Applied Soft Computing 64 (2018) 341–355

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Partial order label decomposition approaches for melanoma diagnosis

Javier Sánchez-Monedero [a,*], María Pérez-Ortiz [a], Aurora Sáez [b], Pedro Antonio Gutiérrez [c], César Hervás-Martínez [c]

[a] Department of Quantitative Methods, Universidad Loyola Andalucía, 14004 Córdoba, Spain
[b] Signal Theory and Communications Department, University of Seville, 41092 Seville, Spain
[c] Department of Computer Science and Numerical Analysis, University of Córdoba, 14071 Córdoba, Spain

ARTICLE INFO

Article history:
Received 19 December 2016
Received in revised form 25 October 2017
Accepted 27 November 2017
Available online 14 December 2017

Keywords:
Melanoma
Computer vision
Machine learning
Ordinal classification
Partial order
Skin cancer

ABSTRACT

Melanoma is a type of cancer that develops from the pigment-containing cells known as melanocytes. Usually occurring on the skin, early detection and diagnosis is strongly related to survival rates. Melanoma recognition is a challenging task that nowadays is performed by well trained dermatologists who may produce varying diagnosis due to the task complexity. This motivates the development of automated diagnosis tools, in spite of the inherent difficulties (intra-class variation, visual similarity between melanoma and non-melanoma lesions, among others). In the present work, we propose a system combining image analysis and machine learning to detect melanoma presence and severity. The severity is assessed in terms of melanoma thickness, which is measured by the Breslow index. Previous works mainly focus on the binary problem of detecting the presence of the melanoma. However, the system proposed in this paper goes a step further by also considering the stage of the lesion in the classification task. To do so, we extract 100 features that consider the shape, colour, pigment network and texture of the benign and malignant lesions. The problem is tackled as a five-class classification problem, where the first class represents benign lesions, and the remaining four classes represent the different stages of the melanoma (via the Breslow index). Based on the problem definition, we identify the learning setting as a partial order problem, in which the patterns belonging to the different melanoma stages present an order relationship, but where there is no order arrangement with respect to the benign lesions. Under this assumption about the class topology, we design several proposals to exploit this structure and improve data preprocessing. In this sense, we experimentally demonstrate that those proposals exploiting the partial order assumption achieve better performance than 12 baseline nominal and ordinal classifiers (including a deep learning model) which do not consider this partial order. To deal with class imbalance, we additionally propose specific over-sampling techniques that consider the structure of the problem for the creation of synthetic patterns. The experimental study is carried out with clinician-curated images from the Interactive Atlas of Dermoscopy, which eases reproducibility of experiments. Concerning the results obtained, in spite of having augmented the complexity of the classification problem with more classes, the performance of our proposals in the binary problem is similar to the one reported in the literature.

1. Introduction

Melanoma is a type of cancer that arises from the pigment-containing cells known as melanocytes. The most common type, the cutaneous melanoma, occurs on the skin. In Europe, approximately 100,000 cases are yearly diagnosed, with a death ratio around 13% [1]. Patient prognosis depends directly on tumour thickness, where mortality can be reduced to a great extent by early detection and diagnosis [2].

To improve survival rates, melanoma must be detected before the tumour has penetrated the epidermis (i.e. before the thickness is higher than 0.76 mm). In the case of early detection, the five-year survival rate is about 99%, otherwise dropping to 15% for patients with advanced disease [3]. The current detection process consists on a visual inspection by trained professionals using a dermatoscope, and the prognosis is evaluated measuring the depth of the melanoma by means of a biopsy. Dermatologists perform this manual visual inspection from dermoscopy images, but this process is time-consuming and error-prone, and it can lead to widely vary-

* Corresponding author at: Department of Quantitative Methods, Universidad Loyola Andalucía, Calle Escritor Castilla Aguayo, 4, 14004 Córdoba, Spain.
E-mail address: jsanchez@uloyola.es (J. Sánchez-Monedero).

https://doi.org/10.1016/j.asoc.2017.11.042
1568-4946/

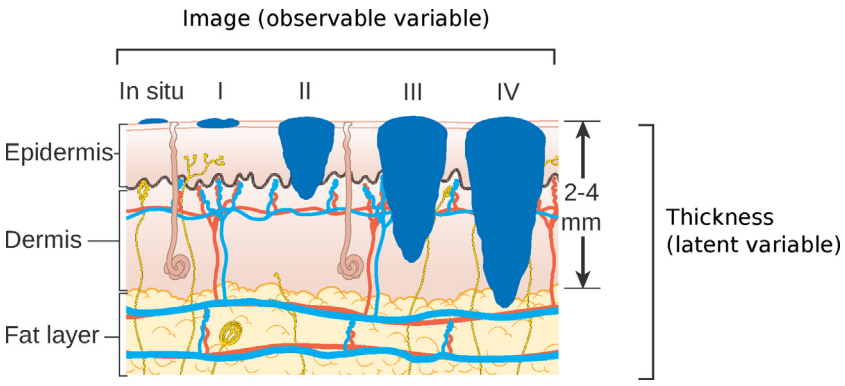


Fig. 1. Graphical representation of the different stages of melanoma, where both the observable data (dermoscopic image) and the unobservable or latent variable (thickness of the tumour) can be analysed.
Image credit: Cancer Research UK/Wikimedia Commons.

ing diagnosis. This motivates automated diagnosed methods [4,5]. Recent works propose new tools to aid or to improve this process [3], mainly based on dermoscopic image analysis. Although there are different lines of undergoing research (e.g. those based on skin temperature variations in the lesion), image analysis methods present the advantage of being cheaper and relatively easy to combine with existing detection procedures.

In the last years, computerised dermoscopy image analysis systems have been proposed to assist pigmented lesions diagnosis [6]. The majority of these works focus on the distinction of melanomas from benign lesions [7–9]. However, a finer grain classification is required for appropriate prognosis. The scarcity of studies on this topic and its inherent difficulty makes it a promising line for research. The first work in that line is the characterisation of two types of melanoma based on their thickness [10]. This study uses 49 features related to colour, geometry and texture, extracted from a private database of 141 images obtained with a company proprietary hardware system. Moreover, a recent study [11] focuses on the classification of three degrees of thickness for melanomas, but it excludes their distinction from benign lesions, which is crucial for constructing a complete detection tool.

In this paper, we propose to simultaneously address the problem of melanoma detection and thickness estimation within a five-class classification problem. To do so, we combine image analysis and machine learning procedures. Now, we summarise the feature extraction process and describe the dataset characteristics which motivate the development of specific machine learning methods. Particularly, the challenging issues found in this problem are: (1) the structure and topology of the classes and (2) the imbalanced nature of the classes that can bias classification performance in favour of majority classes.

Concerning the image analysis, we propose a set of 100 input features to describe images. The extracted features correspond to visual characteristics based on dermatologists clinical findings (see Section 4). Melanoma cases are distinguished from non-melanoma ones using the ABCD method, based on four clinical characteristics that describe a malignant melanoma: asymmetry (A), border irregularity (B), colour variegation (C) and differential structures (D). The rest of the features selected are related to melanoma thickness estimation, and analogously they are based on clinical criteria with respect to visual characteristics present in dermoscopic images [12,11].

When attempting to estimate the severity of a melanoma, it can be seen that the classes are imbued with order information. The Breslow index is modelled as an unobservable latent variable that represents the thickness of the tumour using the dermoscopic image (independent variable). Such latent variable can only be directly observed when performing a biopsy, in which case the actual tumour thickness can be measured and used to validate the prediction. Since the different Breslow index levels correspond to thresholds of the thickness, the corresponding class labels show an order relationship, in such a way that stage II melanomas are thicker than stage I ones, stage III implies a thicker lesion than stage II, and so on. Fig. 1 shows the different stages of a melanoma and analyses the observed and latent variable concepts in the frame of this problem. Please note that in this work we group stages III and IV due to the fact that they have similar clinical properties. This type of problems are known as ordinal classification problems, also referred to as ordinal regression [13]. They differ from nominal (standard) classification problems in the fact that there is an order arrangement between the categories, and they are different from regression because the distance between the values of the dependent variable (the class) is generally unknown. The most common situation in ordinal regression is that the categories come from the discretisation of a latent variable [13], which is exactly the case of the different stages of melanoma. Ordinal methods exploit the ordered nature of the classes to improve learners at the same time that penalise the magnitude of the classification errors (for example, in our case, misclassifying a stage 0 melanoma with a stage I should not be considered the same than confusing it with a stage III melanoma). Ordinal classification has been successfully applied to different areas such as Alzheimer' progression estimation [14] or sovereign ratings [15], among others. Section 2 provides some basic background on ordinal classification.

However, although this order is clear for the different stages of the melanoma (since they reflect different levels of thickness), it cannot be assumed for the benign lesion class. In this sense, the problem can be considered as a partially ordered classification task, for which we propose several machine learning strategies to exploit this characteristic. Fig. 2 illustrates the concept of partial order in a two-dimensional dataset, where it can be seen that $\mathcal{C}_1$ does not follow an order with respect to the rest of classes, while the rest are ordered in the input space ($\mathcal{C}_2$ is closer to $\mathcal{C}_3$ than to $\mathcal{C}_4$, and so on). Note that this structure can be found in very different classification problems, e.g. in medicine (non-disease vs. disease grades). In this case, it can be seen that a unique linear projection (which takes the order of the classes into account) is not feasible, while two projections (one for tackling the binary problem and other for the ordinal one) could separate the data satisfactorily. Ordinal problems with specific data structures (e.g. partial order problems or circular ordinal regression [16,17]) or other more complex label structures (such as multiple output ordinal regression, graded multilabel classification [18] or label ranking [19]) are recently receiving attention from the machine learning community.

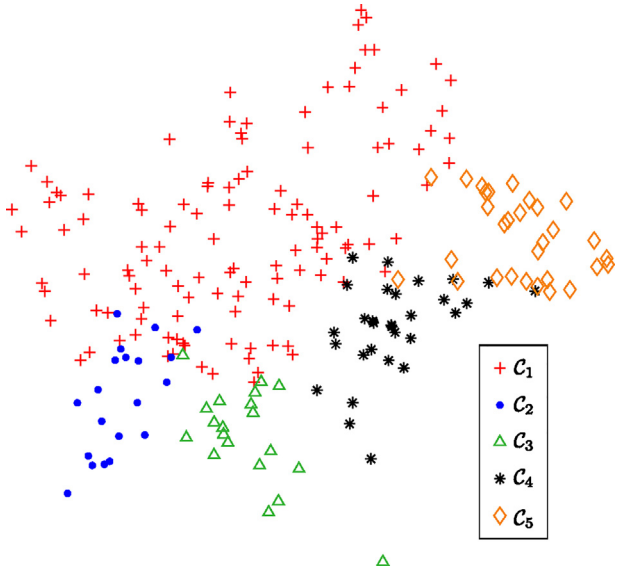


Fig. 2. Example of a partially ordered dataset.

Table 1
Characteristics of the classes in the problem (benign lesions against different stages of melanoma): name of the class, depth of the melanoma lesion, ranking for the class (q), number of patterns per class (N_q) and multiclass imbalance ratio (IR). The best and second-best results are in bold face and italics, respectively.

