Tackling the ordinal and imbalance nature of a melanoma image classification problem

M. Pérez-Ortiz
Dept. of Quantitative Methods
Universidad Loyola Andalucía
14004 Córdoba, Spain
Email: mariaperez@uloyola.es

A. Sáez
Dept. of Signal Theory
and Communications
University of Seville
41092 Seville, Spain
Email: aurorasaez@us.es

J. Sánchez-Monedero
Dept. of Quantitative Methods
Universidad Loyola Andalucía
14004 Córdoba, Spain
Email: jsanchez@uloyola.es

P.A. Gutiérrez
Dept. of Computer Science
and Numerical Analysis
University of Córdoba
14071 Córdoba, Spain
Email: pagutierrez@uco.es

C. Hervás-Martínez
Dept. of Computer Science
and Numerical Analysis
University of Córdoba
14071 Córdoba, Spain
Email: chervas@uco.es

***Abstract*—Melanoma is a type of cancer that usually occurs on the skin. Early detection is crucial for ensuring five-year survival (which varies between 15% and 99% depending on the melanoma stage). Melanoma severity is typically diagnosed by invasive methods (e.g. a biopsy). In this paper, we propose an alternative system combining image analysis and machine learning for detecting melanoma presence and severity. The 86 features selected consider the shape, colour, pigment network and texture of the melanoma. As opposed to previous studies that have focused on distinguishing melanoma and non-melanoma images, our work considers a finer-grain classification problem using five categories: benign lesions and 4 different stages of melanoma. The dataset presents two main characteristics that are approached by specific machine learning methods: 1) the classes representing melanoma severity follow a natural order, and 2) the dataset is imbalanced, where benign lesions clearly outnumber melanoma ones. Different nominal and ordinal classifiers are considered, one of them being based on an ordinal cascade decomposition method. The cascade method is shown to obtain good performance for all classes, while respecting and exploiting the order information. Moreover, we explore the alternative of applying a class balancing technique, presenting good synergy with the ordinal and nominal methods.**

***Index Terms*—melanoma, computer vision, machine learning, ordinal classification, imbalanced classification**

I. Introduction

Melanoma is a type of cancer that develops from the pigment-containing cells known as melanocytes. The most common type is the cutaneous melanoma, which occurs on the skin. In the US, approximately 74,000 melanomas will be diagnosed by the end of 2015, with near 10,000 of those resulting in death [1]. In Europe, approximately 100,000 cases are yearly diagnosed with a similar death ratio [2]. It is well accepted that only early detection can reduce mortality, since the patient prognosis depends on the tumour thickness at the time of the surgical treatment [3].

Melanoma must be detected before the tumour has penetrated the epidermis (i.e. before the thickness is higher than $> 0.76mm$). In that case, five-year survival rate is about 99%, otherwise dropping to 15% for patients with advanced disease [4]. Detection is currently performed by trained professionals and consists in a visual inspection using a dermatoscope. Prognosis and severity are evaluated by measuring the depth of melanoma by a biopsy. However, recent works propose new tools to aid or to improve the prognosis [4], mainly based on dermoscopic image analysis. Although different lines are being researched, image analysis methods present the advantage of being cheap and relatively easy to be combined with existing detection procedures and can be integrated with strategies such as machine learning.

Computerised dermoscopy image analysis systems have been proposed in the last years to assist the diagnosis of pigmented lesions [5]. Most of these works focus on the segmentation and distinction of melanomas from benign lesions [6], [7]. The scarcity of studies on this topic and the inherent difficulty of the problem makes it an imperative avenue for research. In this sense, Rubegni et. al [8] addressed the characterisation of two types of melanoma based on their thickness. Their work uses 49 features related to colour, geometry and texture, extracted from a private database of 141 images obtained with the company hardware system. Finally, a recent work [9] focused on classifying three degrees of thickness for melanomas with promising results, using a set of 81 features. The problem of automatically characterising melanoma thickness by image analysis poses a great research challenge and as said, a finer grain classification will be required for appropriate prognosis.

In this paper, we pursue the objective of developing a system for distinguishing melanoma presence and thickness using a five-class classification problem (the first class representing benign lesions and the rest different levels of melanoma thickness). Up to our best knowledge, it is the first time that such a classification scenario is considered for this challenging problem. To do so, as said, we combine image analysis methods and different machine learning strategies.

The proposed system considers a set of 86 input features, comprising characteristics that have been previously highlighted for the distinction of benign lesions and melanomas and the depth of melanomas. Features for distinguishing melanoma from non-melanoma are based on the ABCD method, a useful tool for the diagnosis of pigmented lesions. This method analyses four clinical characteristics to identify a malignant melanoma: asymmetry (A), border irregularity (B), colour variegation (C) and differential structures (D). Features to estimate the melanoma thickness are based on clinical findings that correlate certain visual characteristics present in dermoscopic images and tumour depth [10].

