{x})$, but low contribution for low values of $p_l(\mathbf{x})$ can differentiate between a benign and a malignant lesion. This can be achieved by fulfilling the criteria

$$p_b(\mathbf{x}) \to 0, \frac{p_l(\mathbf{x})}{p_b(\mathbf{x})} \to \infty \Rightarrow D_{\mathcal{X}_\infty}(p_l, p_b) \to \max(D_{\mathcal{X}_\infty}) \quad \text{(Max)}$$

$$p_l(\mathbf{x}) \to 0 \Rightarrow D_{\mathcal{X}_0}(p_l, p_b) \to \min(D_{\mathcal{X}_0}) \quad \text{(Min)}$$

where $\mathcal{X}_\infty$ is the subspace of $\mathcal{X}$ where $p_l(\mathbf{x})/p_b(\mathbf{x}) \to \infty$, and $\mathcal{X}_0$ is the subspace of $\mathcal{X}$ where $p_l(\mathbf{x}) \to 0$. The criteria cannot be fulfilled simultaneously by a symmetric divergence function.

The non-symmetric Kullback-Leibler information fulfils the two criteria (the calculations are straightforward)

$$Max : I_{\mathcal{X}_\infty}(p_l, p_b) \to \infty = \max(I_{\mathcal{X}_\infty}) \tag{6}$$

$$Min : I_{\mathcal{X}_0}(p_l, p_b) \to 0 = \min(I_{\mathcal{X}_0}) \tag{7}$$

For the symmetric Jensen-Shannon divergence

$$Max : JS_{\mathcal{X}_\infty}(p_l, p_b) \to -\pi_l \log \pi_l \int_{\mathcal{X}_\infty} p_l(\mathbf{x}) d\mathbf{x} \tag{8}$$

$$Min : JS_{\mathcal{X}_0}(p_l, p_b) \to -\pi_b \log \pi_b \int_{\mathcal{X}_0} p_b(\mathbf{x}) d\mathbf{x}, \tag{9}$$

since $p_l(\mathbf{x})/p_b(\mathbf{x}) \to \infty$ gives $\pi_l p_l(\mathbf{x}) + \pi_b p_b(\mathbf{x}) \to \pi_l p_l(\mathbf{x})$. The criteria cannot be fulfilled simultaneously by adjusting $\pi$ since $-\pi \log \pi$ increases for both $\pi \to 0$ and $\pi \to 1$.

In importance sampling, any $g(\mathbf{x})$ in Eq. 4 is asymptotically correct, as long as $\mathcal{X}_g \supseteq \mathcal{X}_h$, but in practice the choice of $g(\mathbf{x})$ has great influence on the result. The region of interest is where $p_b(\mathbf{x})$ has low values, $p_l(\mathbf{x}|\boldsymbol{\theta}_{mj'})$ has high values and $p_l(\mathbf{x}|\boldsymbol{\theta}_{bj})$ has low values, since this is where we can differentiate between a benign and a malignant lesion. In Fig. 2, the region of interest is at the left. By setting $g(\mathbf{x}) = p_b(\mathbf{x})$, samples from this region are heavily weighted. However, since the samples also are taken from $p_b(\mathbf{x})$, they will be sparse. We therefore propose to sample from $p_m(\mathbf{x})$, but weight by $p_b(\mathbf{x})$. We define the Kullback-Leibler-based colour feature as follows

$$d^b = I^*(p_l, p_b) = \frac{1}{N} \sum_{i=1}^{N} \frac{p_l(\mathbf{x}_i)}{p_b(\mathbf{x}_i)} \log \frac{p_l(\mathbf{x}_i)}{p_b(\mathbf{x}_i)}, \tag{10}$$

where the $\mathbf{x}_i$'s are sampled from $p_m(\mathbf{x})$. The feature $d^m$ is defined equivalently. The asterisk signals that this is not a direct approximation of the Kullback-Leibler information, since it is not a proper importance sampling. The proposed feature does not fulfil all the properties of the Kullback-Leibler information, but it fulfils the two criteria Max and Min.