Class	Depth	$\mathcal{C}_q$	N_q	IR
Non-melanoma	–	1	313	0.159
Stage 0	In situ	2	64	**1.556**
Stage I	<0.76 mm	3	102	0.902
Stage II	0.76–1.50 mm	4	54	**1.881**
Stage III	>1.50 mm	5	29	**3.676**
Total number of patterns:			562	

On the other hand, the dataset used in the present work is also characterised by a skewed class distribution (see imbalance ratio in Table 1). In this sense, we present two approaches based on label decompositions to deal with the partial order of the labels and the imbalance nature of the data. The proposed methods comprise: (1) a hierarchical model composed of a binary model to distinguish non-melanoma from melanomas and an ordinal model to refine the classification of the stage of the melanoma; (2) a cascade binary utility ordinal model [20], which has been shown to obtain good results for problems with these characteristics (partially ordered and imbalanced); (3) additionally, we consider data over-sampling techniques to alleviate the imbalance nature of the dataset, specifically the ordinal over-sampling techniques presented in [21].

The findings of the present work are the following:

- First, we propose a set of 100 features to describe dermatoscopic images. Even with baseline methods, the extracted features are proved to be suitable for simultaneously detecting melanomas and predicting the lesion stage.
- Second, we propose to tackle the classification problem in a partial order framework to obtain models that better fit to the data characteristics, i.e. its topology and the imbalanced class distribution. We experimentally demonstrate that both classification and over-sampling methods that consider the partial order nature of the problem present very promising and competitive performance (considering four classification metrics) with respect to nominal and full ordinal methods.
- Third, the proposed features, together with the use of the proposed partial order methods, allow linear probabilistic classifiers to achieve the best performance with respect to a variety of non-linear models. This is, the best model can be examined to evaluate the relevance of each feature and therefore contribute to clinical knowledge.

This paper is a significant extension of a previous work conference [22], which presented an initial proposal to the partial order approach for melanoma detection and stage classification. The present work includes the following new contributions: 14 new image features are added, the problem of partial order is studied more deeply (both proposing new models and methods from the classification and over-sampling points of view), experimental comparisons are strengthen with more methods and configurations, the performance is improved simultaneously considering several classification metrics, we evaluate the impact of feature selection in the models performance, and, finally, the best performing method is analysed to establish the relevance of the different features, which reveals that the new selected features contribute significantly to the classification performance.

The rest of the paper is organised as follows. Section 2 provides some background on ordinal classification and motivates the development of partial ordinal methods. Section 3 presents the clinical problem and some characteristics of the dataset. Section 4 introduces the set of features selected to describe the images. Section 5 presents some previous notions and describes the proposed decomposition methods as well as the over-sampling methods. Section 6 shows the experiments performed and analyses the results. Finally, Section 7 outlines some conclusions and future work.

2. Ordinal classification and partial order problems

This section briefly presents the ordinal classification framework and the family of classifiers that can be used in this context. The limitations of ordinal classification for partial order problems are also analysed. In addition, the basic notation used in the paper is introduced. For further information about ordinal classification, we refer the reader to [13], where a taxonomy of the different proposals is presented.

Ordinal classification is a type of classification problem in which there is an order relationship between the categories to predict, $\mathcal{C}_q$, $q \in \{1, \ldots, Q\}$. Contrary to standard regression, we cannot assume a distance between these categories. 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 qth 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 ($\mathcal{C}_1 \prec \mathcal{C}_2 \prec \ldots \prec \mathcal{C}_Q$, where $\prec$ denotes this order information).

According to Hühn and Hüllermeier [23], this order in the label space is also present in the input space $\mathcal{X}$, an assumption that can be used to improve the classifier. As mentioned in the introduction, the performance metric has to be sensitive to the label order and penalise the magnitude of the errors. Different metrics have been proposed as alternative measures to the well-known accuracy. For a review of those metrics we refer to the study in [24]. We describe the metrics used to evaluate melanoma classification in Section 6.2.

Different ordinal classification methods have been recently proposed. Most of them belong to the category of threshold models. These methods assume that ordinal categories come from the discretisation of a continuous latent variable, $\mathcal{Z} \subseteq \mathbb{R}$, and try to model this discretised variable. In this sense, these methods seek a projection in which the samples are ordered according to their class rank and a set of thresholds that divides the projection into consecutive

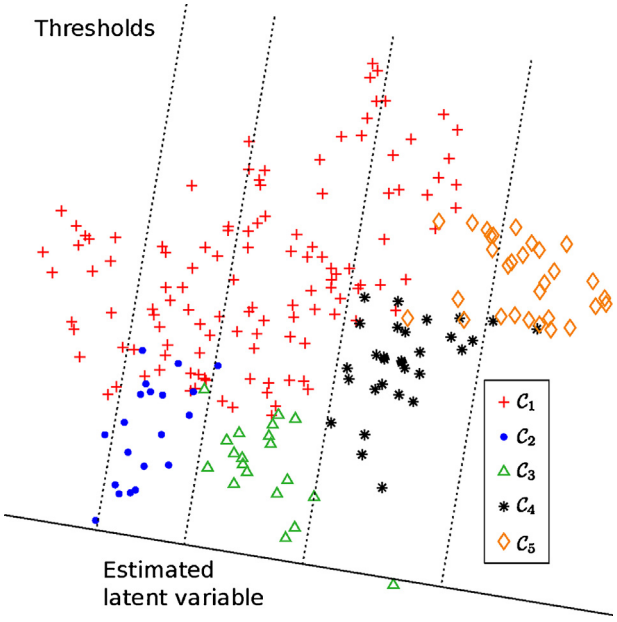


Fig. 3. Example of an ordinal threshold model which has been trained considering the data in Fig. 2. It can be seen that because of the structure of the data the ordinal model is not able of obtaining a suitable solution.

intervals representing ordinal categories. We review now some of the proposed approaches in the literature.

The first threshold model proposal was the Proportional Odds Model (POM), which extends logistic regression to the ordinal case [25]. Support vector machines (SVM) have also been adapted to the threshold model structure. In [26], two new support vector approaches are proposed for ordinal regression. In this case, multiple thresholds are optimised in order to define parallel discriminant hyperplanes for the ordinal classes. In [27], the classification problem is transformed into a regression problem by directly performing a projection from the input space to a one-dimensional space, where the values of the projections are estimated based on the pairwise class distances.

Other alternatives decompose the ordinal problem into a set of binary classification problems. For instance, some previous works [28,29] transform the problem into a set of nested binary classification problems to train several binary models, where the final prediction is obtained by a combination of the binary predictions. A further step in this direction is the extended binary classification framework, which transforms ordinal regression into a binary classification with additional features, in the context of SVMs [30].

In this paper, however, we approach a problem of partial ordering classification. In this type of problems, we have a set of classes that follow a given order (e.g. $\mathcal{C}_2 \prec \mathcal{C}_3 \prec \ldots \prec \mathcal{C}_Q$) and a class or set of classes (e.g. $\mathcal{C}_1$) for which this order cannot be assumed (see Fig. 2). Fig. 3 presents the results of training a standard threshold model (POM), considering the data of Fig. 2. A standard ordinal classifier assumes an order relation for all classes, and, as can been in Fig. 2, this produces poor results. On the other hand, Fig. 4 presents an example of the hierarchical model proposed in Section 5.1, which presents a better performance. This example motivates the development of specific models to address the cases that show this class topology.

Concerning performance evaluation, we believe that, in this case, the same ordinal misclassification errors should hold, since it should be more penalised to misclassify a benign lesion with a stage III melanoma than with a stage I or 'in situ' melanoma. However, there may be specific problems in which other costs may be needed.

3. Data and Breslow index description

As stated, tumour depth is inversely correlated with survival rate. The reason is that thick tumours access lymph capillaries, which is the most common way for cancer to spread. If the melanoma is confined to the epidermis, it is referred to as 'in situ' melanoma, and it is removable by surgery. However, as the cancerous cells propagate to the deepest layer of the skin (the dermis), the melanoma is known as invasive, and the survival rate decreases with the depth of the invasion.

The Breslow index [31] is a method for prognosis of patient survival that measures melanoma depth by a pathological examination after an incisional or excisional biopsy of the lesion [12]. It consists on a vertical measurement in millimetres from the top of the granular layer of the epidermis to its deepest part within the dermis. Moreover, it is the main parameter used to establish the width of the surgical margin excision [32,33], as well as to decide whether to perform sentinel lymph node biopsy (SNB) [32,34] (SNB is a surgical procedure to determine if cancer has spread to the lymphatic system). Therefore, measuring melanoma thickness before surgical excision is crucial in order to assess the risk of progression, and consequently to ensure adequate excision margins avoiding a more complicated operation and SNB.

In this paper, we use 562 images from the Interactive Atlas of Dermoscopy [35], a multimedia project for medical education with pigmented skin lesions images in which all lesions were biopsied and diagnosed histopathologically. As introduced, the images have been classified in five classes: non-melanoma (i.e. benign lesions) and four stages of melanoma depth. The characteristics of these classes can be seen in Table 1, where the imbalanced ratio per class (IR) is also included. The multiclass IR is computed using the formulation in [21]:

$$IR = \frac{1}{Q}\sum_{q=1}^{Q} IR_q, \tag{1}$$

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

$$IR_q = \frac{\sum_{j \neq q} N_j}{Q \cdot N_q}. \tag{2}$$

Classes with an IR higher than 1 are highlighted in bold face, as these are the ones over-sampled in the experiments. All the images have a resolution of 768×512 pixels and have been segmented using the automatic segmentation algorithm proposed in [36], in which an edge based level-set technique is applied together with a perceptually adapted colour gradient [37]. Fig. 5 presents two examples of segmented melanomas.

4. Feature extraction

The feature extraction process 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 properties seen in dermoscopic images and melanoma thickness. A total of 100 descriptors (x_1–x_{100}) based on shape, colour and texture have been extracted. Regarding the ABCD method, asymmetry and border irregularity are characterised by shape features, colour variegation by a feature set that contains the number of colours present in a lesion and differential structures by texture features, especially, by those based on a Markov random field model, that allows to identify different dermoscopic structures, as proposed in [36]).