Note that the label used for melanoma thickness classification is imbued with order information (see Table I), and, consequently, it should be modelled by using ordinal classifiers [11]. Ordinal methods exploit the ordered nature of the classes for constructing the classifier and impose different misclassification errors (misclassifying a stage 0 melanoma with a stage I one should not be considered the same than confusing it with a stage III melanoma). In our case, we also consider a class for benign lesions, which should be treated with caution because it can not be assumed that benign lesions will also follow a natural order with respect to melanoma lesions. However, misclassification costs should be maintained (it is preferred to misclassify a benign lesion with a stage 0 melanoma than with thicker melanomas). Given this partial order of the data and its imbalanced nature, we apply a cascade binary utility ordinal model [12], whose formulation is able to take the partial order in our problem and the imbalanced nature of the labels. Additionally, we consider a strategy for ordinal data over-sampling [13] for tackling the ordinal imbalanced nature of the dataset. The results are compared considering different nominal and ordinal classifiers.

The rest of the paper is organised as follows: Section II presents the clinical problem and the dataset; Section III describes the set of features selected to describe the images; Section IV presents some previous notions and describes the over-sampling technique used and the cascade binary utility model considered; Section V shows the experiments performed and analyses the results; and finally, Section VI outlines some conclusions and future work.

II. Data description

Tumour thickness is directly correlated with greater access to lymph capillaries, which is a common method of cancer spread, which then greatly influences the prognosis. If the melanoma is confined to the epidermis, it is referred as 'in situ' melanoma, and it is curable by removal surgery. However, as the cancerous cells reach the deepest layer of the skin (the dermis), it is known as invasive melanoma, whose survival rate worsens with the depth of invasion.

The Breslow index [14] measures melanoma thickness by means of an incisional or excisional biopsy of the suspected lesion [10]. The maximal thickness of the lesion is measured in millimetres from the top of the granular cell layer to the deepest point of invasion. Tumour thickness is proposed as a valuable tool in prognosing patients survival, as well as to establish the surgical margin excision width [15], [16], and to select patients for prophylactic lymph node dissection [15], [17], which is a surgical procedure to determine if cancer has spread to the lymphatic system. Consequently, it is useful to have reliable preoperative parameters on tumor thickness so as to ensure a correct surgical approach and to assess the risk of progression.

TABLE I: Stages of melanoma according to thickness and number of images for each class

Stage	Location	Class	Number of images
Non-melanoma ($\mathcal{C}_1$)	-	1	313
Stage 0 ($\mathcal{C}_2$)	In situ	2	64
Stage I ($\mathcal{C}_3$)	<0.76 mm	3	102
Stage II ($\mathcal{C}_4$)	0.76 mm - 1.50 mm	4	54
Stage III ($\mathcal{C}_5$)	>1.50 mm	5	29

Table I includes different stages of thickness based on the Breslow index. In this work, we have collected 556 images from the Interactive Atlas of Dermoscopy [18], a multimedia project for medical education with pigmented skin lesions images in which all lesions were biopsied and diagnosed histopathologically. The images have been classified in five classes: non-melanoma (i.e. benign lesions) and four stages of melanoma depth. The number of patterns belonging to each class is also included in Table I, where the imbalanced nature of the data can be appreciated.

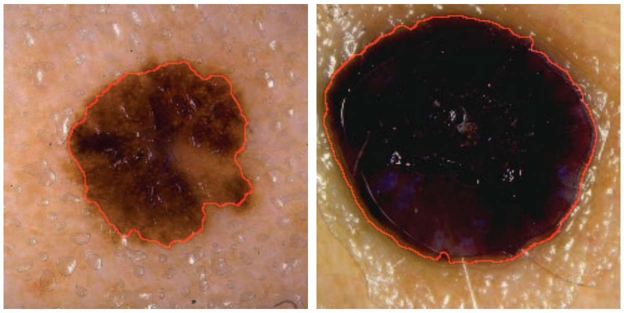
(a) Melanoma <0.76 mm.
(b) Melanoma ≥ 1.5 mm.

Fig. 1: Examples of segmented melanomas

All the images have the same resolution (768x512 pixels). They have been segmented using the automatic segmentation algorithm proposed in [19], based on a level-set technique applied to a perceptually adapted colour gradient [20]. Figure 1 presents two examples of segmented melanomas.

III. Feature extraction

The feature extraction process specifically proposed in this paper aims to mimic dermatologist assessment, using characteristics defined in the clinical ABCD rule (to distinguish between benign lesions and melanomas) and features inspired by the findings derived from clinical studies regarding the correlation between certain characteristics seen in dermoscopic images and melanoma thickness. A total of 86 descriptors (x_1-x_{86}) based on shape, colour and texture have been extracted. Note that, in practice, most classification methods could benefit from a feature selection process. However, for kernel methods (e.g. for the support vector paradigm) the capacity control is equivalent to some form of regularisation so that denoising (or feature selection in this case) might not necessary although it could be very useful for other unregularised methods [21].

A. *Shape features*

Shape features have been extracted to satisfy the asymmetry (A) and irregularity border (B) criteria. We use the circularity index (computed as 4π multiplied by lesion area, divided by its squared perimeter) (x_1) [5], the perimeter normalised by the equivalent perimeter (perimeter of a circle with the same area as the lesion) (x_2), the variance of the distance of the border lesion points from the centroid location (x_3) [22], eccentricity (a measure of elongation) (x_4) [7] and length of major axis of the lesion normalised with respect to the equivalent diameter (diameter of a circle with the same area as the lesion) (x_5). In order to evaluate the lesion asymmetry, first, the major axis orientation of the lesion has been computed, and secondly, it has been rotated the same number of degrees clockwise to align the principal axes with the image (x and y) axes. Then, the lesion has been folded around the x-axis, and the percentage of overlapping area with respect to the total area has been computed to obtain the horizontal asymmetry (x_6). The same procedure has been performed for the y-axis to obtain vertical asymmetry (x_7). If the process is repeated taking into account the percentage of overlapping pixels assigned to the same colour (see Section III-B for the colour assignation), we can compute the colour horizontal asymmetry (x_8) and the colour vertical asymmetry (x_9).

