Knowledge-Based Systems 158 (2018) 9–24

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Dermoscopic assisted diagnosis in melanoma: Reviewing results, optimizing methodologies and quantifying empirical guidelines

Huei Diana Lee[*,a], Ana Isabel Mendes[b], Newton Spolaôr[a], Jefferson Tales Oliva[a,c], Antonio Rafael Sabino Parmezan[a,d], Feng Chung Wu[a,e], Rui Fonseca-Pinto[b,f]

[a] Laboratory of Bioinformatics (LABI), Graduate Program in Electrical Engineering and Computer Science (PGEEC), Western Paraná State University (UNIOESTE), Presidente Tancredo Neves Avenue, 6731, Foz do Iguaçu 85867-900, Brazil
[b] Polytechnic Institute of Leiria, General Norton de Matos Street, 4133, Leiria, 2411-901, Portugal
[c] Bioinspired Computing Laboratory, University of São Paulo, São Carlos, Brazil
[d] Laboratory of Computational Intelligence, University of São Paulo, São Carlos, Brazil
[e] Service of Coloproctology, University of Campinas, Campinas, Brazil
[f] Instituto de Telecomunicações - Multimedia Signal Processing Group, Leiria, Portugal

ARTICLE INFO

Keywords:
Machine learning
Data mining
Computer-aided diagnosis
Dermoscopy
Image analysis

ABSTRACT

Early diagnosis is still the most important factor to deal with skin cancer, a disease that challenges physicians and researchers. It has benefited from computer-aided diagnosis methods that successfully combine dermoscopy, Digital Image Processing, and Machine Learning techniques. This paper aims to approximate medical professionals working with dermoscopy to these methods, to join the challenge of melanoma early detection. Accordingly, a proposal for extracting, selecting and combining texture and shape features from dermoscopic images is presented. The Feature Selection task is added to the learning process to potentiate the quality of classification models. Three classical Machine Learning algorithms were applied to differentiate melanoma from non-melanoma images. The models are evaluated by standard performance measures and a multi-criteria decision analysis method. This is the first time such method is used in melanoma diagnosis. As a result, we found a decision tree that performs well and allows the explicit representation and analysis of the knowledge learned from the images. In addition, the competitiveness of our decision models in comparison with literature approaches reviewed in this work encourages further applications of Machine Learning and Feature Selection to assist computer-aided diagnosis.

1. Introduction

Starting from the simple Laennec's stethoscope to the most recent technology among image advances in medical practice, the improvements regarding medical devices always combined the existence of a practical problem and the development of technology to answer these specific needs. This historical beginning note aims to reflect the valuable dialog between Medicine and Engineering in clinical derived technological improvements, establishing the interdisciplinary approach as a successful strategy to the final goal of improving health care assistance.

Cancer, a major worldwide challenge in basic and clinical research, has gained even more visibility and control with this dialog [1]. Despite important advances in this field to understand the genesis and the control mechanisms, an early diagnosis is still a strong tool for treatment management and prognosis. This is also the case in skin cancer [2].

Epidemiology of skin cancer is frightful. The American Society of Cancer shows that the occurrence of new cases of skin cancer each year is higher than the combined incidence of breast, prostate, lung and colon cancers [3]. Also, it is reinforced in [4] that over the past three decades, more people have had skin cancer than all other cancers combined.

When a histological study of the skin is performed, it is possible to describe three main layers: epidermis, dermis, and hypodermis. The upper layer is itself divided into five more sublayers. Among other types of cells with several functions present in the skin, the melanocytes are specialized cells located in the basal layer of the epidermis. After sun exposure in physiological conditions, these particular cells produce a brown/dark pigment known as melanin. The presence of this pigment

* Corresponding author.
E-mail addresses: huei.lee@unioeste.br (H.D. Lee), aimendes@ipleiria.pt (A.I. Mendes), newtonspolaor@gmail.com (N. Spolaôr), wu.chung@unioeste.br (F.C. Wu), rui.pinto@ipleiria.pt (R. Fonseca-Pinto).

https://doi.org/10.1016/j.knosys.2018.05.016
Received 28 September 2017; Received in revised form 10 May 2018; Accepted 12 May 2018
Available online 15 June 2018
0950-7051/

allows one to classify skin lesions as melanocytic (when it is present) or non-melanocytic (otherwise).

Regarding the classification of skin cancer, it is possible to identify three types according to the origin of the process at the skin layer. Basal Cell Carcinoma (BCC) originates from the basal layer of the epidermis and is the most common type of skin cancer. BCC usually appears on the face, neck or back, presenting a slow-growing locally invasive, thus not presenting metastases.

Spinous Cell Carcinoma (SCC) has its origin in the middle layers of the epidermis (spine layer) and is the second most common type of skin cancer. In this case, the lesion usually is fast growing and is more aggressive than BCC.

Malignant melanoma (or simply melanoma) is the third skin cancer type and the most dangerous one. It is also present in other organs where melanin exists, such as eyes, gastrointestinal tract and meninges. Unlike the two previous cases where the chronic sun exposure is a risk factor, here the acute sun exposure can trigger the carcinogenic process. This fact puts forward the importance of sun exposure prevention campaigns.

The melanoma diagnosis is established after the excision of the lesion and subsequent histological study. The Breslow thickness, the microscopic ulceration, and mitotic rate have been used since 2010 to determine clinical stages of melanoma and to define treatment and prognosis [5].