## IV. Materials and Methods

The data set consisted of dermoscopic images of 752 benign lesions and 80 melanomas. The lesions were excised due to suspicion of malignancy, and the final diagnoses were made by histopathology. For further details on the diagnoses, see [25]. Each image was converted from raw to RGB and then to CIELAB, assuming sRGB. Automatic segmentation was performed [26], and the resulting mask defined the lesion. To reduce noise, the images were binned using a coordinate-wise median with $5 \times 5$ pixel non-overlapping windows. Note that we used binning, not filtering, which downsamples the image and preserves independency of the pixel values. Coordinate-wise median binning was also used in [25]. Then, 1000 lesion pixels were randomly selected from each binned image. Large lesions can be indicative of melanoma. The potential spurious

relationship between the feature value and diagnosis due to lesion size is avoided by sampling a fixed number of pixels. A slight improvement in performance was observed when increasing the number of pixels from 250 to 500 and to 1000.

To estimate the benign model distribution, $p_b(\mathbf{x})$, GMMs with $K = 15, \ldots, 45$ components were fitted for a random sample of 72 benign lesion images (the same number as for the malignant model distribution), and BIC was used for model selection. These images were then excluded for the sake of independence. Due to the low number of melanomas, 10-fold stratified cross-validation was applied to the remaining data. The training sets then consisted of 72 melanomas, and were used to fit $p_m(\mathbf{x})$ in the same manner as $p_b(\mathbf{x})$. To estimate the lesion distributions, $p_l(\mathbf{x})$, GMMs with $K = 1, \ldots 15$ were fitted for each lesion image, and BIC was used for model selection. The whole procedure, from random sampling of benign lesion images to classification, was repeated 15 times, due to variations in random sampling and cross-validation partitioning. The estimated feature values are

Divergence between the lesion and the benign model:

$$\hat{d^b} = \hat{I}^*(\hat{p_l}, \hat{p_b}) = \frac{1}{N_m} \sum_{i=1}^{N_m} \frac{\hat{p_l}(\mathbf{x}_i)}{\hat{p_b}(\mathbf{x}_i)} \log \frac{\hat{p_l}(\mathbf{x}_i)}{\hat{p_b}(\mathbf{x}_i)}, \tag{11}$$

where the $\mathbf{x}_i$'s were sampled from the 72 melanomas in the training set, $\hat{p_l}(\mathbf{x})$ is the GMM fitted to a sample from the lesion image, and $\hat{p_b}(\mathbf{x})$ is the GMM fitted to a sample from the 72 excluded benign lesion images, and

Divergence between the lesion and the malignant model:

$$\hat{d^m} = \hat{I}^*(\hat{p_l}, \hat{p_m}) = \frac{1}{N_b} \sum_{i=1}^{N_b} \frac{\hat{p_l}(\mathbf{x}_i)}{\hat{p_m}(\mathbf{x}_i)} \log \frac{\hat{p_l}(\mathbf{x}_i)}{\hat{p_m}(\mathbf{x}_i)}, \tag{12}$$

where the $\mathbf{x}_i$'s are sampled from the 72 excluded benign lesions, and $\hat{p_m}(\mathbf{x})$ is the GMM fitted to a sample from the 72 melanomas in the training set.

The two features were pooled together with the 59 features previously developed on the same data set [25], [27]. Among the 59 features are features for colour distribution, colour counting, blue-grey area, colour variety and specific colour detection. Correlation-based feature selection (CFS), which is classifier independent [28], was performed. This was done without cross-validation, since there is no testing. A benign model, $p_b(\mathbf{x})$, was fitted to the pixel values from 79 randomly selected benign images. These images were then excluded. For each malignant lesion, the malignant model, $p_m(\mathbf{x})$, was fitted to the other 79 malignant lesion images so that the same lesion did not appear in the fitting of $p_l(\mathbf{x}|\boldsymbol{\theta}_{mj'})$ and $p_m(\mathbf{x})$. The whole procedure was repeated 50 times.