B. *Colour features*

Colour features are one of the most determinant type of features in the estimation of melanoma depth. Different dermoscopic structures, which have been found discriminative for melanoma thickness estimation, are associated with different colours. We have extracted features related to the six colours that physicians assess in the pigmented lesions: black, dark brown, light brown, blue-grey, red and white [23]. These colours appear depending on the depth of the melanoma [24]. To describe these colours, we segment each lesion into their constituting colours by a similar approach to that proposed by Seidenari et al. [25], in which a colour palette formed by 144 patches that present unequivocally one of the six possible colours is developed. This palette was used to extract the colour regions of the lesions from the patches according to a nearest neighbour approach. Each pixel of the image was assigned to the colour patch that minimised its Euclidean distance in the CIE $L^*a^*b^*$ colour space. From this colour identification different descriptors are extracted: the percentage of the lesion area classified as these six colours (x_{10}-x_{15}), number of colours that each lesion presents (colour criterion of ABCD) (x_{16}), and 24 additional statistical descriptors of the colours (mean, standard deviation, kurtosis, and skewness of each colour, x_{17}-x_{40}).

C. *Pigment network features*

Pigment network is a dermoscopic structure represented as a regular grid of brownish lines over a diffuse light-brown background [18]. It is referred by many authors as one of the most discriminative features for melanoma thickness [15], [26], [27], [28], being inversely correlated with melanoma depth [18]. This structure is identified searching for the network 'holes' applying a filtering and thresholding step using the Otsu's method [29]. Once these regions are identified, we considered the two conditions relative to area size and colour proposed in the work of Sadeghi et al. [30] to remove those wrongly detected regions. The features extracted from this detection are network density ratio (x_{41}), number of nodes (x_{42}) and number of links or edges (x_{43}).

D. *Texture features*

Other dermoscopic structures such as vascular pattern [26], [15], blue-grey veil [26], [15], white scar-like areas [27], and dots or globules [28] have been found to have relation with the depth of melanoma. These are usually associated with texture features. To capture properties of different structures, we have extracted three sets of texture features from three different approaches: 19 features from the gray level co-occurrence matrix (GLCM) [31] ($x_{44}-x_{62}$), 18 features based on a Markov random field (MRF) model [19] ($x_{63} - x_{80}$) and 6 features from local binary pattern (LBP) histograms ($x_{81} - x_{86}$).

IV. Methodology

This section presents the over-sampling and classification strategies considered in this paper.

Consider a training sample $T = \{\mathbf{x}_i, y_i\}_{i=1}^N \subseteq \mathcal{X} \times \mathcal{Y}$ generated i.i.d. from a (unknown) joint distribution $P(\mathbf{x}, y)$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_Q\}$. Let N be the number of patterns in the training sample, N_q the number of samples for the q-th class and X_q the set of patterns belonging to class $\mathcal{C}_q$. In the ordinal regression setup, the labelling space is ordered due to the data ranking structure ($\mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_Q$, where $\prec$ denotes this order information). Note that, in our case, this order structure only applies to four of the five classes of the problem.

A. *Previous notions: Support Vector Machines*

The Support Vector Machine paradigm (SVM) [32], [33] is one the most common kernel learning methods for statistical pattern recognition because of its good generalisation ability and absence of local minima. The basic idea is the separation of two classes through a hyperplane specified by a normal vector $\mathbf{w}$ and a bias b. The optimal separating hyperplane is the one that maximises the distance between the hyperplane and the nearest points in both classes (called margin). Beyond the application of kernel techniques to allow non-linear decision discriminants (the kernel trick), another generalisation was made to replace hard margins with soft margins [33], using the so-called slack-variables ξ_i in order to avoid inseparability, relax the constraints and handle noisy data. Therefore, SVM seeks for a classifier $f : \mathbb{R}^d \rightarrow \mathbb{R}$ of the form $f(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x}) + b$ (Φ being the mapping function induced by the kernel) that minimises the objective function:

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i, \tag{1}$$

for some parameter C and subject to the constraints:

$$y_i((\mathbf{w} \cdot \Phi(\mathbf{x}_i)) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall_i \in \{1, \ldots, N\}.$$

Nonetheless, this 1-norm SVM solution for $\mathbf{w}$ is the minimiser of the regularized empirical loss function, thus being also defined as:

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}(1 - y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b))_+,$$

where $(x)_+ = \max(x, 0)$.

As the SVM paradigm was originally proposed for binary classification problems, it had to be reformulated to deal with multiclass problems [34] (one-against-one and one-against-all binary decompositions, among other proposals).