Due to the central role of early diagnosis in skin cancer, a detailed exam of the skin in suspicious lesions is crucial. Hence, the use of digital image processing techniques is the natural approach to avoid the subjectivity of the human eye assessment and to provide an accurate way to quantify important clues to assist a focused diagnosis. Several protocols have been used in dermatology practice. The ABCD rule, 7 point check-list, and the Menzies method are the most known among them [6].

The use of dermoscopy in skin analysis is a practice with documented results in terms of clinical outcomes, when compared with the naked eye examination, by reducing the dermatological surgery workload of false-positive lesions. As a result, it leads to cost savings, reduced morbidity, and less scarring [7]. Also in terms of cost-effective assessment, dermoscopy decreases the number of excised benign lesions and the early detection of melanomas [8,9].

Currently, it is established the advantage of combining dermoscopy with a detailed physical examination of the skin. The dissemination of digital technology in hospitals, joining with the good performance of image processing algorithms, empower the use of Machine Learning (ML) techniques in an integrated and broadly disseminated strategy of Computer-Aided Diagnosis (CAD), as the next natural step in dermoscopy practice. The use of CAD in Medicine began around 1980 [10] and since then, the successful application has proved the relevance of its use [11]. This is the case in the emergency triage [12], related to the automatic detection of polyps in colonoscopy [13–15] and calcifications in mammography [16] as well as in chest imaging [17]. In these successful cases, the low number of false positives indicates high sensitivity levels, being determinant to the adherence of the medical community to these tools, avoiding to waste time in classifying bad hits.

Although the number of CAD publications increased in the last decades, the comparison of results between pieces of work is an awkward task. This limitation is due to some methodological constraint, such as redundancies, sampling, under and overfitting, as well as due to the absence of some of the standard metrics to evaluate the performance of the classifiers in terms of Sensitivity, Specificity, and Accuracy, as discussed in [18].

In this work, the use of ML in dermoscopy to assist diagnosis of melanocytic images is addressed. In particular, the extraction and selection of classical texture and shape features are followed by the use of three ML algorithms. Afterward, the performance of these algorithms in the classification problem melanoma *versus* non-melanoma is evaluated. This strategy of using the most promising features and well-known classification algorithms for this particular challenge accomplishes two objectives. First, the use of conventional algorithms will approach those who apply digital techniques only as a user. Second, it is aligned with the major goal of setting optimized algorithms and parameters to achieve high levels of sensitivity and specificity. Moreover, this strategy will put forward these tools in the track of its use as a truly CAD strategy in dermoscopy.

2. Review on melanoma classification results

The starting point for this work is the optimization of classic techniques in ML in dermoscopy, as well as establishing a way to evaluate its performance. In order to assess the obtained results, it is mandatory to perform the state of the art in what refers to the classification results within recent publications enrolling ML methodologies and to evaluate the dependence of these different approaches in the performance of the classifiers.

With the purpose of literature comparison, a comprehensive review of 13 papers studying the melanoma classification issue on dermoscopy images, published in the last five years (2013–2017), was conducted. Table 1 compares our proposal with the approaches reported in the 13 papers. This table also defines new abbreviations regarding method names. The following criteria were considered:

- Image set size (M + NM): number of images, including Melanoma (M) and Non-Melanoma (NM) cases, used in the study;
- #Features: number of single descriptors or descriptors categories belonging to a collection;
- Feature categories: categories of the dermoscopy image descriptors;
- Dimensionality reduction method: Feature Selection or feature construction[1] method applied in the study;
- Classification algorithm(s): supervised ML algorithm(s) used to induce classifiers;
- Performance measure(s): measures considered to evaluate the performance of classifiers;
- Evaluation strategy: sampling method adopted to estimate the ML algorithm(s) performance;
- Statistical test(s): test(s) employed to verify the occurrence of significant differences among classifiers.

A classification method based on a local image descriptor, widely known as Bag-of-Features, was evaluated in [19] to differentiate melanoma from benign lesions. The paper compares distinct strategies to extract regions of interest and achieves results showing that texture-based detectors generally outperform the dense sampling strategy.

With the same purpose, a Bag-of-Features model was also applied in [20] and [21] to identify the role of different descriptors. In the first study, the authors investigated a new feature set based on color keypoint detectors and Color-SIFT. Their outcomes evidenced that the color extensions of SIFT are more discriminative than the SIFT variation computed in the luminance images. In the second study, the authors compared different local texture and color descriptors. The obtained results indicated that the latter feature type outperformed the former one. Another characteristic of these studies is that both of them dealt with the class imbalance problem in the selected dermoscopic images. To do so, the authors have created artificial instances by repeating the feature values of melanoma cases and adding Gaussian noise.

In [22], the applicability of color constancy was analyzed by using two color-based approaches. The idea was to reduce the influence of the acquisition devices and settings on the color features extracted from the

[1] Feature construction, *a.k.a* feature extraction by the data mining community, builds new and more expressive features from the existing ones by mapping the input dimension space into another one usually smaller.

Table 1
Some properties of computerized approaches for dermoscopic image analysis published in related work.