## V. Results

The performance of a feature or a classifier can be reported in terms of sensitivity (proportion of melanomas classified as malignant) and specificity (proportion of non-melanomas classified as benign). Receiver operating characteristic (ROC) curves are more informative than a single sensitivity/specificity pair. We report the new features' performance in three ways: (1) ROC curves for the individual features, (2) feature selection, and (3) contribution to a skin lesion classifier.

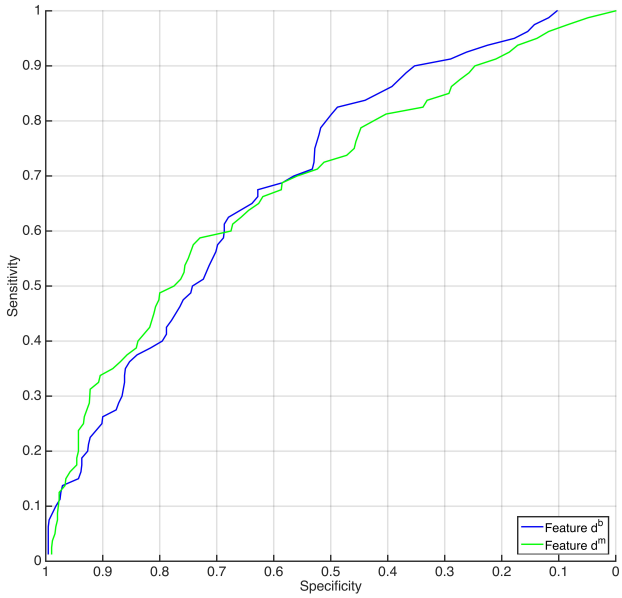



Fig. 3: ROC curves for the individual features: $d^b$ and $d^m$ reflects the divergence between a lesion and a benign model or a malignant model, respectively.

To calculate the sensitivities and specificities of $\hat{d^b}$, the images in the test sets are classified according to a threshold $t$, such that a lesion is classified as malignant if $\hat{d^b} > t$, and benign otherwise. The calculations are done equivalently for $\hat{d^m}$. Fig. 3 shows the ROC curves for $\hat{d^b}$ and $\hat{d^m}$. $\hat{d^b}$ performs better than $\hat{d^m}$ for sensitivities above 70%.

Fig. 4 shows the number of times that $\hat{d^b}$ and $\hat{d^m}$ were selected. Both are selected almost every time. The high frequencies indicate that the two new features' values are correlated to the class labels, but not highly correlated to the existing features, and not highly correlated to each other.

Finally, the features' contribution to a skin lesion classifier is measured. For each cross-validation training set, CFS is used on the 59 previously proposed features. A linear discriminant analysis (LDA) classifier is trained with the selected features, and the sensitivities and specificities are calculated from the test set. A second LDA classifier is trained with $\log(\hat{d^b})$ and $\log(\hat{d^m})$ added to the selected features, and tested accordingly. The logarithm is used since the feature values are not Gaussian distributed, which is the assumption of LDA. The ROC curves for the two classifiers are shown in Fig. 5.

## VI. Discussion

Divergence functions are used for many aspects of image analysis, e.g. segmentation by region merging [16] and image retrieval [29]. There is a wide range of colour feature algorithms for melanoma detection, but to our knowledge, there

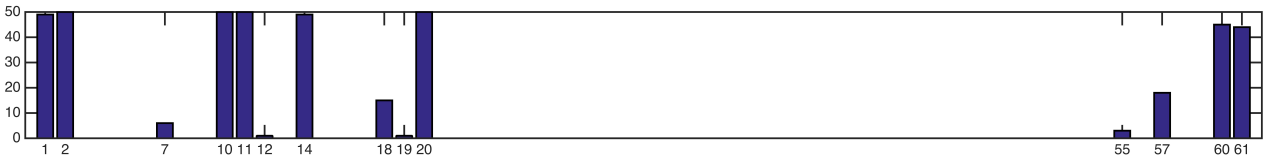



Fig. 4: $f_1$ and $f_2$ are asymmetry features, $f_{10}$ and $f_{11}$ are colour distribution features, $f_{14}$ and $f_{18}$ are border features, $f_{20}$ and $f_{57}$ are specific colour features, and $f_{60} = \hat{d^b}$ and $f_{61} = \hat{d^m}$.