B. Ordinal cascade classification model

A modification of the binary decomposition method known as the cascade linear utility model [35] is used in this paper. This procedure considers $Q-1$ models (each model D_i will be comprised in this case of a projection $\mathbf{w}_i$ and a threshold b_i), in such a way that model q separates class $\mathcal{C}_q$ from classes $\mathcal{C}_{q+1} \vee \ldots \vee \mathcal{C}_Q$, so that not all the classes are considered in the computation of each model (as can be seen in Fig. 2, where the decomposition is graphically described). This methodology is used to balance the projections (due to the imbalanced character of the sample considered in this paper) and to consider the partial order of the classes (in this sense, we will only consider a model with $\mathcal{C}_1$, i.e. the first model which discriminates between $\mathcal{C}_1$ and the rest, which does not assumes any ordinal constraint). This approach is usually known as one-against-followers in the literature [11].

The training set for model or decision maker $D_q = \{\mathbf{w}_q, b_q\}$ is specified by $\{\mathbf{X}_{(i|i=q)}, \mathbf{X}_{(j|j>q)}\}$. Therefore, a coding matrix $\mathbf{M}_{(Q-1 \times Q)}$ associated to the $Q-1$ binary decompositions of the cascade utility model can be defined as follows:

$$\mathbf{M} = \begin{pmatrix} -1 & +1 & +1 & +1 & +1 \\ 0 & -1 & +1 & +1 & +1 \\ 0 & 0 & -1 & +1 & +1 \\ 0 & 0 & 0 & -1 & +1 \end{pmatrix},$$

where the label -1 is assigned to patterns corresponding to the negative class, the label $+1$ to patterns belonging to the positive class, and finally, the patterns associated with label 0 are excluded for the training process. In this way, the approach considered is the same than in [13] but using a one-against-followers approach. A matrix of predictions can be obtained by means of a single model (e.g. using artificial neural networks) or by multiple models (training a binary classifier for each subproblem, as in this paper) [11]. Once the model or models have been trained, a set of $Q-1$ decision values $\mathbf{f}(\mathbf{x})$ are obtained for pattern $\mathbf{x}$. Concerning the prediction phase, usually a hierarchical approach is followed. In our case, we consider a different strategy (as done in [13]): the Error-Correcting Output Codes framework (ECOC). The principal idea is to associate each class $\mathcal{C}_q \in \mathcal{Y}$ with a column of the binary coding matrix $\mathbf{M}$ previously introduced. The binary classifier is trained once for each row of the matrix using the induced binary classification problem, f yielding $Q-1$ hypotheses. Prediction is then accomplished by choosing the column of $\mathbf{M}$ closest to the set of decision values $\mathbf{f}(\mathbf{x}) = f_1(\mathbf{x}), \ldots, f_{Q-1}(\mathbf{x})$. When the coding matrix contains a 0, this leads to an indifferent condition in the prediction phase [36]. According to this, the final decision function is the following one:

$$C(\mathbf{x}) = \mathcal{C}_q, \text{where } \ q = \arg\min_{q=1\ldots Q} d(\mathbf{M}_q, \mathbf{f}(\mathbf{x}))$$

where $\mathbf{M}_q$ is the q-th row of matrix $\mathbf{M}$ and d is the loss function considered. The main issue within this paradigm is the choice of a loss function which corresponds with the loss function used for deriving the binary classifier. In this case, due to the choice of the 1-norm SVM paradigm as the base methodology, the hinge-loss function is chosen, which is the most commonly used for SVM. To see this, we formulate Eq.

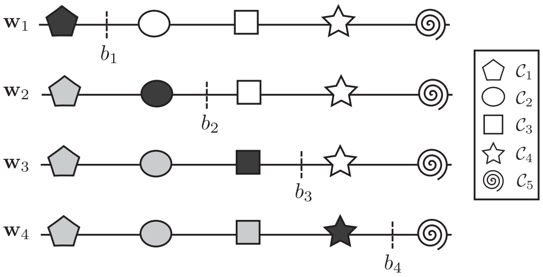

Fig. 2: Binary decompositions performed for a five-class problem, where $\mathbf{w}_i$ represents the i-th projection and b_i the bias associated to that projection. White-shadowed figures represent the negative class, black-shadowed ones the positive one and grey ones the classes not used in each model. The shape represents the original category. Note that this is an ordinal decomposition, as adjacent classes are grouped together (except for the first class, which by the problem definition presents a different relationship with the rest of classes).

1 in terms of the error:

$$\underbrace{\frac{1}{2}||\mathbf{w}||^2}_{minimiser} \underbrace{+C\sum_{i=1}^{N}\text{loss}(y_i, f(\mathbf{x}_i)),}_{error}$$

where the error function chosen is usually the hinge-loss (for L1-SVM), or its square (for L2-SVM):

$$\begin{aligned}\text{loss}(y_i, f(\mathbf{x}_i)) &= (1 - y_i \cdot f(\mathbf{x}_i))_+ = \xi_i = \\ &= \max(0, 1 - y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b)).\end{aligned} \quad (2)$$

The different possibilities for $y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b)$ are:

- $y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) > 1$: the point is well-classified and outside the margin.
- $y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) = 1$: the point is on the margin.
- $y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) < 1$: the point is within the margin or misclassified.