Paper	Image set size (M + NM)	#Features	Feature categories	Dimensionality reduction method	Classification algorithm(s)	Performance measure(s)	Evaluation strategy	Statistical test(s)
[19]	176 (25 + 151)	100, 200, 300	Color and texture features	–	NN	SE, SP, and a Cost Function	10-fold Stratified Cross-Validation (SCV)	–
[20]	176 (25 + 151)	100, 200, 300, ..., 600	Scale-Invariant Feature Transform (SIFT) and Color-SIFT	–	NN	SE, SP, and Acc	SCV	–
[21]	176 (25 + 151)	100, 200, 300	Local texture and color features	–	NN	SE and SP	SCV	–
[22]	200 (40 + 160)	15, 16, ..., 50	Color features	–	SVM	SE, SP, and Acc	SCV	–
[23]	176 (25 + 151)	15, 16, ..., 50	Color and texture features	–	AdaBoost, NN, and SVM	SE, SP, and a Cost Function	SCV	–
[24]	482 (241 + 241)	25, 50, ..., 300	SIFT	–	NN and SVM	SE, SP, and Acc	SCV	–
[25]	256 (128 + 128)	150	Blob, color, texture, and location features	–	Logistic Regression	SE, SP, and Acc	Holdout Validation (HV)	Chi-square Pearson
[26]	5130 (90 + 4090 + 950 Dysplastic)	446	Shape, color, and texture features	Principal Component Analysis	Gradient Boosting, Random Forest, and SVM	SE and SP	HV	–
[27]	200 (40 + 80 + 80 Atypical)	84	2-D fast Fourier transform, 2-D discrete cosine transform, complexity feature set, color feature set, pigment network feature set, lesion shape feature, lesion orientation feature, lesion margin feature, and lesion intensity pattern feature	–	SVM	SE, SP, and Acc	HV	–
[28]	250 (83 + 167)	81	Shape, color, pigment network, and texture features	Feature selection during the learning phase	Nominal classifiers – Logistic Regression Using Initial Variables and Product Units (LIPU), Product Unit Neural Networks, Logistic Regression (LR), Kernel Discriminant Analysis, SVM – , and Ordinal classifiers – Support Vector Ordinal Regression with IMplicit constraints (SVORIM), REDuction from cost-sensitive ordinal ranking to weighted binary classification framework for SVM (REDSVM), and Kernel Discriminant Learning for Ordinal Regression (KDLOR)	Acc, Minimum Sensitivity (MS), and Average Mean Absolute Error (AMAE)	10-fold cross-validation	—
[29]	562 (185 + 377)	86	Shape, color, pigment network, and texture features	—	Nominal classifiers (Multinomial Logistic Regression, SimpleLogistic, LIPU, SVM), and Ordinal classifiers (Proportional Odds Model (POM), KDLOR, REDSVM, SVORIM)	Acc, MS, Geometric Mean (GM), and AMAE	10-fold cross-validation	—
[30]	562 (185 + 377)	86	Shape, color, pigment network, and texture features	—	SVM, LR, SVORIM, POM, Frank and Hall decomposition approach for ordinal classification using SVM, and Ordinal Cascade binary utility model using Error-Correcting Output Codes	Acc, GM, AMAE, and Maximum Mean Absolute Error	10-fold cross-validation	Wilcoxon
[31]	297 (184 + 113)	Undefined	Ridge/furrow width ratio	—	Acral Lentiginous Melanoma vs nevus classifier	SE, SP, and Acc	HV	—
This paper	104 (58 + 46)	166	Shape and geometric features, Neighborhood Gray-tone Difference Matrix, Haralick's descriptors, fractal dimension based features, Laws' texture energy measures, Local Binary Patterns	ReliefF	J48, NN, SVM	SE, SP, and Acc	10-fold stratified cross-validation	Kruskal-Wallis with Tukey post-hoc test

dermoscopy images. To apply color constancy, the authors adopted the Shades of Gray method. Although other color constancy algorithms have been examined, all of them achieved similar results. In contrast, the authors compared in [23] the role of color and texture features in lesion classification and determined which set of features were more discriminative. They concluded that color features generally outperform texture features when used alone and that both descriptors provide good results. In these two papers, the class imbalance problem was also treated.

Four color constancy algorithms – Gray World, max-RGB, Shades of Gray, and General Gray World – were investigated in [24] to correct the color variations from images generated by multiple sources, *i.e.*, the multisource problem. The outcomes showed that color constancy improves the classification of multisource dermoscopy images, increasing the sensitivity of a bag-of-features system from 71.0% to 79.7% and the specificity from 55.2% to 76% using only 1-D RGB histograms as features.

In [25], a real-time supervised detection of three shades of pink areas (light pink, dark pink, and orange pink) in dermoscopic melanoma images was conducted using color analysis techniques in five color palettes (red, green, blue, hue, and saturation). An LR model provided up to 87.9% accuracy for discriminating melanoma in 256 images. According to the authors, the lower accuracy of this models was obtained when omitting any of the following features: individual pink shades, texture information, and color location.

An automatic framework to differentiate melanoma lesions from dysplastic nevi was proposed in [26]. The system comprises of three steps: automatic segmentation, feature extraction, feature modification, and classification. Using texture features and Random Forest, the framework achieved the highest sensitivity at 98% and specificity at 70%. It should be emphasized that this proposal uses a dimensionality reduction strategy named Principal Components Analysis (PCA). It transforms the input space where images are represented into a smaller space with new dimensions. However, despite the ability to transform the feature space and provide a momentaneous reduced number of dimensions used to describe an image (the principal components), PCA will need all the original extracted features every time a new image comes in and needs to be classified.

In [27], two major components of a system for noninvasive real-time analysis of skin lesion were proposed. Its goal was melanoma early detection and prevention. The first component is a real-time alert to help users prevent skin burn caused by sunlight. In turn, the second component is an automated image analysis module that performs image acquisition, hair detection and exclusion, lesion segmentation, feature extraction and classification. From a usage standpoint, the users are able to use the system on their own smartphones and analyze their eventual skin lesions from pictures taken with the phone's camera. The system then analyzes the image and informs the user if it is a benign lesion, atypical lesion or melanoma. Using a dataset provided by the academic community, a system evaluation showed that the accuracy achieved by combining all feature sets is higher than the one obtained when features are used individually or in partial combinations.