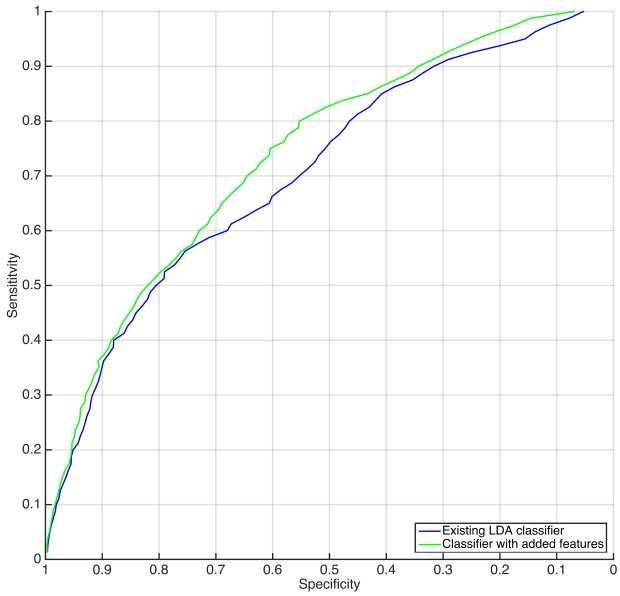



Fig. 5: ROC curves for LDA classifiers: The existing LDA classifier with previously proposed features, and the LDA classifier when adding the two new features $d^b$ and $d^m$.

are none that applies divergence between distributions of pixel values. GMM was used in [30] to estimate the distribution of separate colours, but not of the lesions themselves. In [31], the Kullback-Leibler information was used for nearest-neighbour classification. A drawback of the Kullback-Leibler information is its instability for numerical integration. If $p_b(\mathbf{x}_i) = 0$ while $p_l(\mathbf{x}_i) > 0$ for a single $\mathbf{x}_i$, then $I(p_l, p_b) = \infty$ regardless of all other $\mathbf{x}_i$'s. This occurred for about $1\%$ of the images. By letting $\min(p_b(\mathbf{x}_i)) = \epsilon$, where $\epsilon > 0$ (e.g. the machine epsilon), and since $p_l(\mathbf{x}_i)/p_b(\mathbf{x}_i)$ appears inside the logarithmic function, the Kullback-Leibler information retains stability. Fitting several GMMs for each lesion image is time consuming. An alternative is to pre-define the number of components, $K$, as done in [25].

The assumptions in Section III about the rate at which $p_l(\mathbf{x}|\boldsymbol{\theta}_{bj}) \to 0$ and $p_l(\mathbf{x}|\boldsymbol{\theta}_{mj'}) \to 0$ compared to $p_b(\mathbf{x}) \to 0$ are not true for all lesions. However, the high sensitivities and specificities for $\hat{d^b}$ and $\hat{d^m}$ suggest that they are true for a majority of the lesions.

Many CAD systems for melanoma detection report sensitivity and specificity close to 100% [32], but if feature selection is done on the whole data set, before cross-validation partitions, the observed performance is overly optimistic [21]. The data set at hand impacts the observed performance and direct comparison between systems is not possible. A CAD system based on 53 of the 59 features and an LDA classifier has been tested, and the performance did not deviate from that of three dermatologists [27], which puts it in the same range as state-of-the-art systems [4], [33], [34].

The sensitivities and specificities of single features have limited interpretive value for the features' relevance to melanoma detection. If a new feature is highly correlated with existing features, adding it to the classifier can lower the classifier's performance [35, p.52]. A feature with low sensitivity can be a valuable contribution to a classifier if the melanomas detected by the new feature are those that are misclassified by the existing classifier. However, the sensitivities and specificities indicate how general a feature is for the melanoma class. The proposed colour features are very general, as expected. The result from the classifier-independent feature selection indicates that the proposed features are not highly correlated with the existing features. Finally, the increased sensitivities and specificities when adding the new features to an existing classifier show their value in melanoma detection.