Some of the extracted features are based on several previous works that the reader can check for more details [11,38,22]. More-

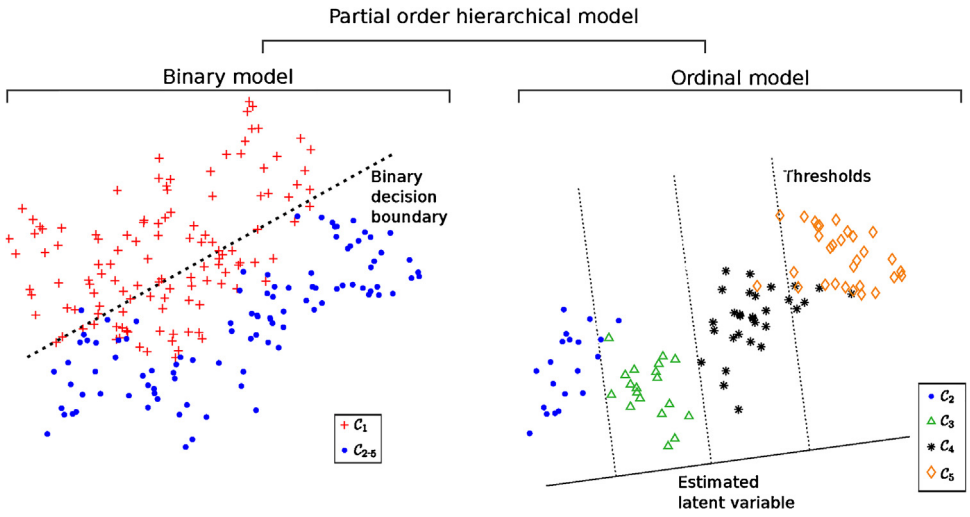


Fig. 4. Example of a partial order hierarchical model fitted to the partial order problem in Fig. 2. The method divides the problem in two subproblems: fitting a binary model to separate the non-ordered class and an ordinal model to order the rest of classes. This figure illustrates the idea of the hierarchical model presented in Section 5.1.

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

Fig. 5. Examples of segmented melanomas.

over, in the current study, we include 14 additional shape and colour features.

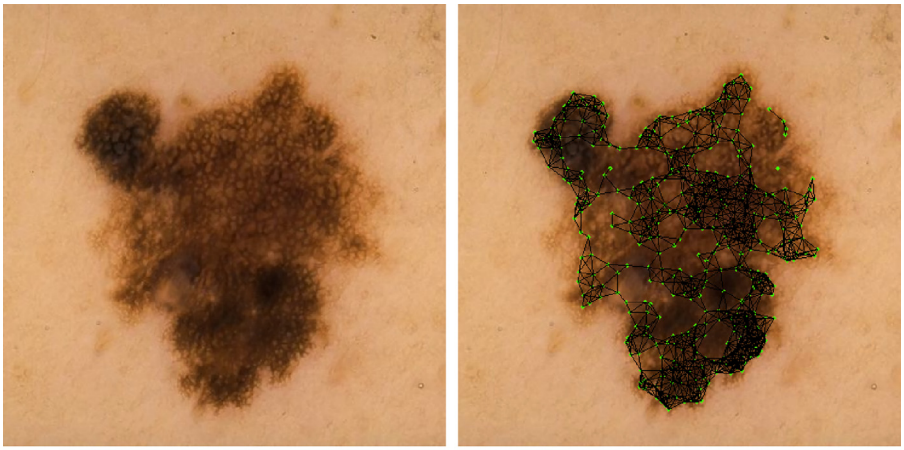
4.1. Shape features

As previously stated, 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) [6], 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) [39], the eccentricity (a measure of elongation) (x_4) [8] and length of major and minor axis of the lesion normalised with respect to the equivalent diameter (diameter of a circle with the same area as the lesion) (x_5, x_6), and difference between these two values (x_7).

In order to evaluate the lesion asymmetry, first, the major axis orientation of the lesion has been computed, and secondly, it has been rotated clockwise the same number of degrees 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_8). The same procedure has been performed for the y-axis to obtain vertical asymmetry (x_9). If the process is repeated taking into account the percentage of overlapping pixels assigned to the same colour (see Section 4.2 for the colour assignation), we can compute the colour vertical asymmetry (x_{10}) and the colour horizontal asymmetry (x_{11}).

4.2. Colour features

Colour features are one of the most determinant features for estimating melanoma depth. Different dermoscopic structures, which have been found discriminative for melanoma thickness, are associated with different colours. We have extracted features related to the six colours present in the pigmented lesions: black, dark brown, light brown, blue-grey, red and white [40]. The presence of these colours depend on the depth of the melanoma: black melanin appears when it is located in the stratum corneum and upper epidermis; brown is associated with a deeper location in the epidermis; grey and blue are related to its presence in the dermis; red is associated with dilation of blood vessels; and white with



(a) Original image. (b) Pigment network detection.

Fig. 6. Illustration of pigment network detection.

scaring and/or regression. To describe these colours, we segment each lesion into their constituting colours by a similar approach to that proposed by Seidenari et al. [41]. We have developed a colour palette formed by 144 patches that present unequivocally one of the six possible colours. This palette is used to extract the colour regions of the lesions from the patches according to a nearest neighbour approach. Each pixel of the image has been assigned to the colour patch that minimises its Euclidean distance in the CIE $L^*a^*b^*$ colour space. From this colour identification, we have extracted six descriptors (x_{12}–x_{17}) that represent the percentage of the lesion area classified as these colours, one (x_{18}) that represents the number of colours that each lesion presents (colour criterion of ABCD), and 36 additional statistical descriptors of the colours (mean, standard deviation, kurtosis, skewness, entropy, and average of local standard deviation of each colour, x_{19}–x_{54}).

4.3. Pigment network features

Pigment network is a dermoscopic structure, referred by many authors as one of the most discriminative features for melanoma thickness [32,42,43], being inversely correlated with melanoma depth [35]. A pigment network is a regular grid of brownish lines over a diffuse light-brown background [35]. We have identified this structure searching for the network 'holes' by applying a filtering and thresholding step using the Otsu's method [44]. Finally, we have considered the two conditions relative to area size and colour proposed in the work of Sadeghi et al. [45] to remove those wrongly detected areas. The features extracted from this detection are network density ratio (x_{55}), number of nodes (x_{56}) and number of links or edges (x_{57}). Fig. 6 shows an example of pigment network detection.

4.4. Texture features

Other dermoscopic structures have been found to be correlated with the depth of melanoma such as vascular patterns [42,32], blue-grey veil [42,32], white scar-like areas [43] and dots or globules. These are usually associated with texture features. For instance, vascular patterns are associated to the presence of a vascular vessel with a line shape, and grey-blue areas and white scar-like areas are found as homogeneous areas [43]. To capture properties of different structures, we have extracted three sets of texture features from three different approaches: 19 features from the grey level co-occurrence matrix (GLCM) [46] (x_{58}–x_{76}), 18 features based on a Markov random field (MRF) model [36] (x_{77}–x_{94}) and 6 features from local binary pattern (LBP) histograms (x_{95}–x_{100}).

5. Decomposition approaches for partial order classification problems

This section presents the preliminary concepts and classification strategies proposed in this paper to deal with partially ordered classification problems, which, in addition, present class imbalance. More specifically, two different decomposition methods are derived for this type of learning problems.

Decomposition methods have been one of the first proposals both for multiclass classification and for ordinal classification, because of their simplicity and their good performance (given that they are, essentially, ensemble methods). Concerning multiclass classification, most common approaches are the one-against-one and the one-against-all proposals. With regard to ordinal classification, there exist different strategies based on decomposition methods in the literature: ordered partitions [28] (where the classes are joined taking the order of the classes into account), ordinal one-against-all partitions [47] (where each class is separated from the previous and following classes), one-against-next [13] (or one-against-previous, where each class is separated from the previous or the following class or classes) or the cascade utility model [20] (where each class is separated from the remaining ones, taking the scale of order into account).

The partial order problem is similar to the standard ordinal classification one, where there exist different misclassification costs and where the order of the classes has to be taken into account for constructing a fair and robust classifier. However, the order of the classes is not total but partial, in the sense that not all the classes in the problem are ordered. This is a common setting in biomedicine applications, such as the one presented in this work. Despite the number of applications, up to our best knowledge, this setting has not been specifically tackled before in the ordinal classification literature. In this sense, this paper aims to establish the difference in performance between standard ordinal classification strategies and partially ordered ones in a problem that we hypothesise is partially ordered. Given also the imbalanced nature of the dataset, we consider two different classification approaches for partial order problems (a hierarchical decomposition and a cascade binary utility model) and a reformulation of an over-sampling technique where the order of the classes can also be included partially or completely, used as a preprocessing step.

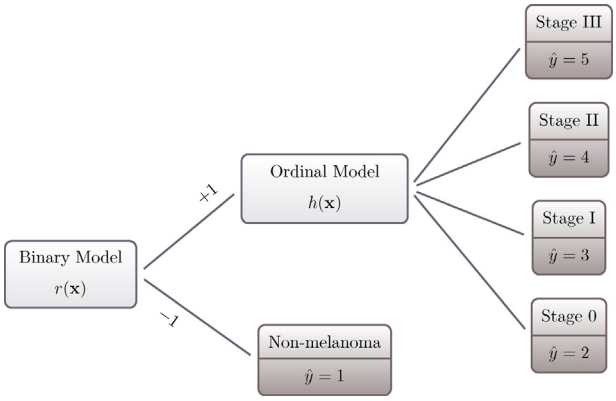


Fig. 7. Hierarchical classifier prediction process.

For melanoma severity classification, we hypothesise that this order structure only applies to four of the five classes of the problem, in such a way that $\mathcal{C}_2 \prec \ldots \prec \mathcal{C}_5$ follow a natural order between them but do not present an ordering with respect to $\mathcal{C}_1$ (see Table 1).

5.1. Hierarchical decomposition

As stated before, it cannot be assumed that the benign lesion class presents an order relationship with the rest of the classes, although the same misclassification costs than in ordinal regression should hold for this case. Moreover, given that the features extracted to distinguish between benign and malignant lesions and the depth of melanoma are different, the relevance of the features could also be different for each model. These reasons, together with the imbalance nature of the problem, motivate the use of hierarchical classification models. To do so, we propose to first learn a binary model to distinguish between benign lesions and melanoma. Secondly, we train an ordinal classification model to determine the stage of the melanoma. Fig. 7 presents the hierarchical model composed of a binary and an ordinal classifier. Note also that the imbalanced nature of the data is alleviated by this approach since the minority classes are specifically separated from the majority one.

In this work, we have implemented two hierarchical models. The first is based on Logistic Regression (LR) for the binary model and on the Proportional Odds Model for the ordinal one [25] (POM adapts the standard LR to the ordinal case). The second is a kernel version where Support Vector Machines are used for the binary model and the reformulation of SVM for ordinal regression with implicit constraints (SVORIM) [26] is used as the ordinal one.

For binary classification, LR and SVM solve the following unconstrained optimisation problem with different loss functions $\xi(\mathbf{w};\mathbf{x}_i, y_i)$ [48]:

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi(\mathbf{w};\mathbf{x}_i, y_i), \tag{3}$$

where $C > 0$ is a cost parameter. The common loss function for SVM, that is referred to as L1-SVM, is $\max(1 - y_i, \mathbf{w}^T\mathbf{x}_i, 0)$. In the case of LR the loss function $\log(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i})$ is derived from a probabilistic model. In our case, we use L2 regularised logistic regression. Though the standard LR does not include the C penalty parameter, implementations such as `LIBLINEAR` [48] include this cost. From an experimental point of view, in this problem, the results improved when this parameter was optimised. Finally, the well known LR probabilistic prediction model for the binary case gives the pattern probability of belonging to the positive class:

$$P(y|\mathbf{x}_i, \mathbf{w}) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}_i)}}. \tag{4}$$

The previous model assigns a class to a pattern using the following function:

$$\hat{y} = r(\mathbf{x}) = \begin{cases} +1, & \text{if } P(y|\mathbf{x}_i, \mathbf{w}) \geq 0.5, \\ -1, & \text{if } P(y|\mathbf{x}_i, \mathbf{w}) < 0.5, \end{cases} \tag{5}$$

where +1 denotes the positive class (melanoma presence) and −1 the negative class (melanoma absence).