The feature extraction strategy proposed in [28] is based on the clinical findings that correlate certain characteristics present in dermoscopic images and tumor depth. In particular, the strategy considers all the original features as input and includes new dimensions based on feature space transformations. Afterwards, during the classifier building, the approach removes dimensions that contribute little to specific image classification schemes. It should be emphasized that, as is the case for PCA, the approach will need all the original features extracted from every new image as input to classification schemes.

Specifically, two supervised classification schemes were developed: a binary scheme in which melanomas are classified into thin or thick, and a three-class scheme (thin, intermediate, and thick). Five nominal classification methods were applied to the binary problem. One of them is LIPU, a learner that combines logistic regression with artificial neural networks. Regarding the three-class scheme, seven classifiers (four nominal and five ordinal) were applied. Nominal classification disregard the order relation between classes in the data. In turn, ordinal classification takes advantage of class ordering information. For the binary case, LIPU outperforms all the other methods with an accuracy of 77.6%. On the other hand, for the second scheme, although LIPU reports the highest overall accuracy, the ordinal classification methods achieve a better balance among the performance achieved for each class. Besides the experimental highlight, the authors emphasized a LIPU advantage: the building of an interpretable model that provides probabilistic classes assignment and performs Feature Selection during the learning phase.

In [29] is presented a hybrid approach based on computational image analysis and ML to develop a tool for automatic detection of melanoma presence and severity. Image features related to shape, color, pigment network and texture were extracted to describe benign and several malign stage cases. The authors found promising the performance results achieved, indicating that the type of features extracted describes the lesions accurately.

As opposed to the previous study, which focuses on the differentiation between melanoma and non-melanoma, a finer-grain classification problem based on five categories was considered in [30]: a group of benign lesions and four different stages of melanoma. The problem is described by using a partially ordered labeling structure, which is also naturally imbalanced. To deal with these issues, the authors proposed the use of an ordinal cascade decomposition method and an ordinal oversampling technique. The outcomes demonstrated that the features selected, identical to the ones used in the preliminary study, describe the lesions properly and that the use of more complex and specialized ML techniques can be the road to build accurate models. A relevant finding for our work is also given by the authors. In particular, they agree that classification methods could benefit from an FS process. However, according to them, FS might be unnecessary for kernel methods, such as SVM. The reason is that this algorithm has a regularization parameter that establishes a weight to error minimization in the training set, concerning the minimization of the model complexity.

A novel algorithm was proposed in [31] for an automated diagnostic system. This system, in turn, considers the analysis of dermoscopy findings of Acral Lentiginous Melanoma (ALM), which can assist dermatologists in an early diagnosis. The proposed algorithm is able to precisely distinguish the furrow and ridge patterns of pigmentation on dermoscopy images using the width ratio of dark and bright patterns. According to the authors, the proposed method is sufficiently accurate, robust, and computationally fast for the discrimination of ALM, which helps dermatologists in a screening support system or in a teledermatology system.

This work presents a robust proposal for extracting, selecting and combining texture and shape features from dermoscopic images. A contribution of this proposal involves the use of a filter Feature Selection algorithm named ReliefF to support image classification [32–34]. As a result, ReliefF efficiently finds subsets of the original features independently of any learning algorithm. Thus, the best feature subsets for the classification algorithms evaluated in this work – J48, NN and SVM – could also be useful for other classification approaches. It should be emphasized that, different from [26,28], our proposal discovers subsets of features that fully characterize the training set of images only once. Consequently, only the features selected in a subset are extracted from a new image before its classification, saving computational time.

As summarized in the last row of Table 1, our proposal differs from the literature not only by incorporating the Feature Selection task to the learning process to potentialize the melanoma classification, but also includes a more rigorous experimental design supported by a new multi-criteria decision analysis method. Besides identifying relevant attributes to avoid the curse of dimensionality, we used three classical

Machine Learning classifiers from distinct paradigms. The resulting classifiers were evaluated with the support of a stratified sampling method and statistical significance tests.

3. Material and methods

The dataset used in this work is composed of 104 dermoscopic images, from which 58 and 46 are melanoma and non-melanoma images, respectively. These images are RBG true colored (24-bit color) and JPEG compressed with a minimum resolution of 300 dpi. Their acquisition was conducted according to clinical protocols in dermoscopy, following all legal requirements [35]. These images were already used in related work to validate image processing techniques in dermoscopic image processing, as presented in [18] and [36].

Image processing algorithms were implemented in MATLAB (The Mathworks, Inc., US, http://www.mathworks.com) and its Image Processing Toolbox. The Neighborhood Gray-Tone Difference Matrix features, described in Section 3.1, were developed in Java (Oracle, US, https://www.oracle.com/java). The ReliefF Feature Selection algorithm and the learning methods used in this paper (Sections 3.2 and 3.3, respectively), implemented in the Weka library (University of Waikato, New Zealand, [37]), were applied with default parameter values. As a preprocessing step, all images were cropped at 450x600 resolution and a segmentation process was followed according to [18] and [38,39].

In what follows, Sections 3.1 and 3.2 respectively describe the techniques applied to extract and select features that characterize the input images. The selected features are then submitted to classification algorithms (Section 3.3), aiming to learn a model able to differentiate non-melanoma from melanoma images. Finally, Section 3.4 reports the experimental setting applied to evaluate the proposed methodology.