In this way, the final decision function is $C(\mathbf{x}) = \mathcal{C}_q$, where:

$$q = \arg\min_{q=1\ldots Q} \sum_{i=1}^{Q-1} \max\left(0, (1 - \mathbf{M}_{qi} \cdot f_i(\mathbf{x}))\right)$$

One of the main advantages of this methodology over a purely hierarchical approach is that real values for prediction are used instead of binary predicted class values; thus the model is provided with additional information which may be useful for improving its performance. Note that the real values used are related with the distance to the threshold, a measure usually used for estimating the probability of belonging to a class.

C. *Ordinal pattern over-sampling*

Imbalanced data arise naturally in ordinal classification problems, since there are usually classes that are naturally of lower probability (e.g. extreme classes) [13]. The application considered in this paper is one example: there are significantly more patterns associated to benign lesions with respect to melanomas (specially when considering thick ones). Because of this, we consider the application of a recently proposed method for class balancing in ordinal classification problems [13]. This technique creates synthetic patterns considering the data distribution of minority classes and the data ordering. The main assumption of this method is that the ordering of the classes should be considered when resampling patterns for an ordinal classification problem and that this order is generally represented by a latent manifold. To exploit this manifold, the structure of the data is captured constructing a pattern graph, and the paths that preserve the ordinal constraints of the data are considered for over-sampling. Moreover, new patterns are created in the borderline between adjacent classes, in order to smooth the ordinal nature of the dataset.

In this paper, we consider one of the proposals of [13], named as: ordinal graph-based over-sampling via shortest paths using a probability function for the intra-class edges (OGO-SP), as suggested in [13]. The classes over-sampled are the ones which present an imbalance ratio higher than a considered threshold (in our case 1.5 as suggested in other works [37]) and the number of synthetic new patterns is that needed to obtain an imbalance ratio lower than this threshold. The mean imbalance ratio is defined as:

$$IR = \frac{1}{Q}\sum_{q=1}^{Q} IR_q, \quad (3)$$

where IR_q is the imbalance ratio associated to $\mathcal{C}_q$:

$$IR_q = \frac{\sum_{j\neq q} N_j}{Q \cdot N_q}. \quad (4)$$

For more information about this procedure refer to [13]. Given the previously assumed partial ordering of the data, the over-sampling technique is only considered for melanoma classes (obviating the benign lesion class). In this sense, the classes over-sampled are the following: $\mathcal{C}_2, \mathcal{C}_4, \mathcal{C}_5$ (the number of patterns per class is included in Table 1), given that the imbalance ratio of $\mathcal{C}_3$ is lower than 1.5.

V. Experiments

This section includes a description of the experiments performed and analyses the results obtained. As stated before, different classifiers (nominal and ordinal) are compared in this section. More specifically, we compare our ordinal proposal (using a cascade binary utility model) with two nominal classifiers (which are based on a decomposition of the labelling space) and three ordinal classifiers (two of them being the ordinal counterpart of the nominal methods and the other performing also a labelling decomposition). The methods tested are the following:

- Support Vector Classifier (SVC) using the one-against-one approach [38].
- Logistic Regression (LR) where the classification model is composed of several binary LR models using the one-against-all scheme.
- The reformulation of SVMs for ordinal classification with implicit constraints (SVORIM) [39].
- The Proportional Odds Model (POM), which adapts the standard logistic regression to the ordinal case.
- Frank & Hall decomposition approach for ordinal classification (FHSVM) [40] using SVMs as base classifiers and the weighting scheme proposed in [41].
- Ordinal Cascade binary utility model using the ECOC framework (OC-ECOC) [12]. The model is described in Section IV-C, and, in contrast to SVORIM and POM, it is able to consider the potential partial ordering of the classes.

The performance of these 6 classifiers is evaluated on the original dataset and the over-sampled one.

Different metrics have been proposed for measuring the performance of machine learning classifiers. In our case, we have selected different metrics that evaluate the global performance, the balance of performance for the different classes and the ordinal magnitude of the errors. The metrics selected are the following:

- Accuracy (Acc) is the percentage of correctly classified patterns:
$$Acc = 100 \cdot \frac{1}{N} \sum_{i=1}^{N} [\![\hat{y}_i = y_i]\!],$$
where $[\![\cdot]\!]$ is the indicator function (being 1 if the condition is true, and 0 otherwise) and $\hat{y}_i$ is the predicted target for $\mathbf{x}_i$.
- The geometric mean of the sensitivities (GM) is typically used in imbalanced problems [42]:
$$GM = 100 \cdot \sqrt[Q]{\prod_{q=1}^{Q} S_q},$$
where S_q is the sensitivity (accuracy ratio) of the classifier for class q. If $GM = 0$ the classifier is totally misclassifying at least one class.
- The Mean Absolute Error (MAE) is the average deviation in absolute value of the predicted class from the true class. It is the most commonly used ordinal classification metric. For imbalanced datasets, this measure is modified to consider the relative frequency of the classes, resulting in the Average MAE ($AMAE$) and Maximum MAE ($MMAE$) [43]:
$$AMAE = \frac{1}{Q} \sum_{q=1}^{Q} MAE_q = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{N_q} \sum_{i=1}^{N_q} e(\mathbf{x}_i), \quad (5)$$
$$MMAE = \frac{1}{Q} \max_{q=1}^{Q} MAE_q, \quad (6)$$
where $e(\mathbf{x}_i) = |\mathcal{O}(y_i) - \mathcal{O}(\hat{y}_i)|$ is the distance between the true and the predicted ranks and $\mathcal{O}(\mathcal{C}_q) = q$ is the position of the q-th label. $AMAE$ values range from 0 to $Q - 1$, and so do $MMAE$ values.