In the case of ordinal classification, the Proportional Odds Model (POM) [25] extends binary logistic regression to ordinal regression. It belongs to a family of methods known as threshold models [13]. These models assume that an unobserved continuous variable underlies the ordinal response variable, so that they estimate:

- A function $g(\mathbf{x})$ that project the data into real-valued outcomes.
- A set of thresholds $\mathbf{b} = (b_1, b_2, \ldots, b_{Q-1})$ to represent intervals in the range of $g(\mathbf{x})$, which must satisfy the constraints $b_1 \leq b_2 \leq \cdots \leq b_{Q-1}$.

In this way, patterns $\mathbf{x}$ are projected to the latent space $\mathcal{Z}$ and then classified depending on the set of thresholds.

The POM is a member of a wider family of models referred to as Cumulative Link Models (CLMs) [49]. CLMs predict probabilities of group of contiguous categories considering the ordinal scale so that cumulative probabilities $P(y \preceq \mathcal{C}_j|\mathbf{x})$ are estimated, which can be directly related to standard probabilities:

$$P(y \preceq \mathcal{C}_q|\mathbf{x}) = P(y = \mathcal{C}_1|\mathbf{x}) + \cdots + P(y = \mathcal{C}_q|\mathbf{x}),$$
$$P(y = \mathcal{C}_q|\mathbf{x}) = P(y \preceq \mathcal{C}_q|\mathbf{x}) - P(y \preceq \mathcal{C}_{q-1}|\mathbf{x}),$$

with $q \in \{1, \ldots, Q\}$ and considering that $P(y \preceq \mathcal{C}_Q|\mathbf{x}) = 1$. Stochastic ordering of space $\mathcal{X}$ is satisfied by the following general model form [50]:

$$g^{-1}(P(y \preceq \mathcal{C}_q|\mathbf{x})) = b_q - \mathbf{w}^T\mathbf{x}, \quad 1 \leq q \leq Q, \tag{6}$$

where $g^{-1} : [0, 1] \to (-\infty, +\infty)$ is a monotonic function, typically referred to as the inverse link function, and b_q is the threshold defined for each class $\mathcal{C}_q$. As mentioned, this structure is associated to latent variable and threshold models, where $\mathbf{w}^T\mathbf{x}$ is a linear transformation.

Finally, the decision function to assign a pattern to a class is:

$$\hat{z} = h(\mathbf{x}) = \begin{cases} 2, & \text{if } g(\mathbf{x}) \leq b_1, \\ 3, & \text{if } b_1 < g(\mathbf{x}) \leq b_2, \\ \vdots & \\ Q, & \text{if } g(\mathbf{x}) > b_{Q-1}, \end{cases} \tag{7}$$

where $g : \mathcal{X} \to \mathcal{R}$ is the function that projects data space onto the one-dimensional latent space $\mathcal{Z}$. Note that in the context of this partial order model the lowest class number for the ordinal model is two.

The hierarchical model based on LR and POM is formalised as:

$$\hat{y} = f(\mathbf{x}) = \begin{cases} 1, & \text{if } r(\mathbf{x}) = -1, \\ h(\mathbf{x}), & \text{if } r(\mathbf{x}) = +1, \end{cases} \tag{8}$$

where $r(\mathbf{x})$ is the LR decision function defined at Eq. (5) and $h(\mathbf{x})$ is the POM classification rule defined at Eq. (7).

The training process consists of two steps: (1) solving the optimisation problem of Eq. (3) with patterns labelled as −1 (non-melanoma, $\mathcal{C}_1$) and +1 (rest of the classes); (2) training the

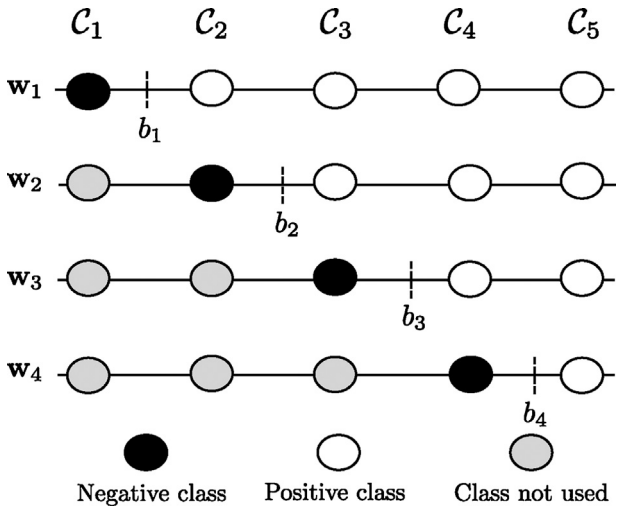


Fig. 8. Binary decompositions performed for a five-class problem, where $\mathbf{w}_i$ represents the ith projection and b_i the bias associated to that projection. White-shadowed shapes represent the negative class, black-shadowed ones the positive one and grey ones the classes ignored in each model. This is an ordinal decomposition since adjacent classes are grouped together (with the exception of the first class, which, by the problem definition, presents a different relationship with the rest of classes).

POM model of Eq. (6) with patterns belonging to classes in set $\{\mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_5\}$.

In the case of SVMs, a similar idea is applied, but we consider the standard binary *C*-SVC to build the binary classifier [51] and SVORIM for the ordinal one [26].

5.2. Cascade utility model

This section presents a modification of the binary decomposition method known as the cascade linear utility model [52]. This procedure considers $Q-1$ binary models, where each model D_i is comprised of a projection $\mathbf{w}_i$ and a threshold b_i. Model q separates class $\mathcal{C}_q$ from classes $\mathcal{C}_{q+1} \vee \ldots \vee \mathcal{C}_Q$ and only a portion of the classes are considered for the computation of each model. Fig. 8 graphically describes this decomposition. This methodology is also naturally well-suited for the problem considered, because it alleviates the imbalanced class distribution, and because it considers the partial order of the classes, not assuming any order constraint for $\mathcal{C}_1$ ($\mathcal{C}_1$ is only considered by the first model, which discriminates $\mathcal{C}_1$ from the rest of classes). This approach is also known as one-against-followers [13].

The training set for model or decision maker $D_q = \{\mathbf{w}_q, b_q\}$ is specified by $\{\mathbf{X}_{(i|i=q)}, \mathbf{X}_{(j|j>q)}\}$ (see Fig. 8). Therefore, a coding matrix $\mathbf{M}_{(Q-1\times Q)} = [M_{iq}]$, $i = 1, \ldots, Q-1$, $q = 1, \ldots, Q$ 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 corresponds to negative class patterns, the label +1 to patterns belonging to the positive class, and finally, the patterns associated with label 0 are excluded from the training process of that binary classifier. In this way, the approach considered is the same than in [20], but using a one-against-followers approach. A matrix of predictions can be obtained by means of a single multiclass model (e.g. using artificial neural networks) or by multiple models (training a binary classifier for each subproblem, as in this paper) [13]. Once the models have been trained, a set of $Q-1$ decision values $\mathbf{f}(\mathbf{x}) = f_1(\mathbf{x}), \ldots, f_{Q-1}(\mathbf{x})$ are obtained for pattern $\mathbf{x}$. For the prediction phase, two different approaches can be considered (both tested in the experiments of this paper): a hierarchical approach or an approach based on the Error-Correcting Output Codes framework (ECOC). The hierarchical approach is the most commonly used with the cascade binary utility model. In this case, $\mathbf{w}_1$ is used in the first place, and all the patterns that are not predicted as positive (i.e. $\mathcal{C}_1$) but rather as negative (i.e. belonging to the set $\{\mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_5\}$) are used for $\mathbf{w}_2$, and so on. In this sense, this approach emphasises more the first computed models, so that when these models fail, the final predictions are wrong without considering the rest of models. Concerning the ECOC framework [20], 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). Prediction is then accomplished by choosing the column of $\mathbf{M}$ closest to the set of decision values $\mathbf{f}(\mathbf{x})$. When the code contains a 0, this leads to an indifferent condition in the prediction phase. According to this, the final decision function is the following one:

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

where $\mathbf{M}_q$ is the qth 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 should correspond with the loss function used for deriving the binary classifier). For example, for the case of the 1-norm SVM paradigm (one of the base methodologies used in this paper), the hinge-loss function could be chosen:

$$\text{loss}(y_i, f(\mathbf{x}_i)) = \max(0, 1 - y_i(\mathbf{w}\cdot\Phi(\mathbf{x}_i) + b)).$$

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

$$q = arg\,\min_{q=1\ldots Q} \sum_{i=1}^{Q-1} \max(0, (1 - \mathbf{M}_{iq}\cdot f_i(\mathbf{x}))) \tag{9}$$

and for LR:

$$q = arg\,\min_{q=1\ldots Q} \sum_{i=1}^{Q-1} \log(1 + e^{-\mathbf{M}_{iq}\cdot f_i(\mathbf{x})}). \tag{10}$$

One of the main advantages of this methodology over a purely hierarchical approach is that all real values are used for prediction instead of binary predicted class values. Consequently, the model is provided with additional information which may be useful for improving its performance. Note that the decision values $\mathbf{f}(\mathbf{x})$ represent the distance to the threshold, which is a measure usually considered for estimating class probabilities.

5.3. Preprocessing data by ordinal pattern over-sampling

Imbalanced data arise naturally in ordinal classification problems. The reason is that there are classes that naturally present lower a priori probability (typically, extreme classes) [21], as occurs in the problem considered in this paper: there are significantly more patterns associated to benign lesions with respect to melanomas (specially when considering thicker ones, see Table 1). Because of this reason, we consider the application of a recently proposed method for class balancing in ordinal classification problems [21]. In contrast to other over-sampling techniques, 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

and therefore exploited. Moreover, new patterns are created in the borderline between adjacent classes, in order to smooth the ordinal nature of the dataset and prevent that minority classes are obviated in the classifier.

In this paper, we consider one of the proposals of [21], named as ordinal graph-based over-sampling via shortest paths using a probability function for the intra-class edges (OGO-SP). The classes over-sampled are the ones which present an imbalance ratio (see Eq. (1)) higher than a considered threshold (in our case, 1), and the number of synthetic new patterns is that needed to obtain an imbalance ratio lower than this threshold.

For more information about this procedure refer to [21]. As said before, our experiments consider both the total order and partial order approaches (only a subset of the classes $\mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$, and $\mathcal{C}_5$ follow this order). Since the over-sampling strategy exploits the order of the classes to resample data, this is an important consideration. The number of patterns needed to balance the distribution is, however, computed using all the data, so that both datasets present the same number of patterns. 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 $\mathcal{C}_1$ and $\mathcal{C}_3$ are lower than 1.