3.1. Feature extraction

Dermoscopic signs represented by a set of features (texture, shape, and color) are used together to give clinical clues to diagnose melanoma. In the context of CAD, computers mostly characterize skin lesions by shape and texture features [40]. By approximating morphology properties verified by medical experts, the former type of features is crucial to differentiate melanoma from non-melanoma images. On the other hand, the latter type comprises features which represent patterns usually complex to be directly observed by experts. Methods that consider both types could combine their benefits to obtain a more complete sight of the skin lesions, supplementing expert skills and potentially improving the overall performance. In this work, texture and shape characteristics from 104 images are described by 166 features from different categories (see Table 2). Each category is addressed in what follows:

3.1.1. Shape and geometric features

Within this category, the following features from the images were extracted:

Approximate entropy. The nonlinear dynamic index Approximate Entropy (*ApEn*) can be used to distinguish useful information from noise [41]. AppEn finds irregularity in pixel patterns and this information can be relevant to determine whether a spatial signal intensity distribution varies with regular pattern (thus being a feature) or is highly random (in that case it can be seen as noise in the context of feature identification).

In order to formalize the definition, given an array of size N and an integer m; $(0 < n \leq N)$, a sequence of real numbers $u = (u_1, u_2, u_3, ..., u_N)$, and a real number $r \geq 0$, let the distance $d(x_i, x_j)$ between two subsequences $x_i = (u_i, u_{i+1}, u_{i+2}, ..., u_{i+m-1})$ and $x_j = (u_j, u_{j+1}, u_{j+2}, ..., u_{j+m-1})$ be defined by Eq. (1).

$$d(x_i, x_j) = \max_{(p=1,2,\cdots,m)} (|u_{(i+p-1)} - u_{(j+p-1)}|). \tag{1}$$

Then let $C_i^m(r)$ be the number of $j \leq (N - m + 1)$, such that $d(x_i, x_j) \leq (N - r + 1)$. Afterwards, $\Phi^m(r)$ is defined by Eq. (2).

$$\Phi^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \log C_i^m(r). \tag{2}$$

Finally, *ApEn* is defined by Eq. (3).

$$ApEn(m, r, N)(u) = \Phi^m(r) - \Phi^{m+1}(r). \tag{3}$$

In this paper, the *ApEn* definition is settled for 1D vectors, using $m = 1$ and $r = 0$. To obtain the original array, a normalized image histogram with 256 bins were used.

Eccentricity. The lesion shape is a measure of growth regularity, and the degree of fitness to a circular shape can be measured to quantify irregularity in growth. This measure is normally quantified by the eccentricity, which is defined by the ratio of the distance between the foci of the fitted ellipse and its major axis length. The output for the Eccentricity ranges from 0, in the case of a circle, to 1 when the lesion is similar to a line segment [42].

Circularity and compactness. The ratio defined by Eq. (4), a.k.a. circularity, is constant and equal to 4π in the case of a circular lesion.

$$Circ = \frac{P^2}{A}, \tag{4}$$

Compactness, in turn, is defined by Eq. (5).

$$Comp = \sqrt{\frac{P}{A}}, \tag{5}$$

where P is the Perimeter and A is the region Area. It should be emphasized that these measures may be useful to distinguish circular objects from oval ones. As both shapes are frequent in biomedical images, the idea behind *Circ* and *Comp* is considered important in the corresponding literature [42].

Two-dimensional moment invariant. The Two-Dimensional Moment Invariant (TDMI) method uses measures to identify shape-based features based on algebraic invariants. Eq. (6) calculates the $(i + j)$th order moments. Let $f(x, y)$ be a density function and $i, j = 0, 1, 2, 3, ..., \infty$.

$$m_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^{i*} y^{j*} f(x, y)\, dx\, dy. \tag{6}$$

For digital image processing, Eq. (6) is replaced by Eq. (7) [43]:

$$m_{ij} = \sum_x \sum_y x^{i*} y^{j*} f(x, y). \tag{7}$$

In Table 3, the central moments until level 3 are defined by Eqs. (8)–(17), where $w = \frac{m_{10}}{m_{00}}$, and $z = \frac{m_{01}}{m_{00}}$. In this table, except for μ_{10} and μ_{01}, all central moments have a physical meaning, as detailed in [43].

Lastly, by using the central moments, seven TDMI measures can be obtained – Eqs. (18)–(24). In these equations, $M_1 = \mu_{21} - \mu_{03}$, $M_2 = \mu_{21} + \mu_{03}$, $M_3 = \mu_{30} + \mu_{12}$, and $M_4 = \mu_{30} - \mu_{12}$.

$$Hu_1 = \mu_{20} + \mu_{02}. \tag{18}$$

$$Hu_2 = (\mu_{20} - \mu_{02})^2 + 4^*\mu_{11}^2. \tag{19}$$

$$Hu_3 = (\mu_{30} - 3^*\mu_{12})^2 + 3^*M_1^2. \tag{20}$$

Table 2
Feature categories extracted from dermoscopic images.

Category	Descriptors	Number
1)	Shape and geometric features	11
2)	Neighborhood Gray-tone Difference Matrix (NGTDM)	5
3)	Haralick's descriptors	13
4)	Fractal dimension based features	3
5)	Laws' texture energy measures	75
6)	Local Binary Patterns (LBP)	59

Table 3
Definition of the central moments.