The ROC curves for the LDA classifier with and without the two new features are approximately the same for low sensitivity values. A classifier with low sensitivity is not clinically relevant, due to the cost of misclassifying a melanoma. Sensitivity of minimum $95\%$ has been suggested [36], and at that level, the two new features increase the specificity from $16\%$ to $24\%$. At $20\%$ specificity, adding the two new features increases the sensitivity from $94\%$ to $97\%$. The increases might seem small, but the cost of misclassifying a melanoma can be huge, both in terms of patient survival and treatment costs [37], and even a small increase has a great impact. Increasing sensitivity without decreasing specificity becomes more difficult the higher the sensitivity is. Excision of a lesion carries low risk and has little disadvantage for the patient. However, for the health care system, excising a large number of benign lesions is a burden, since each lesion is examined by an expert pathologist. By increasing the specificity level, valuable resources can be made available for other tasks [38].

Decreasing the size of the data set for feature selection gave more unstable results. Since feature selection is performed for every partition in the cross-validation, variation in the selected feature sets gives variation in the trained classifiers.

10-fold cross-validation, which gives larger training sets than, for example, 5-fold cross-validation, was used. The resulting test sets consist of only 8 melanomas, and small test sets give large variations in the observed performance. The confidence intervals for the ROC curves overlap, and we are not able to conclude that the two new features actually increase the performance of the classifier. Ongoing data collection will provide an independent test set, which can verify the new features' relevance in melanoma detection in the near future.

In summary, the proposed divergence-based colour features are relevant to melanoma detection. This is shown by high frequencies for classifier-independent feature selection, and by increased performance when adding them to an existing LDA classifier, but a final independent verification is needed.

## REFERENCES


[1] Cancer Registry of Norway, "Cancer in Norway 2013 - cancer incidence, mortality, survival and prevalence in Norway," Cancer Registry of Norway, Tech. Rep., 2015.

[2] W. Stolz, O. Braun-Falco, P. Bilek *et al.*, *Color Atlas of Dermatoscopy*, 2nd ed. Berlin: Blackwell Wissenschafts-Verlag, 2002.

[3] G. Argenziano, Soyer, S. Chimenti *et al.*, "Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet," *Journal of the American Academy of Dermatology*, vol. 48, no. 5, pp. 679–693, May 2003.

[4] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artificial Intelligence in Medicine*, vol. 56, no. 2, pp. 69–90, Oct. 2012.

[5] M. E. Celebi and A. Zornberg, "Automated quantification of clinically significant colors in dermoscopy images and its application to skin lesion classification," *Systems Journal, IEEE*, vol. 8, no. 3, pp. 980–984, 2014.

[6] S. Seidenari, C. Grana, and G. Pellacani, "Colour clusters for computer diagnosis of melanocytic lesions." *Dermatology (Basel, Switzerland)*, vol. 214, no. 2, pp. 137–143, 2007.

[7] H. M. Kim and J. M. Mendel, "Fuzzy basis functions: comparisons with other basis functions," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 2, pp. 158–168, May 1995.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[9] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[10] G. McLachlan and D. Peel, *Finite Mixture Models*, ser. Wiley Series in Probability and Statistics, N. A. C. Cressie, N. I. Fisher, I. M. Johnstone *et al.*, Eds. John Wiley & Sons, Inc., 2000.

[11] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.

[12] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[13] L. Pardo, *Statistical Inference Based on Divergence Measures*, ser. Statistics. Chapman and Hall/CRC, Oct. 2006, vol. 185.

[14] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, Dec. 1989.

[15] J.-P. Baudry, A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo, "Combining mixture components for clustering," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 332–353, Jun. 2010.

[16] F. Calderero and F. Marques, "General region merging approaches based on information theory statistical measures," in *15th IEEE International Conference on Image Processing*, Oct. 2008, pp. 3016–3019.