6. Experiments

This section covers the methods, performance metrics and experimental design used, together with an analysis of the results and models obtained. Source code of proposed methods is available in the website associated to the paper.[1] Experiments for all the methods but TensorFlow are performed using ORCA framework.[2]

6.1. Comparison methods

Different classifiers (nominal, ordinal and partial order proposals) are compared in this paper. The methods included are single model methods as well as binary approaches to multi-class/ordinal classification. More specifically, the methods tested are the following:

- Kernel Discriminant Learning for Ordinal Regression (KDLOR) [13], which extends the Kernel Discriminant Analysis (KDA) to ordinal classification using a rank constraint.
- Multinomial Logistic Regression (MLR), which applies the one-against-all scheme.
- The Proportional Odds Model (POM), which adapts standard logistic regression to the ordinal case. MLR and POM are implemented with the `mnrfit` function in Matlab.
- Regularised Multinomial Logistic Regression (RMLR), where the classification model is composed of several binary models using the one-against-all scheme (implemented in `LIBLINEAR` [48]).
- RED-SVM [13], which applies the reduction from cost-sensitive ordinal ranking to weighted binary classification framework to SVM.
- Support Vector Classifier using the one-against-one (SVM-1v1) and one-against-all (SVM-1vA) approaches [53].
- The reformulation of SVM for ordinal classification with implicit constraints (SVORIM) [26].
- Weighted Support Vector Machine with Ordered Partitions (WSVMOP), which considers a binary decomposition method using weight-based SVMs [29].
- Extreme Learning Machines for nominal (ELMNO) and ordinal classification (ELMOR) [54], which are randomised algorithms for training neural networks.
- The hierarchical models described in Section 5.1 based on logistic regression (H-LR) and support vector machines (H-SVM).
- The ordinal Cascade binary utility model proposed in Section 5.2, where two different prediction approaches are used: the hierarchical approach (OC-H) and the ECOC framework (OC-E) [20]. We test these models using SVMs (OC-E-SVM and OC-H-SVM) and RMLR (OC-E-LR and OC-H-LR).

These methods make use of the vector of features extracted from each picture (see Section 4). On the other hand, deep learning based methods automatically extract features in image processing as part of the model optimisation process [55]. Specifically in the field of image processing, deep convolutional neural networks (CNN) have obtained significant performance improvement over previous approaches [56]. CNNs are artificial neural networks composed of several layers that automatically extract features from images from lower to higher abstraction levels by performing non-linear transformations in each layer. The first layers of the CNN perform an autoencoding of raw images, whereas the last layers use the previous data transformations to perform high level classification of images, i.e. image labelling.

Deep learning models need large amounts of data and computational resources to build the above-mentioned complex models. However, many pre-trained models can be re-trained to adapt them to a new classification task. This is formally known as transfer learning [57], and it is specifically suitable for two non-excluding situations: first, complex deep neural network models are computationally costly to tune, and transfer learning computational cost is significantly lower; and second, transfer learning obtains competitive results when the new task amount of data is not enough to train the model from scratch.

Recently, Esteva et al. [58] applied the GoogleNet Inception v3 CNN architecture [59,60] to the classification of several skin diseases. In that work, the authors used transfer learning to re-train a model that was pre-trained on approximately 1.28 million images and 1000 object categories from 2014 ImageNet Large Scale Visual Recognition Challenge [61]. The re-trained model results were aligned with dermatologist performance. In [58], malignant melanomas are distinguished from benign nevi (non-melanoma), but the thickness of the melanoma is not included in the classification levels.

In order to compare our approach to a CNN based approach, we have re-trained GoogleNet Inception v3 CNN in a similar way as in [58]. We have removed the final classification layer and re-trained it with the raw images of our dataset. Results are included in Table 2 but not in the rest of the experiments, because over-sampling and feature selection methods work with the feature vectors but not with raw images.

6.2. Evaluation metrics and experimental design

To take into account different aspects of classification performance evaluation, we have selected different metrics focused on the global performance, the balance of performance for the different classes and the ordinal magnitude of the errors:

- Accuracy (*Acc*) is the percentage of correctly classified patterns:

$$Acc = \cdot \frac{1}{N}\sum_{i=1}^{N}[\![\hat{y}_i = y_i]\!],$$

[1] http://www.uco.es/grupos/ayrna/partial-order-melanoma.
[2] https://github.com/ayrna/orca.

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:

$$GM = \sqrt[Q]{\prod_{q=1}^{Q} S_q},$$

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*) [24]:

$$AMAE = \frac{1}{Q}\sum_{q=1}^{Q} MAE_q = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{N_q}\sum_{i=1}^{N_q} e(\mathbf{x}_i), \tag{11}$$

$$MMAE = \frac{1}{Q}\max_{q=1}^{Q} MAE_q, \tag{12}$$

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 qth label. *AMAE* values range from 0 to 4, and so do *MMAE* values.

The first class ($\mathcal{C}_1$) is also considered for the ordinal errors, given that the partial order assumption is considered in the input space, but the misclassification costs of $\mathcal{C}_1$ with respect to $\mathcal{C}_2$ to $\mathcal{C}_5$ can be assumed to be the same that the ones applied in the case of ordinal classification.

The experiments have been performed with the original dataset, and two datasets with synthetic patterns generated as described in Section 5.3. The experimental design consists on a stratified 10-fold partition procedure, and the metrics are calculated using the sum of all generalisation confusion matrices from the 10 folds. To adjust the kernel width and cost parameters for the SVM-based methods (REDSVM, SVM-1v1, SVM-1vA, SVORIM, WSVMOP, H-SVM, OC-E-SVM and OC-H-SVM), a nested cross-validation is applied to the training data, with a grid search with parameter values within the range $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. In the case of SVM based hierarchical model (H-SVM), we have two parameters (C and γ, width of the kernel) corresponding to *C*-SVC and SVORIM. Since adjusting these parameters would lead to a four dimensions grid search, we use the same parameters C and γ for the binary and the ordinal model. The kernel width of KDLOR is optimised using the same range than SVM-based methods with regularisation parameter values in the range $u \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. The cost parameter of RMLR, H-LR, OC-E-LR and OC-H-LR is adjusted using the same values of the SVM-based methods. For ELMNO and ELMOR the number of sigmoid hidden neurons is optimised from the set {5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100}. The criteria for selecting the parameters is *AMAE*, which had a positive impact on the performance of the metrics related to imbalance. The POM does not have hyperparameters to optimise. CNN Inception v3 was re-trained with the default parameters, as indicated in the project website.[3]

Table 2
Experimental results obtained for the different methods considered in the original dataset. The best and second-best results are in bold face and italics, respectively.

Method	*Acc*	*GM*	*AMAE*	*MMAE*
Nominal methods				
MLR	0.632	0.393	0.857	1.172
RMLR	0.632	0.375	0.852	1.167
ELMNO	0.610	0.212	1.194	1.897
SVC1V1	**0.665**	0.402	0.910	1.448
SVC1VA	0.649	0.377	0.920	1.352
Inception-v3	0.635	0.000	1.292	2.714
Ordinal methods				
POM	0.616	0.320	0.852	1.241
ELMOR	0.555	0.202	1.068	1.862
SVORIM	0.633	0.401	0.816	1.207
REDSVM	0.635	0.359	0.827	1.138
WSVMOP	0.653	0.333	0.922	1.414
KDLOR	0.546	0.357	0.883	1.345
Proposed methods – partial order				
H-LR	**0.665**	**0.416**	*0.770*	1.093
H-SVM	0.642	0.355	0.812	1.138
OC-E-LR	0.582	0.384	0.839	1.125
OC-E-SVM	0.575	*0.407*	**0.764**	**0.922**
OC-H-LR	0.639	0.406	0.773	*1.000*
OC-H-SVM	*0.660*	0.394	0.822	1.138

Table 3
Experimental results obtained for the different methods considered in the total order over-sampled dataset. The best and second-best results are in bold face and italics, respectively.

Method	*Acc*	*GM*	*AMAE*	*MMAE*
Nominal methods				
MLR	0.623	0.402	0.862	1.207
RMLR	*0.641*	*0.404*	0.854	1.148
ELMNO	0.557	0.283	1.070	1.370
SVC1V1	0.626	0.334	0.947	1.481
SVC1VA	0.616	0.323	0.977	1.414
Ordinal methods				
POM	0.598	0.378	**0.815**	**1.016**
ELMOR	0.569	**0.410**	0.861	1.167
SVORIM	0.560	0.356	0.856	1.276
REDSVM	0.548	0.338	0.871	1.276
WSVMOP	0.630	0.375	0.899	1.276
KDLOR	0.475	0.322	0.914	1.414
Proposed methods – partial order				
H-LR	**0.644**	0.381	0.865	1.333
H-SVM	0.637	0.358	*0.847*	1.148
OC-E-LR	0.625	0.344	0.929	1.345
OC-E-SVM	0.548	0.362	0.861	*1.094*
OC-H-LR	*0.641*	0.339	0.940	1.448
OC-H-SVM	0.621	0.326	0.885	1.370

6.3. Experimental results

The results of the experiments performed can be seen in Tables 2–4, where the best performing method is highlighted in bold face and the second one in italics. *Acc* and *GM* are to be maximised, whereas *AMAE* and *MMAE* have to be minimised. From those tables, several conclusions can be drawn. Table 2 shows the results for the original dataset (without over-sampling). Firstly, it can be inferred that the problem can be addressed as standard ordinal regression, as the performance of state-of-the-art methods is relatively satisfactory (specially that of SVM-based methods). Comparing the nominal SVM with other ordinal methods (e.g. SVORIM and REDSVM), it can be seen that the ordinal approaches obtain better performance in the ordinal metrics, thus validating the need of considering this problem as an ordinal one. This is also applicable to ELM (comparing ELMNO and ELMOR). The *Acc* metric can be thus misleading, in such a way that the best accuracy method (SVM-1v1) obtains the fifth worst *AMAE* results and the third worst *MMAE* values. In this way, high *Acc* values hide significant errors

[3] https://www.tensorflow.org/tutorials/image_retraining.

Table 4
Experimental results obtained for the different methods considered in the partial order over-sampled dataset. The best and second-best results are in bold face and italics, respectively.

Method	*Acc*	*GM*	*AMAE*	*MMAE*
Nominal methods				
MLR	0.625	0.402	0.855	1.172
RMLR	0.630	0.390	0.880	1.333
ELMNO	0.569	0.320	1.053	1.481
SVC1V1	0.625	0.366	0.894	1.333
SVC1VA	0.617	0.352	0.904	1.333
Ordinal methods				
POM	0.612	*0.413*	**0.756**	**0.953**
ELMOR	0.596	**0.424**	*0.824*	*1.103*
SVORIM	0.539	0.316	0.896	1.345
REDSVM	0.537	0.302	0.869	1.379
WSVMOP	**0.651**	0.388	0.875	1.310
KDLOR	0.512	0.348	0.875	1.414
Proposed methods – partial order				
H-LR	0.639	0.368	0.916	1.426
H-SVM	0.601	0.291	0.896	1.241
OC-E-LR	0.625	0.369	0.945	1.296
OC-E-SVM	0.562	0.348	0.903	1.276
OC-H-LR	*0.644*	0.368	0.945	1.352
OC-H-SVM	0.637	0.363	0.877	1.296

for some of the classes (specially, for minority classes) and do not take the order information into account.