Equations		Physical meaning
$\mu_{00} = m_{00}$	(8)	Binary objective region
$\mu_{10} = 0$	(9)	–
$\mu_{01} = 0$	(10)	–
$\mu_{20} = m_{20} - w^{*}m_{10}$	(11)	Abscissa variance
$\mu_{11} = m_{11} - z^{*}m_{10}$	(12)	Abscissa covariance
$\mu_{02} = m_{02} - z^{*}m_{01}$	(13)	Ordinate variance
$\mu_{30} = m_{30} - 3^{*}w^{*}m_{20} + 2^{*}w^{2*}m_{10}$	(14)	Abscissa skew intensity
$\mu_{21} = m_{21} - 2^{*}w^{*}m_{11} - z^{*}m_{20} + 2^{*}w^{2*}m_{01}$	(15)	Distribution intensity towards the left side compared to the right side in the abscissa
$\mu_{12} = m_{12} - 2^{*}z^{*}m_{11} - w^{*}m_{02} + 2^{*}z^{2*}m_{10}$	(16)	Distribution intensity towards the lower side compared to the upper side in the ordinate
$\mu_{03} = m_{03} - 3^{*}z^{*}m_{02} + 2^{*}z^{2*}m_{01}$	(17)	Ordinate skew intensity

$$Hu_4 = (\mu_{30} + 3^{*}\mu_{12})^2 + (3^{*}M_2)^2. \tag{21}$$

$$Hu_5 = (\mu_{30} - 3^{*}\mu_{12})^{*}M_3^{*}[M_3^2 - 3^{*}M_2^2] + 3^{*}M_1^{*}M_2^{*}[3^{*}M_3^2 - M_2^2]. \tag{22}$$

$$Hu_6 = (\mu_{20} - \mu_{02})^{*}[M_3^2 - M_2^2] + 4^{*}\mu_{11}^{*}M_3^{*}M_2. \tag{23}$$

$$Hu_7 = 3^{*}M_1^{*}M_3^{*}[M_4^2 - 3^{*}M_2^2] - (\mu_{30} - 3^{*}\mu_{12})^{*}M_2^{*}[3^{*}M_3^2 - M_2^2]. \tag{24}$$

3.1.2. Neighborhood gray-tone difference matrix

Neighborhood Gray-Tone Difference Matrix (NGTDM), a feature extraction method able to measure the difference among grayscale tones in an image, was proposed in [44]. In this method, each image pixel is initially converted into a grayscale tone. Then, the *i*th NGTDM element is given by the difference between gray tones equal to *i* and the grayscale tone averaged across the neighbor pixels in an image represented by a matrix, considering a distance *d*.

The average of the neighbors of the tone located at position (*a, b*) is defined by Eq. (25), where *M* is the matrix of a grayscale image and $W = (2^{*}d + 1)^2$.

$$A(a, b) = \frac{1}{(W-1)} \sum_{p=-d}^{d} \sum_{q=-d}^{d} M(a + p, b + q). \tag{25}$$

Afterwards, each NGTDM element, $S(i)$, is summarized by Eq. (26). Let G_i be the grayscale tone with value *i*.

$$S(i) = \begin{cases} \sum_i |i - A_i|, \ \forall_i \in G_i \\ 0, \text{ if } G_i = 0 \end{cases} \tag{26}$$

Thereby, five features can be measured by NGTDM [44]: busyness (Am_{bus}), coarseness (Am_{coa}), complexity (Am_{com}), contrast (Am_{con}), and texture strength (Am_{tst}). Busyness is a measure of how fast is the change in pixel intensity between one pixel and its neighbors. Coarseness is usually identified with the common visual perception of texture. This leads to there being smaller differences between a gray tone value of a pixel and the average gray tone of its neighboring pixels. Complexity is a measure of the visual information content of the image. Complexity is increased when the basic texture patterns are highly present or even when the pattern has a varying degree of average intensities. Contrast is increased when there is a large difference in the gray tone level of a pixel and its neighbors. Regarding strength, a strong texture is one where the basic patterns are clearly definable and visible. These features are defined by Eqs. (27)–(31). Let $P(i) = \frac{N_i}{n^2}$ be the probability to occur the *i*th grayscale tone in an $N \times N$ image, $n = N - 2d$, $max(G)$ be the highest value of grayscale tone found in the image, and ϵ be a small constant number that prevents a feature to have an infinite value.

$$Am_{bus} = \frac{[\sum_{i=0}^{max(G)} P(i)^{*}S(i)]}{\left[\sum_{i=0}^{max(G)} \sum_{j=0}^{max(G)} (i^{*}P(i) - j^{*}P(j))\right]}. \tag{27}$$

$$Am_{coa} = \left[\epsilon + \sum_{i=0}^{max(G)} P(i)^{*}S(i)\right]^{-1}. \tag{28}$$

$$Am_{com} = \sum_{i=0}^{max(G)} \sum_{j=0}^{max(G)} \left\{\frac{(|i-j|)}{(n^{2*}(P(i) + P(j)))}\right\} {}^{*}\{P(i)^{*}S(i) + P(j)^{*}S(j)\}. \tag{29}$$

$$Am_{con} = \left[\frac{1}{N_g^{*}(N_g - 1)} \cdot \sum_{i=0}^{max(G)} \sum_{j=0}^{max(G)} P(i)^{*}P(j)^{*}(i-j)^2\right] {}^{*}\left[\frac{1}{n^2} \sum_{i=0}^{max(G)} S(i)\right]. \tag{30}$$

$$Am_{tst} = \frac{[\sum_{i=0}^{max(G)} \sum_{i=0}^{max(G)} (P(i) + P(j))^{*}(i-j)^2]}{[\epsilon + \sum_{i=0}^{max(G)} S(i)]}. \tag{31}$$

3.1.3. Haralick's descriptors

Haralick et al. proposed in [45] a set of widely used measures to extract second order texture information from images, already used in the context of medical image processing [40]. In this work, we extracted from the dermoscopic images the 13 Haralick's descriptors implemented in a MATLAB toolbox [46]: Energy (E), Correlation (Cr), Inertia (I), Entropy (En), Inverse Difference Moment (IDM), Sum Average (Sa), Sum Variance (Sv), Sum Entropy (Se), Difference Average (Da), Difference Variance (Dv), Difference Entropy (De), and two information measures of correlation (C1) and (C2).