Note that, although the classification problem here considered is hypothesised to be partially ordered in the input space, ordinal metrics should be used, given that ordinal misclassification costs should also be considered for the first class ($\mathcal{C}_1$).

The experiments have been performed using a 10-fold partition procedure. The metrics are calculated using the sum of all generalisation confusion matrices from the 10 folds. To adjust the hyper-parameters (kernel width and cost parameter) for the SVM-based methods (SVC, SVORIM, FHSVM and OC-ECOC), a nested cross-validation is applied to the training data, with a grid search with parameter values within the range $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. The k parameter associated to the k-nearest neighbour for OGO-SP oversampling method (see Section IV-C and [13]) is set to 5. The criteria for selecting the parameters is $AMAE$.

TABLE II: Experimental results obtained for the original dataset and the over-sampled one.

Method	Acc	GM	$AMAE$	$MMAE$
		Original dataset results		
SVC	**65.66**	35.92	0.876	1.315
LR	62.81	36.41	0.937	1.552
SVORIM	63.70	34.30	*0.803*	1.379
POM	63.35	*39.13*	0.855	*1.241*
FHSVM	*65.48*	32.94	0.900	1.448
OC-ECOC	58.19	**41.15**	**0.736**	**1.063**
		OGO-SP preprocessed dataset results		
SVC	**66.90**	*42.02*	0.850	1.241
LR	63.88	41.52	0.845	*1.207*
SVORIM	61.75	**43.00**	**0.789**	1.241
POM	*64.06*	39.98	*0.820*	*1.207*
FHSVM	**66.90**	39.77	0.877	1.448
OC-ECOC	58.36	40.05	0.828	**1.063**

The results of the experiments performed can be seen in Table II, where the best performing method is highlighted in bold face and the second one in italics. Note that Acc and GM are to be maximised, while $AMAE$ and $MMAE$ should be minimised. From these results, several conclusions can be drawn. Firstly, the proposed strategy is the one that presents the best performance for the ordinal and imbalanced metrics (GM, $AMAE$ and $MMAE$) when considering the original dataset, meaning this that tackling this problem as a partially ordered one is the best approach and that the considered cascade binary utility model is suitable for this purpose. Secondly, both the nominal and ordinal SVM classifiers obtain quite similar results (compare SVC with SVORIM), as opposed to the case of LR, where the ordinal version (POM) obtains better performance. This might be because of the demonstrated superiority of the one-against-one approach of SVC over the one-against-all considered for LR. Concerning the application of the OGO-SP over-sampling, the use of this technique generally improves the results for Acc and GM. It could be said that the ordinal oversampling presents good synergy with all methods except OC-ECOC, helping to order the classes in a more appropriate manner. The combination of OC-ECOC and the ordinal SMOTE is not satisfactory, as it leads to similar or worse results. This could indicate that the use of the proposed cascade classifier is enough to tackle the imbalance nature of the dataset without the need of using class balancing techniques. Note that this model has been proposed to consider imbalanced data, in such a way that the decompositions considered are those that balance the distribution of classes. Nonetheless, the over-sampling approach is clearly beneficial for the rest of classifiers, e.g. for SVORIM, which obtains competitive results in Acc and GM but worst results than our proposal for $AMAE$ and $MMAE$ (the metrics that capture the class ordering).

Each pair of algorithms was compared by means of the Wilcoxon test [44] using the 10 different partitions, where a level of significance of $\alpha = 0.05$ was considered. The results of these tests are shown in TABLE III, where the number of wins (W), draws (D) and loses (L) is shown. It can be appreciated that most algorithms benefit from the use of the over-sampling technique (except OC-ECOC which already tackles the imbalanced nature of the data in an algorithmic manner). This is specially true for GM, but also applies to $AMAE$ and $MMAE$. Note also that our proposal (OC-ECOC) shows the best results in terms of $AMAE$ and

TABLE III: Wilcoxon statistical test obtained (W or wins, D or draws and L or loses) for the original dataset and the over-sampled one.

Metric	GM			$AMAE$			$MMAE$		
Method	W	D	L	W	D	L	W	D	L
SVC	0	5	6	1	8	2	0	9	2
SVC-OGO-SP	3	8	0	1	9	1	0	9	2
LR	0	7	4	0	4	7	0	6	5
LR-OGO-SP	4	7	0	0	10	1	0	10	1
SVORIM	0	10	1	5	6	0	1	9	1
SVORIM-OGO-SP	6	5	0	3	8	0	2	9	0
POM	0	8	3	1	8	2	0	10	1
POM-OGO-SP	0	11	0	3	8	0	1	10	0
FHSVM	0	6	5	0	6	5	0	9	2
FHSVM-OGO-SP	4	7	0	0	8	3	0	8	3
OC-ECOC	1	10	0	7	4	0	8	3	0
OC-ECOC-OGO-SP	2	8	1	1	9	1	5	6	0

$MMAE$, metrics that represent the ordering of the classes. In fact, both variants of OC-ECOC show also competitive results in GM, given that there are no statistically significant differences in the comparisons (neither loses or wins). In general, the maximum number of statistically significant wins was obtained by SVORIM-OGO-SP in GM and by OC-ECOC in $AMAE$ and $MMAE$.