[17] J. Lin, "Divergence measures based on the Shannon Entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[18] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[19] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer, Aug. 2004.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, 2nd ed., ser. Springer Series in Statistics. New York: Springer, 2009.

[21] P. Smialowski, D. Frishman, and S. Kramer, "Pitfalls of supervised feature selection," *Bioinformatics*, vol. 26, no. 3, pp. 440–443, Feb. 2010.

[22] M. A. Babyak, "What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models." *Psychosomatic medicine*, vol. 66, no. 3, pp. 411–421, May 2004.

[23] J. Shao, "Linear model selection by cross-validation," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 486–494, Jun. 1993.

[24] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

[25] K. Møllersen, M. Zortea, K. Hindberg *et al.*, *Improved skin lesion diagnostics for general practice by computer aided diagnostics*, ser. Digital Imaging and Computer Vision. CRC Press/Taylor & Francis, in press.

[26] M. Zortea, S. O. Skrøvseth, T. R. Schopf, H. M. Kirchesch, and F. Godtliebsen, "Automatic segmentation of dermoscopic images by iterative classification," *International Journal of Biomedical Imaging*, vol. 2011, pp. 1–19, 2011.

[27] M. Zortea, T. R. Schopf, K. Thon *et al.*, "Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists," *Artificial Intelligence in Medicine*, vol. 60, no. 1, pp. 13–26, Jan. 2014.

[28] M. A. Hall and L. A. Smith, "Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper," in *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*. AAAI Press, 1999, pp. 235–239.

[29] P. Piro, S. Anthoine, E. Debreuve, and M. Barlaud, "Image retrieval via Kullback-Leibler divergence of patches of multiscale coefficients in the KNN framework," in *Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on*. IEEE, Jun. 2008, pp. 230–235.

[30] C. Barata, M. Figueiredo, M. Emre Celebi, and J. S. Marques, "Color identification in dermoscopy images using Gaussian mixture models," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, May 2014, pp. 3611–3615.

[31] C. Barata, M. Ruela, T. Mendonça, and J. Marques, "A bag-of-features approach for the classification of melanomas in dermoscopy images: The role of color and texture descriptors," in *Computer Vision Techniques for the Diagnosis of Skin Cancer*, ser. Series in BioEngineering, J. Scharcanski and M. E. Celebi, Eds. Springer Berlin Heidelberg, 2014, pp. 49–69.

[32] A. Blum, I. Zalaudek, and G. Argenziano, "Digital image analysis for diagnosis of skin tumors," *Seminars in Cutaneous Medicine and Surgery*, vol. 27, pp. 11–15, 2008.

[33] B. Rosado, S. Menzies, A. Harbauer *et al.*, "Accuracy of computer diagnosis of melanoma: A quantitative meta-analysis," *Archives of Dermatology*, vol. 139, no. 3, pp. 361–367, Mar. 2003.

[34] M. E. Vestergaard and S. W. Menzies, "Automated diagnostic instruments for cutaneous melanoma." *Seminars in cutaneous medicine and surgery*, vol. 27, no. 1, pp. 32–36, Mar. 2008.

[35] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, Hamilton, New Zealand, 1999.

[36] G. Monheit, A. B. Cognetta, L. Ferris *et al.*, "The Performance of MelaFind: A Prospective Multicenter Study," *Archives of dermatology*, vol. 147, no. 2, pp. 188–194, Feb. 2011.

[37] S. N. Markovic, L. A. Erickson, R. D. Rao *et al.*, "Malignant melanoma in the 21st century, part 2: Staging, prognosis, and treatment," *Mayo Clinic Proceedings*, vol. 82, no. 4, pp. 490–513, Apr. 2007.

[38] B. Lindelöf, M.-A. Hedblad, and U. Ringborg, "Nevus eller malignt melanom? Rätt kompetens vid diagnostik ger lägre kostnader," *Läkartidningen*, vol. 105, no. 39, pp. 2666–2669, 2008.