3.1.4. Fractal dimension

An alternative to measuring the extent of self-similarity of an object at different scales is the fractal dimension [37]. To apply this texture descriptor, one can project a three-dimensional surface from an input image by using a set of skyscrapers, *i. e.*, rectangular columns with a square of edge ϵ on the top [47]. Specifically, the third dimension is computed according to the gray level, such that the higher the gray values inside a $\epsilon \times \epsilon$ window, the higher the corresponding skyscraper is. The fractal dimension is then given by the slope of the plotted $\log(\text{area}(\epsilon)) \times \log(\epsilon)$. As previously done in the context of mammographic image processing in [48,49], the following sets of ϵ values were used to obtain three features: (2, 4, 8), (3, 6, 9), and (3, 5, 7).

3.1.5. Laws' measures

Laws' masks are filters to emphasize level, edge, spot, wave and ripple structures in images. To this end, the author defined five vectors of integer values, each one mapping a different structure [50]. By multiplying pairs of vectors and ensuring rotational invariance, 15 masks were generated. As proposed in [51], after filtering the input image according to each mask, 15 Texture Energy images (TE) are submitted to five first-order statistics: mean, standard deviation, range, skewness, and kurtosis. As a consequence, 75 features (15 masks × 5 statistics) were extracted from each dermoscopic image.

3.1.6. Local binary patterns

Local Binary Patterns (LBP) are a texture operator obtained by labeling to a binary format each image pixel after the comparison of its neighborhoods [52]. The main idea is to use local spatial patterns and grayscale differences to describe a 2D representation. One differential characteristics of LBP, whose interest is central in dermoscopy imaging, is its robustness against illumination variations.

To obtain the LBP sequence for each pixel, Eq. (32) must be applied, where P and R stand for pixel neighborhoods regarding P sampling points on a circle of radius R.

$$LBP_{(P,R)} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \tag{32}$$

where

$$s(x) = \begin{cases} 1, & x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

3.2. Feature selection

The feature categories previously presented provide different views of an image. By combining all of them, it is possible to explore complementarities and achieve a more complete description of biomedical image contents. In addition, positive results were found in dermoscopy after combining features from distinct modalities [26,53].

Despite the mentioned benefits, extracting a large number of image features may also bring disadvantages, such as the presence of redundant and irrelevant features to specific classes. These problems are inherent to the "curse of dimensionality", which computers can experience when learning differentiation models (classifiers) from data described by a high number of descriptors. By applying Feature Selection (FS) algorithms, one tackles this effect by reducing data dimensionality [33]. As an additional result, it is possible to speed up learning algorithms and sometimes improve the discrimination performance.

In this work, different feature groups were tested and specified before applying FS. The idea was to identify relevant dermoscopic characteristics within specific sets of features. Table 4 specifies the four promising feature combinations investigated in this paper.

The selection of a subset of relevant features able to describe well melanocytic clues was conducted by ReliefF [32,34]. In particular, this algorithm was applied to rank features within each of the identified groups. The main advantage of ReliefF over other strictly univariate measures (e.g., information gain and chi-square) is that it takes into account the effect of interacting attributes. The idea of ReliefF and its derivatives is to reward an attribute for having different values on a pair of nearest examples from distinct classes and penalize it for having different values on examples from the same class. In other words, ReliefF values features that support learners to separate examples from distinct classes. For each feature, ReliefF outputs a value w, with large positive w assigned to more important features. In the end, a threshold value t can be employed to obtain a feature subset composed of the t best-ranked features.

3.3. Classification algorithms

The feature subsets obtained from ReliefF application were used to construct models through three ML methods: J48 decision tree, Nearest Neighbors (NN) and Support Vector Machines (SVM). These learning algorithms represent three alternative paradigms: symbolic, lazy learning and statistical, respectively. They were chosen in order to give a wider range of different representations.

Decision Tree (DT) is a learning method that builds symbolic classifiers, which in turn are structured by algorithms based on the divide and conquer strategy. In a tree structure, the classes are represented by a set of rules obtained by the tree derivation. A DT classifier labels a new example by traversing the corresponding tree, starting from the root and ending in a leaf node (a class). Then, the class from the leaf is assigned to the example [54]. The J48 algorithm is a well-known implementation of DT [37].

Nearest Neighbors (NN) is an instance-based learning method used to classify data (test or/and new examples) by computing their similarity with instances (training examples) previously labeled. As NN is based on lazy learning, it does not build a classifier. In other words, the set of training examples itself is considered in the classifier [55].

Support Vector Machines (SVM) is a learning method that constructs one or more hyperplanes in order to separate patterns in a multidimensional space. In particular, a hyperplane specifies margins to separate the input data into classes. A new example is then classified according to its localization in the dimensional space. In this sense, the SVM method can be effectively applied in linear and nonlinear problems. To build nonlinear classifiers, SVM can use kernel functions to enable the representation of input data into multidimensional spaces [56].

3.4. Experimental setting

The experimental setup designed in this paper was organized in four steps, as shown in Fig. 1. As can be seen in this figure, after acquisition and cropping, each dermoscopic image in the dataset was initially converted to the corresponding grayscale version by selecting the channel with the highest entropy (Step 1 – Medical Images Acquisition).