VI. Conclusions

This paper presents a novel approach for automatic melanoma characterisation (considering both presence and thickness), via computational image analysis and machine learning. The problem is described by using a partially ordered labelling structure, which is also naturally imbalanced. To deal with these issues, we proposed the use of an ordinal cascade decomposition method and an ordinal oversampling technique. The performance results obtained from the experiments here considered are promising, demonstrating that the features selected describe the lesions properly and that the use of more complex and specialised machine learning techniques is crucial for constructing accurate models in this case. Moreover, this paper has shown that the characterisation of melanoma thickness and their distinction from other benign lesions is suitable using image analysis and machine learning. Both options (ordinal cascade decomposition and ordinal oversampling methods) are shown to be promising tools for alleviating the imbalanced nature of the dataset and the partially ordered structure of the data, although their combination does not lead to a performance improvement.

As future work, an interesting avenue for research would be the analysis of the most discriminative features from the set here considered, as well as the use of additional features.

Acknowledgment

This work was partly financed by a grant provided by the TIN2014-54583-C2-1-R project of the Spanish Ministry of Economy and Competitively (MINECO), by FEDER Funds and by the P11-TIC-7508 project of the Junta de Andalucía, Spain.

References

[1] National Cancer Institute, "Seer stat fact sheets: Melanoma of the skin." 2015, accessed December 15, 2015. Available on http://seer.cancer.gov/statfacts/html/melan.html. [Online]. Available: http://seer.cancer.gov/statfacts/html/melan.html

[2] International Agency for Research on Cancer. World Health Organization, "Cancer factsheet. malignant melanoma of skin." 2015, accessed December 15, 2015. Available on http://eco.iarc.fr/eucan/Cancer.aspx?Cancer=20.

[3] M. Pizzichetta, G. Argenziano, R. Talamini, D. Piccolo, A. Gatti, G. Trevisan, G. Sasso, A. Veronesi, A. Carbone, and H. Peter Soyer, "Dermoscopic criteria for melanoma in situ are similar to those for early invasive melanoma," *Cancer*, vol. 91, no. 5, pp. 992–997, 2001.

[4] C. Herman, "Emerging technologies for the detection of melanoma: Achieving better outcomes," *Clinical, Cosmetic and Investigational Dermatology*, vol. 5, pp. 195–212, 2012.

[5] I. Maglogiannis and C. N. Doukas, "Overview of advanced computer vision systems for skin lesions characterization," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 5, pp. 721–733, 2009.

[6] R. Garnavi, M. Aldeen, and J. Bailey, "Computer-aided diagnosis of melanoma using border- and wavelet-based texture analysis," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1239–1252, 2012.

[7] M. Celebi, H. Kingravi, B. Uddin, H. Iyatomi, Y. Aslandogan, W. Stoecker, and R. Moss, "A methodological approach to the classification of dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 31, no. 6, pp. 362–373, 2007.

[8] P. Rubegni, G. Cevenini, P. Sbano, M. Burroni, I. Zalaudek, M. Risulo, G. Dell'Eva, N. Nami, A. Martino, and M. Fimiani, "Evaluation of cutaneous melanoma thickness by digital dermoscopy analysis: a retrospective study," *Melanoma research*, vol. 20, no. 3, pp. 212–217, 2010.

[9] A. Sáez, J. Sánchez-Monedero, P. A. Gutiérrez, and C. Hervás-Martínez, "Machine learning methods for binary and multiclass classification of melanoma thickness from dermoscopic images," *IEEE Transactions on Medical Imaging*, no. Accepted, 2015.

[10] M. Amouroux and W. Blondel, "Non-invasive determination of Breslow index," in *Current management of malignant melanoma*, M. Y. Cao, Ed. InTech, 2011, pp. 29–44. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00626482

[11] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: Survey and experimental study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, Jan 2016.

[12] M. Pérez-Ortiz, M. Cruz-Ramírez, M. Ayllón-Terán, N. Heaton, R. Ciria, and C. Hervás-Martínez, "An organ allocation system for liver transplantation based on ordinal regression," *Applied Soft Computing*, vol. 14, Part A, pp. 88 – 98, 2014.

[13] M. Pérez-Ortiz, P. Gutiérrez, C. Hervás-Martínez, and X. Yao, "Graph-based approaches for over-sampling in the context of ordinal regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1233–1245, May 2015.

[14] A. Breslow, "Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma." *Annals of Surgery*, vol. 172, no. 5, pp. 902–908, 1970.

[15] M. Stante, V. De Giorgi, P. Cappugi, B. Giannotti, and P. Carli, "Non-invasive analysis of melanoma thickness by means of dermoscopy: A retrospective study," *Melanoma Research*, vol. 11, no. 2, pp. 147–152, 2001.

[16] M. Lens, P. Nathan, and V. Bataille, "Excision margins for primary cutaneous melanoma: Updated pooled analysis of randomized controlled trials," *Archives of Surgery*, vol. 142, no. 9, pp. 885–891, 2007.

[17] M. Brady and D. Coit, "Sentinel lymph node evaluation in melanoma," *Archives of Dermatology*, vol. 133, no. 8, pp. 1014–1020, 1997.