Considering our proposals for partial order problems, very interesting results can be found. Firstly, H-LR obtains the same performance for *Acc* than SVM-1v1 (i.e. the highest result) but also finds a suitable balance between the rest of metrics. This shows that, although a nominal approach could lead to acceptable results, these can be optimised by the use of a specific ordinal approach (in this case, for partial order). This also shows that the combined use of two linear methods (in H-LR) can help to improve the classification of a kernel one (e.g. in SVM-1v1), which could come from the high number of input features, which simplifies the class discrimination.

Secondly, from the cascade binary utility model variants proposed in this paper (hierarchical prediction and ECOC method, i.e. OC-H against OC-E), it seems that the hierarchical method works better for LR (where very competitive results are obtained for all the metrics), while the ECOC framework is better suited for SVM, where minority classes are very well-classified. Moreover, H-LR and OC-H-LR frameworks significantly improve their performance considering all the performance metrics with respect to POM and RMLR. This allows the use of two linear models for melanoma detection and thickness classification with feature relevance analysis purposes. The differences of partial order proposals with respect to SVORIM and REDSVM are more clear for ordinal metrics (*AMAE* and *MMAE*). This is due to the fact that, given than partial order is not exploited by standard ordinal regression methods, they can misclassify class $\mathcal{C}_1$ with a higher probability. As $\mathcal{C}_1$ is an extreme class, the errors committed for this class tend to be of higher magnitude.

Thirdly, the CNN Inception-v3 performance in *Acc* was acceptable. However, it presents the lowest possible *GM* and a large magnitude of errors in *AMAE* and *MMAE*. The confusion matrix revealed that it was not able of correctly classify any pattern of stage 0 and have numerous errors in stage II and stage III (which are the minority classes, see Table 1). On the other hand, it correctly classified patterns of classes Non-melanoma and stage I (majority classes). Then we can conclude that Inception-v3 need more patterns of minority classes to improve the model performance for those type of situations. Note that this model is not specifically designed to deal with imbalanced or ordinal data.

In order to compare our results with the ones reported in the literature, which are generally based on binary classification approaches, we provide the performance of the melanoma presence (positive class) or absence (negative class) classification task. The H-LR performance is 86.61%/90.58% (sensitivity/specificity) in the original dataset, 89.24%/85.25% in the total oversampled dataset, and 89.19%/85.00% in the partial order oversampled dataset. The performance is aligned with that reported in previous binary classification works 89%/89% in [9], 84.09%/96.61% in [7], and 91.2%/81.7% in [62]. The reader should consider that our proposals solve a more difficult five-class classification problem instead of a binary classification task. In addition, the features extracted differ.

Tables 3 and 4 present the results for the over-sampled datasets, where Table 3 shows the performance obtained when using the full-ordinal regression over-sampling approach (considering the total order of the data) and Table 4 represents the partial order of the data. This means that the full-ordinal regression over-sampling approach obviates the partial order of the problem, trying to reconstruct the original ordinal manifold using all the classes (including $\mathcal{C}_1$), whereas the partial order over-sampling only considers those classes that hypothetically would represent an order between them. The conclusions that can be extracted from these tables are the following. Firstly, the over-sampling with partial order (Table 4) shows better performance in general than the full-order over-sampling (Table 3), as shown by the highlighted best models, which motivates again the necessity of developing specific methods which consider the structure of the data. Secondly, the use of over-sampling in an imbalanced domain improves the results in the case of nominal and ordinal methods, but not in partial order ones. This could be due to the fact that the proposed methods already take into account the imbalance nature of the dataset, proposing decompositions that alleviate this degree of imbalance (e.g. separating benign lesions from melanoma). When introducing in this setting new synthetic patterns, great emphasis is put in these minority classes, then producing worse results (probably due to over-fitting). Finally, it can be seen that the application of the over-sampling method is more beneficial for POM, ELMOR and ELMON. However, these results are worse than the one obtained by decomposition methods without over-sampling. This means that, although the results of baseline methods can be improved by the use of over-sampling techniques, a method specifically designed for a given problem (as in this case, a method designed for imbalanced learning and partially ordered classes) has a greater potential than a generic method combined with over-sampling.

6.4. Feature selection results

The high number of dimensions (100 features) motivates the use of feature selection (FS) methods to discard potentially useless features. FS can be divided into individual feature ranking (FR) and feature subset selection (FSS). The former measures feature-class relevance to create a rank of features, and the top-ranked ones are selected. On the other hand, FSS aims at finding a set of features which presents good performance. In [63], the fast correlation-based filter (FCBF) is proposed as an hybrid model which takes advantage of both approaches. In this work we use the FCBF implementation available in Weka [64].

The FCBF method was applied to the original dataset and it reduced the number of features to 18 (x_1–x_3, x_{10}, x_{13}, x_{16}, x_{25}, x_{29}–x_{33}, x_{37}, x_{38}, x_{55}, x_{61}, x_{64}, x_{96}). This subset of features is included in Fig. 10 in the following section to allow comparison between the selected features and the feature weights of the H-LR model in the original dataset. Table 5 presents the results of all the methods in the reduced dataset. The CNN was excluded since it works with raw images. The comparison of results in Tables 2 and 5 is illustrated in Fig. 9, which shows the performance difference for all the metrics.

From Table 5, we can conclude that the best results are achieved by KDLOR (best results in *GM*, *AMAE* and second best result in *MMAE*). The performance of partial ordering methods is still very

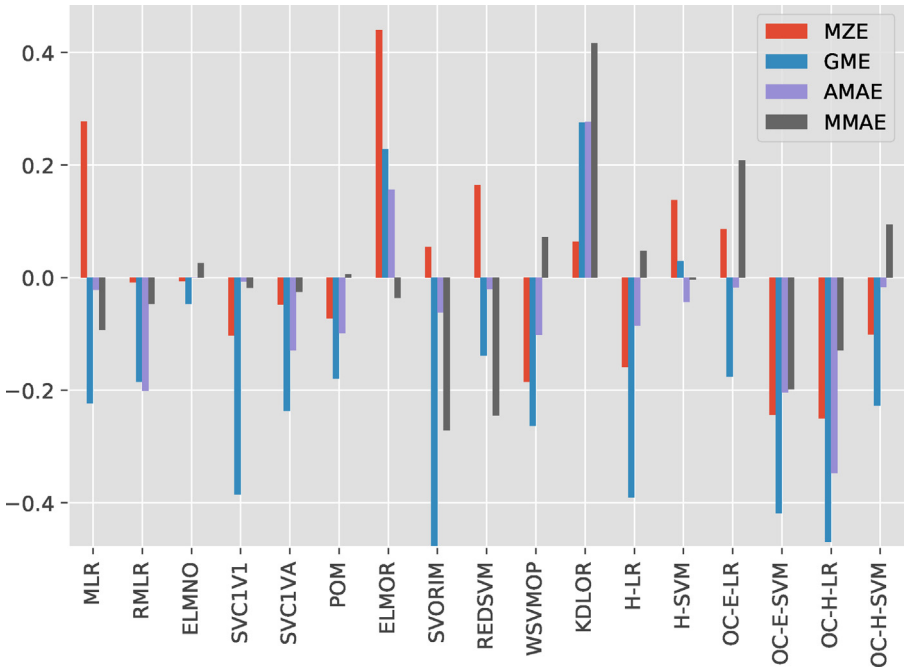


Fig. 9. Performance difference between the dataset with all the original features and that including only the ones selected by FCBF. To ease the comparison, *Acc* and *GM* are expressed as errors (Mean zero-one error, $MZE = 1 - Acc$, and $GME = 1 - GM$, respectively) and all the metrics are scaled in the range [0, 1]. Positive values represent performance differences favouring the FCBF method.

Table 5
Experimental results obtained for the different methods considered in the original dataset after performing feature selection. The best and second-best results are in bold face and italics, respectively.

Method	*Acc*	*GM*	*AMAE*	*MMAE*
Nominal methods				
MLR	**0.662**	0.341	0.866	1.296
RMLR	0.626	0.331	0.920	1.259
ELMNO	0.603	0.197	1.112	1.741
SVC1V1	0.649	0.315	0.901	1.444
SVC1VA	0.639	0.322	0.948	1.379
Ordinal methods				
POM	0.601	0.277	0.887	1.276
ELMOR	0.601	0.246	0.967	1.759
SVORIM	0.635	0.294	0.848	1.448
REDSVM	0.651	0.325	0.843	1.379
WSVMOP	0.626	0.272	0.939	1.352
KDLOR	0.544	**0.412**	**0.789**	*1.059*
Proposed methods – partial order				
H-LR	0.642	0.328	*0.821*	1.138
H-SVM	*0.655*	*0.357*	0.839	1.207
OC-E-LR	0.585	0.342	0.851	**1.047**
OC-E-SVM	0.536	0.313	0.855	1.188
OC-H-LR	0.603	0.301	0.908	1.196
OC-H-SVM	0.644	0.341	0.838	1.138

competitive (see H-LR, H-SVM and OC-E-LR), and it is better than that of nominal and ordinal classification methods, with the exception of KDLOR for *GM*, *AMAE* and *MMAE*.

Fig. 9 leads to the following conclusions. In general, the performance of most of the methods is reduced by feature selection. The results in *Acc* are not significantly affected. However, the performance of the metrics which are sensitive to class imbalance is decreased in the majority of cases. This behaviour can be explained because many feature selection methods can suppress features/attributes which may not affect the global performance, but are important to distinguish the minority classes.

In conclusion, the models produced after FS are much more simple but lead to a general performance decay. A wider study including more filters could improve the efficiency for the minority classes. However, such a study is out of the scope of the present work.

6.5. Features and model analysis

This section presents an analysis of the relevance of the features used by H-LR models as well as a comparison of them with the features subset selected by FCBF.

6.5.1. Analysis of the original dataset H-LR model

This section presents an analysis of the H-LR models corresponding to experiments summarised in Table 2. We analyse some of the variables because a whole feature analysis is out of the scope of the current work. Since these experiments correspond to a 10-fold set up, we have 10 resulting models. Fig. 10 presents box plots of these 10 models together with the best performing model. Each hierarchical model is composed of a binary model (upper plot in the figure) and an ordinal model (bottom plot of the figure represented in logarithmic scale). Each figure represents the absolute w_i values of each model, so that we can analyse the relevance and robustness of each variable for each classification task. As a general comment, feature weights with low values and small variance do not have high relevance to the model. On the other hand, weights with higher values have greater contribution to the model. The high variability of some weights can be interpreted as a lack of robustness of corresponding feature, whose relevance highly depends on the train data sample of the fold.

Fig. 10 also indicates which features were designed to contribute to each one of the two classification tasks (melanoma distinction and thickness estimation). For instance, we can see that circularity index ($x1$) is more relevant in the binary case than in the ordinal one. On the other side, it should be noticed that some variables designed to measure the thickness of the melanoma are also relevant for the binary classification task (see x_{15}, x_{19}, x_{20}, x_{22} among others). As expected, variables such as the percentage of the lesion area of reference colours (x_{12}–x_{17}) have higher relative weight in the POM model that in the LR model.

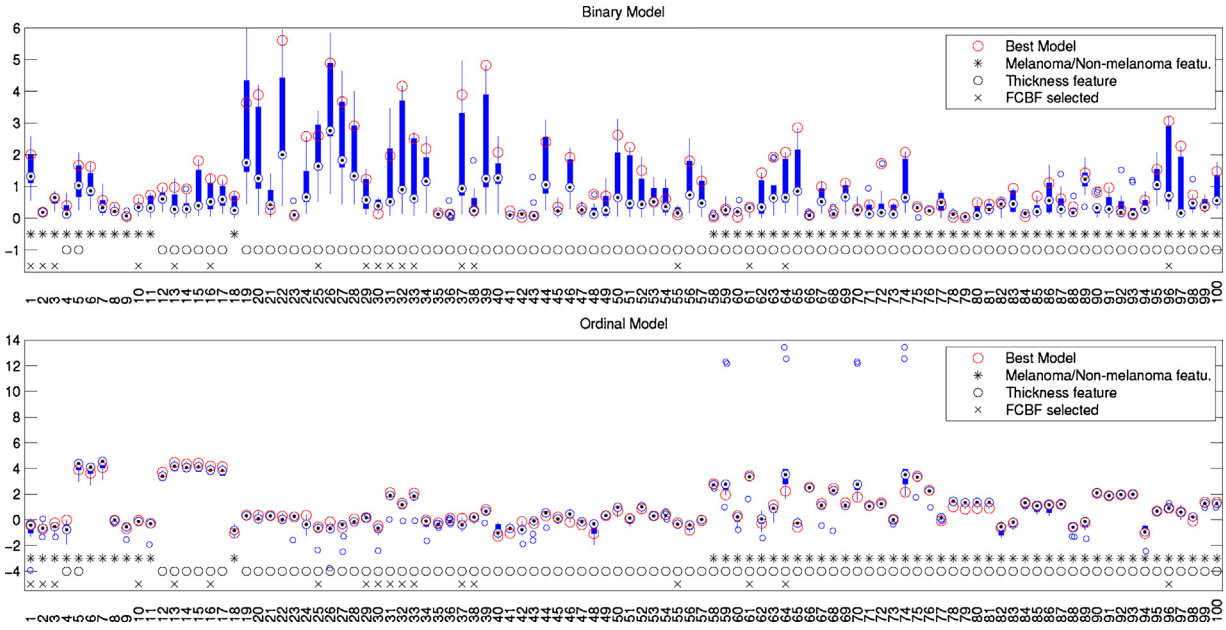


Fig. 10. Box plots of the 10 H-LR models (binary and ordinal models) together with the best performing model. The figure represents the weight w_i of each model corresponding to every feature. Since the features are standardised, we use the absolute value of w_i. The ordinal model is represented in logarithmic scale. The features selected by the FCBF filter are marked with a "×" symbol (see Section 6.5.2).

Considering the features corresponding to shape (x_1–x_{11}) and percentage of colour area (x_{12}–x_{17}), these seem more robust than the ones based on statistical descriptors of colours (x_{19}–x_{54}). That is, for the last ones, even though these variables highly influence the models, their influence is affected by the training sample. This behaviour could be explained by the fact that these features, in comparison to others, are not as clearly defined as clinical criteria for both problems. In addition, we can hypothesise that with a larger training sample the variance of some of the weights could be reduced so more robust conclusions could be achieved. The analysis of these descriptors reveals that some channels are more relevant than others. For instance, features x_{43}–x_{48} are the entropy of RGB and $L^*a^*b^*$ colour spaces. In this case, entropy of green (x_{44}) and L (x_{46}) is more relevant to both models than the entropy of other components. Regarding pigment network features (x_{55}–x_{57}), network density ratio (x_{55}) is not relevant for the binary model, but the other two features are relevant for both models. The texture features (x_{58}–x_{100}) are in general more robust variables where some of them have a very low weight, specially in the case of the binary model. Finally, in the POM model, the following blocks of variables based on MRF model [36] are outlined: x_{78}–x_{81}, x_{84}–x_{87} and x_{90}–x_{93}.

Finally, the features added with respect to previous works also contribute to both models. These are: shape features x_6 and x_7 (specially in the ordinal models) and colour features x_{43}–x_{54} (with more relevance in the binary models).

6.5.2. Comparison of H-LR model to FCBF feature selection

In this last section, we compare the relevance that H-LR assigns to features in the original dataset to the FCBF subset obtained in Section 6.4. Fig. 10 represent the FCBF selected features to ease the comparison.

Considering the binary model, some of the FCBF features are relevant (for instance x_1, x_{13}, x_{16}, x_{29}–x_{33} and x_{37}) whereas x_{38} and x_{55} have a low contribution to the H-LR model. Regarding the ordinal model, the FCBF features x_1–x_3 are not relevant (these variables seem to be important only for the binary model). Features x_{13}, x_{16}, x_{55} are known to be clinically relevant to distinguish deeper melanomas. Indeed, x_{13} and x_{16} are relevant for the ordinal model but x_{55} has a smaller contribution to this model. Finally, other variables such as x_{61} and x_{64} are relevant variables in the ordinal model and have also been selected by the FCBF algorithm.

7. Conclusions

This paper presents a novel approach for automatic melanoma characterisation via computational image analysis and machine learning methods. 100 features based on clinical insights are extracted to describe each image. Learning is tackled using classifiers which simultaneously differentiate melanomas from benign lesions and, in the case of melanomas, predict the stage. The stage (depth) is characterised by using a set of ordered labels, representing different stages according to the Breslow index. We experimentally confirm the hypothesis that topology-aware methods improve different aspects of the classification (evaluated through different performance metrics such as *GM*, *AMAE* and *MMAE*). More specifically, we propose three decomposition based approaches (a hierarchical model and two methods based on the ordinal cascade utility model). For the ordinal cascade utility model, two different strategies for fusing the predictions of the binary models are used.

We experimentally confirm that this problem is a partially ordered dataset, since the best performance was achieved by the partial order methods with respect to a wide collection of nominal and ordinal classifiers. The results outline that this partial order should be taken into account to minimise the magnitude of the errors. Also, the optimisation of model hyper-parameters with *AMAE* as selection criteria contributed to reduce the magnitude of errors and to balance the prediction of all the classes. Finally, we have applied over-sampling schemes that take into account the topology of the classes in the input space, which are able to improve the performance of standard nominal and ordinal methods (especially if the partial order is considered during synthetic data generation).

Our approaches have been also compared to a recent convolutional deep neural network (CNN) technique specifically designed for classifying skin lesions. The global accuracy was aligned with the rest of the methods, but the CNN model presented worse performance for minority classes. This can be due to the fact that CNN

models need large amounts of data to be trained, and the number of patterns for this problem is low in some cases. However, the CNN performance could be improved in several ways. The first is increasing the training data, not only including new images, but also testing the effect of generating new images by applying rotation, zoom and translation operations to the training pictures. Besides, specific architecture and parameter tuning could be explored.

We have included a study on feature relevance (weights) assigned by the best performing method and a set of experiments with feature filtering. In this case, the performance was clearly affected for the minority classes. As future work, an extended analysis of feature selection could bring a trade-off between model complexity (in terms of number of variables) and performance for all the classes. In addition to this, non-negative linear models [65] can be explored to focus on model interpretability.

Future work lines mostly rely on data acquisition. Although the use of a public dataset has advantages (such as experiment reproducibility or to avoid dependencies on specific hardware), more images acquired with modern dermatoscopes are needed. A larger and more modern dataset would strengthen the conclusions regarding the expected performance of the system (as well as model analysis conclusions) and would allow the construction of a more general skin lesion detection framework. Finally, partial order classification methods could also be tested in other application fields, to validate their usefulness under the assumption of partial order.

Acknowledgements

This work has been subsidised by the projects TIN2014-54583-C2-1-R and TIN2015-70308-REDT of the Spanish Ministry of Economy and Competitiveness (MINECO) and FEDER funds (FEDER EU).