[18] G. Argenziano, H. Soyer, and et al., *Interactive atlas of dermoscopy.* Milan: EDRA-Medical Publishing and New Media, 2000.

[19] A. Sáez, C. Serrano, and B. Acha, "Model-based classification methods of global patterns in dermoscopic images," *IEEE Transactions on Medical Imaging*, vol. 33, no. 5, pp. 1137–1147, 2014.

[20] A. Sáez, C. S. Mendoza, B. Acha, and C. Serrano, "Development and evaluation of perceptually adapted colour gradients." *IET Image Processing*, vol. 7, no. 4, p. 355 – 363, 2013.

[21] A. J. Smola, B. Schölkopf, and K.-R. Müller, “The connection between regularization operators and support vector kernels,” *Neural Networks*, vol. 11, no. 4, pp. 637–649, Jun. 1998.

[22] A. Bono, S. Tomatis, C. Bartoli, G. Tragni, G. Radaelli, A. Maurichi, and R. Marchesini, “The abcd system of melanoma detection: A spectrophotometric analysis of the asymmetry, border, color, and dimension,” *Cancer*, vol. 85, no. 1, pp. 72–77, 1999.

[23] H. Soyer, g. argenziano, R. Hofmann-Wellenhof, and R. Johr, *Color Atlas of Melanocytic Lesions of the Skin*. Springer Berlin Heidelberg, 2010.

[24] K. Weismann and H. F. Lorentzen, “Dermoscopic color perspective,” *Archives of Dermatology*, vol. 142, no. 9, p. 1250, 2006.

[25] S. Seidenari, G. Pellacani, and C. Grana, “Computer description of colours in dermoscopic melanocytic lesion images reproducing clinical assessment,” *British Journal of Dermatology*, vol. 149, no. 3, pp. 523–529, 2003.

[26] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, and M. Delfino, “Clinical and dermatoscopic criteria for the preoperative evaluation of cutaneous melanoma thickness,” *Journal of the American Academy of Dermatology*, vol. 40, no. 1, pp. 61–68, 1999.

[27] H. Lorentzen, K. Weismann, and F. Grønhøj Larsen, “Dermatoscopic prediction of melanoma thickness using latent trait analysis and likelihood ratios,” *Acta Dermato-Venereologica*, vol. 81, no. 1, pp. 38–41, 2001.

[28] V. da Silva, J. Ikino, M. Sens, D. Nunes, and G. Di Giunta, “Dermoscopic features of thin melanomas: A comparative study of melanoma in situ and invasive melanomas smaller than or equal to 1mm [características dermatoscópicas de melanomas finos: Estudo comparativo entre melanomas in situ e melanomas invasivos menores ou iguais a 1mm],” *Anais Brasileiros de Dermatologia*, vol. 88, no. 5, pp. 712–717, 2013.

[29] N. Otsu, “Threshold selection method from gray-level histograms.” *IEEE Trans Syst Man Cybern*, vol. SMC-9, no. 1, pp. 62–66, 1979, cited By 10522.

[30] M. Sadeghi, M. Razmara, T. Lee, and M. Atkins, “A novel method for detection of pigment network in dermoscopic images using graphs,” *Computerized Medical Imaging and Graphics*, vol. 35, no. 2, pp. 137–143, 2011.

[31] R. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. smc 3, no. 6, pp. 610–621, 1973.

[32] B. Boser, I. Guyon, and V. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, D. Haussler, Ed. Pittsburgh, PA: ACM Press, 1992, pp. 144–152.

[33] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[34] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multi-class support vector machines,” *IEEE Transaction on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[35] H. Wu, H. Lu, and S. Ma, “A practical svm-based algorithm for ordinal regression in image retrieval,” in *Proceedings of the eleventh ACM international conference on Multimedia (Multimedia2003)*, 2003, pp. 612–621.

[36] E. L. Allwein, R. E. Schapire, and Y. Singer, “Reducing multiclass to binary: a unifying approach for margin classifiers,” *Journal of Machine Learning Research*, vol. 1, pp. 113–141, Sep. 2001.

[37] F. Fernández-Navarro, C. Hervás-Martínez, and P. A. Gutiérrez, “A dynamic over-sampling procedure based on sensitivity for multi-class problems,” *Pattern Recognition*, vol. 44, p. 1821–1833, 2011.

[38] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, May 2011.

[39] W. Chu and S. S. Keerthi, “Support Vector Ordinal Regression,” *Neural Computation*, vol. 19, no. 3, pp. 792–815, 2007.

[40] E. Frank and M. Hall, “A simple approach to ordinal classification,” in *Proc. of the 12th Eur. Conf. on Machine Learning*, 2001, pp. 145–156.

[41] W. Waegeman and L. Boullart, “An ensemble of weighted support vector machines for ordinal regression,” *International Journal of Computer Systems Science and Engineering*, vol. 3, no. 1, pp. 1–7, 2009.

[42] M. Kubat and S. Matwin, “Addressing the curse of imbalanced training sets: one-sided selection,” in *Proceedings of the 14th International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 179–186.

[43] S. Baccianella, A. Esuli, and F. Sebastiani, “Evaluation measures for ordinal regression,” in *Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications (ISDA ’09)*, Pisa, Italy.

[44] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.