References

[1] International Agency for Research on Cancer, World Health Organization, Cancer Factsheet. Malignant Melanoma of Skin, 2015, Available on http://eco.iarc.fr/eucan/Cancer.aspx?Cancer=20 (accessed 15.12.2015).
[2] M. Pizzichetta, G. Argenziano, R. Talamini, D. Piccolo, A. Gatti, G. Trevisan, G. Sasso, A. Veronesi, A. Carbone, H. Peter Soyer, Dermoscopic criteria for melanoma in situ are similar to those for early invasive melanoma, Cancer 91 (5) (2001) 992–997.
[3] C. Herman, Emerging technologies for the detection of melanoma: achieving better outcomes, Clin. Cosmet. Investig. Dermatol. 5 (2012) 195–212.
[4] M. Binder, M. Schwarz, A. Winkler, A. Steiner, A. Kaider, K. Wolff, H. Pehamberger, Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists, Arch. Dermatol. 131 (3) (1995) 286–291.
[5] L. Yu, H. Chen, Q. Dou, J. Qin, P.A. Heng, Automated melanoma recognition in dermoscopy images via very deep residual networks, IEEE Trans. Med. Imaging 36 (4) (2017) 994–1004.
[6] I. Maglogiannis, C.N. Doukas, Overview of advanced computer vision systems for skin lesions characterization, IEEE Trans. Inf. Technol. Biomed. 13 (5) (2009) 721–733.
[7] R. Garnavi, M. Aldeen, J. Bailey, Computer-aided diagnosis of melanoma using border- and wavelet-based texture analysis, IEEE Trans. Inf. Technol. Biomed. 16 (6) (2012) 1239–1252.
[8] M. Celebi, H. Kingravi, B. Uddin, H. Iyatomi, Y. Aslandogan, W. Stoecker, R. Moss, A methodological approach to the classification of dermoscopy images, Comput. Med. Imaging Graph. 31 (6) (2007) 362–373.
[9] L. Li, Q. Zhang, Y. Ding, H. Jiang, B.H. Thiers, J.Z. Wang, Automatic diagnosis of melanoma using machine learning methods on a spectroscopic system, BMC Med. Imaging 14 (1) (2014) 1–12.
[10] P. Rubegni, G. Cevenini, P. Sbano, M. Burroni, I. Zalaudek, M. Risulo, G. Dell'Eva, N. Nami, A. Martino, M. Fimiani, Evaluation of cutaneous melanoma thickness by digital dermoscopy analysis: a retrospective study, Melanoma Res. 20 (3) (2010) 212–217.
[11] A. Sáez, J. Sánchez-Monedero, P.A. Gutiérrez, C. Hervás-Martínez, Machine learning methods for binary and multiclass classification of melanoma thickness from dermoscopic images, IEEE Trans. Med. Imaging 35 (4) (2016) 1036–1045.
[12] M. Amouroux, W. Blondel, Non-invasive determination of Breslow index, in: M.Y. Cao (Ed.), Current Management of Malignant Melanoma, InTech, 2011, pp. 29–44 https://hal.archives-ouvertes.fr/hal-00626482.
[13] P.A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernandez-Navarro, C. Hervás-Martínez, Ordinal regression methods: survey and experimental study, IEEE Trans. Knowl. Data Eng. 28 (1) (2016) 127–146.
[14] O.M. Doyle, E. Westman, A.F. Marquand, P. Mecocci, B. Vellas, M. Tsolaki, I. Kloszewska, H. Soininen, S. Lovestone, S.C.R. Williams, A. Simmons, Predicting progression of Alzheimer's disease using ordinal regression, PLOS ONE 9 (8) (2014) 1–10.
[15] J. Sánchez-Monedero, P. Campoy-Muñoz, P.A. Gutiérrez, C. Hervás-Martínez, A guided data projection technique for classification of sovereign ratings: the case of European Union 27, Appl. Soft Comput. 22 (2014) 339–350.
[16] D. Devlaminck, W. Waegeman, B. Bauwens, B. Wyns, P. Santens, G. Otte, From circular ordinal regression to multilabel classification, Proceedings of the 2010 Workshop on Preference Learning (European Conference on Machine Learning, ECML) (2010) 15.
[17] K. Fernandes, J.S. Cardoso, Discriminative directional classifiers, Neurocomputing 207 (2016) 141–149.
[18] W. Cheng, E. Hüllermeier, K.J. Dembczynski, Graded multilabel classification: the ordinal case, Proceedings of the 27th International Conference on Machine Learning (ICML-10) (2010) 223–230.
[19] E. Hüllermeier, J. Fürnkranz, W. Cheng, K. Brinker, Label ranking by learning pairwise preferences, Artif. Intell. 172 (16) (2008) 1897–1916.
[20] M. Pérez-Ortiz, M. Cruz-Ramírez, M. Ayllón-Terán, N. Heaton, R. Ciria, C. Hervás-Martínez, An organ allocation system for liver transplantation based on ordinal regression, Appl. Soft Comput. 14 (Pt A) (2014) 88–98.
[21] M. Pérez-Ortiz, P.A. Gutiérrez, C. Hervás-Martínez, X. Yao, Graph-based approaches for over-sampling in the context of ordinal regression, IEEE Trans. Knowl. Data Eng. 27 (5) (2015) 1233–1245.
[22] M. Pérez-Ortiz, J. Sánchez-Monedero, A.A. Sáez, P.A. Gutiérrez, C. Hervás-Martínez, Tackling the ordinal and imbalance nature of a melanoma image classification problem, International Joint Conference on Neural Networks (IJCNN) (2016) 2156–2163.
[23] J.C. Hühn, E. Hüllermeier, Is an ordinal class structure useful in classifier learning? Int. J. Data Min. Model. Manag. 1 (1) (2008) 45–67.
[24] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero, P.A. Gutiérrez, Metrics to guide a multi-objective evolutionary algorithm for ordinal classification, Neurocomputing 135 (2014) 21–31.
[25] P. McCullagh, Regression models for ordinal data, J. R. Stat. Soc. Ser. B (Methodol.) 42 (2) (1980) 109–142.
[26] W. Chu, S.S. Keerthi, Support vector ordinal regression, Neural Comput. 19 (3) (2007) 792–815.
[27] J. Sánchez-Monedero, P.A. Gutiérrez, P. Tiňo, C. Hervás-Martínez, Exploitation of pairwise class distances for ordinal classification, Neural Comput. 25 (9) (2013) 2450–2485.
[28] E. Frank, M. Hall, A simple approach to ordinal classification, Proc. of the 12th Eur. Conf. on Machine Learning (2001) 145–156.
[29] W. Waegeman, L. Boullart, An ensemble of weighted support vector machines for ordinal regression, Int. J. Comput. Syst. Sci. Eng. 3 (1) (2009) 47–51.
[30] H.-T. Lin, L. Li, Reduction from cost-sensitive ordinal ranking to weighted binary classification, Neural Comput. 24 (5) (2012) 1329–1367.
[31] A. Breslow, Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma, Ann. Surg. 172 (5) (1970) 902–908.
[32] M. Stante, V. De Giorgi, P. Cappugi, B. Giannotti, P. Carli, Non-invasive analysis of melanoma thickness by means of dermoscopy: a retrospective study, Melanoma Res. 11 (2) (2001) 147–152.
[33] M. Lens, P. Nathan, V. Bataille, Excision margins for primary cutaneous melanoma: updated pooled analysis of randomized controlled trials, Arch. Surg. 142 (9) (2007) 885–891.
[34] M. Brady, D. Coit, Sentinel lymph node evaluation in melanoma, Arch. Dermatol. 133 (8) (1997) 1014–1020.
[35] G. Argenziano, H.P. Soyer, V.D. Giorgio, D. Piccolo, P. Carli, M. Delfino, A. Ferrari, R. Hofmann-Wellenhof, D. Massi, G. Mazzocchetti, M. Scalvenzi, I.H. Wolf, Interactive Atlas of Dermoscopy, EDRA-Medical Publishing and New Media, Milan, 2000.
[36] A. Sáez, C. Serrano, B. Acha, Model-based classification methods of global patterns in dermoscopic images, IEEE Trans. Med. Imaging 33 (5) (2014) 1137–1147.
[37] A. Sáez, C.S. Mendoza, B. Acha, C. Serrano, Development and evaluation of perceptually adapted colour gradients, IET Image Process. 7 (4) (2013) 355–363.
[38] J. Sánchez-Monedero, A. Sáez, M. Pérez-Ortiz, P.A. Gutiérrez, C. Hervás-Martínez, Classification of melanoma presence and thickness based on computational image analysis, in: F. Martínez-Álvarez, A. Troncoso, H. Quintián, E. Corchado (Eds.), Proceedings of the 11th International Conference Hybrid Artificial Intelligent Systems, Springer International Publishing, 2016, pp. 427–438.
[39] A. Bono, S. Tomatis, C. Bartoli, G. Tragni, G. Radaelli, A. Maurichi, R. Marchesini, The ABCD system of melanoma detection: a spectrophotometric analysis of the asymmetry, border, color, and dimension, Cancer 85 (1) (1999) 72–77.
[40] H. Soyer, G. Argenziano, R. Hofmann-Wellenhof, R. Johr, Color Atlas of Melanocytic Lesions of the Skin, Springer Berlin Heidelberg, 2010.
[41] S. Seidenari, G. Pellacani, C. Grana, Computer description of colours in dermoscopic melanocytic lesion images reproducing clinical assessment, Br. J. Dermatol. 149 (3) (2003) 523–529.
[42] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, M. Delfino, Clinical and dermatoscopic criteria for the preoperative evaluation of cutaneous melanoma thickness, J. Am. Acad. Dermatol. 40 (1) (1999) 61–68.

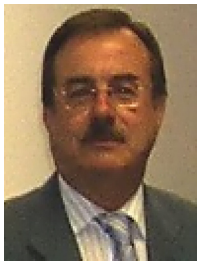
[43] H. Lorentzen, K. Weismann, F. Grønhøj Larsen, Dermatoscopic prediction of melanoma thickness using latent trait analysis and likelihood ratios, Acta Dermato-Venereol. 81 (1) (2001) 38–41.
[44] N. Otsu, Threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.
[45] M. Sadeghi, M. Razmara, T. Lee, M. Atkins, A novel method for detection of pigment network in dermoscopic images using graphs, Comput. Med. Imaging Graph. 35 (2) (2011) 137–143.
[46] R. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. 3 (6) (1973) 610–621.
[47] M. Pérez-Ortiz, P.A. Gutiérrez, C. Hervás-Martínez, Projection-based ensemble learning for ordinal regression, IEEE Trans. Cybern. 44 (5) (2014) 681–694.
[48] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[49] A. Agresti, Categorical Data Analysis, 2nd edition, John Wiley and Sons, 2002.
[50] R. Herbrich, T. Graepel, K. Obermayer, Large margin rank boundaries for ordinal regression, in: A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, 2000, pp. 115–132.
[51] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[52] H. Wu, H. Lu, S. Ma, A practical svm-based algorithm for ordinal regression in image retrieval, Proceedings of the Eleventh ACM International Conference on Multimedia (Multimedia 2003) (2003) 612–621.
[53] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011), 27:1-27:27.
[54] W.-Y. Deng, Q.-H. Zheng, S. Lian, L. Chen, X. Wang, Ordinal extreme learning machine, Neurocomputing 74 (1-3) (2010) 447–456.
[55] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (1) (2009) 1–27.
[56] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[57] S. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[58] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2) (2017) 115–125.
[59] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, 2015 http://download.tensorflow.org/paper/whitepaper2015.pdf.
[60] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016 (January 2016) 2818–2826.
[61] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[62] M. Rastgoo, G. Lemaitre, J. Massich, O. Morel, F. Marzani, R. Garcia, F. Meriaudeau, Tackling the problem of data imbalancing for melanoma classification, in: Bioimaging, Rome, Italy, 2016.
[63] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (2004) 1205–1224.
[64] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, Spec. Interest Group Knowl. Discov. Data Min. Explor. Newslett. 11 (2009) 10–18.
[65] L. Wu, Y. Yang, Nonnegative elastic net and application in index tracking, Appl. Math. Comput. 227 (2014) 541–552.

Javier Sánchez-Monedero was born in Córdoba, Spain, in 1982. He received the B.S in Computer Science from the University of Granada, Spain, in 2008 and the M.S. in Multimedia Systems from the University of Granada in 2009. In 2013 he obtained the Ph.D. degree on Information and Communication Technologies of the University of Granada. He is working as Associate Professor with the Department of Quantitative Methods at the Universidad Loyola Andalucía. His current research interests include computational intelligence methods and their applications, as well as distributed systems.

María Pérez-Ortiz was born in Cordoba, Spain. She received the B.S. degree in Computer Science in 2011, the M.Sc. degree in Intelligent Systems in 2012 and the Ph.D. degree in Computer Science and Artificial Intelligence at the University of Cordoba. She is working now as a postdoctoral research associate with the Department of Computer Science and Technology at the University of Cambridge. Her current interests include a wide range of topics concerning machine learning and computer vision.

Aurora Sáez received the degree in electronic engineering from the University of Cordoba, Spain, in 2003, the Bachelor degree in Communications Engineering in 2008 and the Ph.D. degree in Communication Engineering in 2014, both from the University of Seville, Spain. She conducts research at Signal Processing and Communications Department, University of Seville. Her research interest is mainly focused on image processing, in particular on colour image processing and its medical applications.

Pedro A. Gutiérrez-Peña was born in Córdoba (Spain). He received the B.S. degree in Computer Science from the University of Seville (Spain, 2006), and the Ph.D. degree in Computer Science and Artificial Intelligence from the University of Granada (Spain), in 2009. He is a Lecturer with the Department of Computer Science and Numerical Analysis, University of Cordoba (Spain). His current research interests include machine learning, ordinal classification, evolutionary artificial neural networks and their applications in different areas.



César Hervás-Martínez was born in Cuenca, Spain. He received the B.S. degree in Statistics and Operations Research from the Universidad Complutense, Madrid, Spain, in 1978, and the Ph.D. degree in Mathematics from the University of Seville, Spain, in 1986. He is currently Professor of Computer Science and Artificial Intelligence in the Department of Computer Science and Numerical Analysis, University of Córdoba, and an Assessor in the Department of Quantitative Methods at Universidad Loyola Andalucía. His current research interests include pattern recognition, neural networks, evolutionary computation, and the modelling of natural systems.