Then, according to Table 4, features based on texture, shape, and local binary patterns were extracted from each one of the preprocessed images.

Regarding Feature Selection with ReliefF, for each feature group previously defined (TSL, TS, TL, and T), a threshold t was established to yield four feature subsets. These subsets were composed of 10%, 20%, 40% and 80% of the best-ranked features (Step 2 – Data Preprocessing). As a result, 16 feature subsets (4 feature groups × 4 threshold settings) were generated. Each subset was then submitted to J48, NN, and SVM for Classification (Step 3 – Classification).

To evaluate the classification performance in terms of Accuracy (Acc), Specificity (SP) and Sensitivity (SE), a 10-fold stratified cross-validation was applied to the image database. In particular, the training folds were submitted for Feature Selection and classifier building, while the corresponding testing folds were considered to evaluate the learner. The results achieved by each classification algorithm were initially

Table 4
Feature combinations and their properties.

Group ID's	Short description	Features					
		Shape and geometric features	NGTDM	Haralick's descriptors	Fractal dimension	Laws' texture energy measures	LBP
TSL	Texture, shape and LBP	✓	✓	✓	✓	✓	✓
TS	Texture and shape	✓	✓	✓	✓	✓	
TL	Texture and LBP			✓	✓	✓	✓
T	Texture			✓	✓	✓	

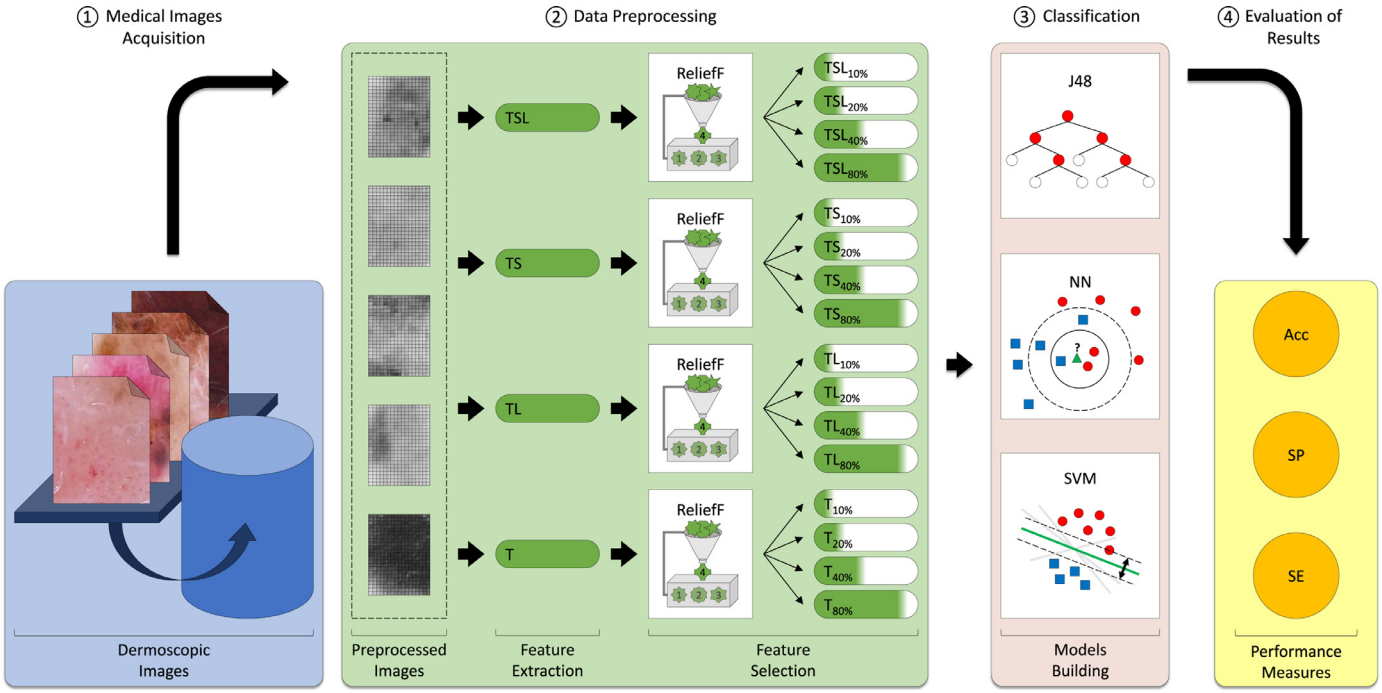

Fig. 1. Experimental setup.

averaged across the folds to resume information from each of the three classic metrics. These averages were then used to assess the performance of evaluated three classification algorithms (Step 4 – Evaluation of Results).

4. Results and discussion

The success of the previously described process of feature extraction and selection is vital to achieving a proper performance of automatic classification, and thus to contribute to the main objective of this work. In fact, the optimization of ML related factors, the selection and tailored combination of the most discriminant features to capture specific clues in melanocytic images, are vital to the effectiveness of the automatic classification. These combinations of suited and optimized methodologies in well-known algorithms will approach medical community to CAD systems. They will also work as a tool to improve the classification output, *i.e.*, the well succeeded diagnosis, and in particular to early diagnosis. The analysis of the results is presented following a classical approach and a Multi-Criteria Decision Analysis (MCDA) to broadly explore different perspectives. The idea is to find the most adequate methodology and criteria to assess ML algorithms performance to be used in dermoscopy.

4.1. Classical analysis

In order to adequately describe the classification performance achieved by each ML algorithm used, and also for comparison purposes with other works (Sections 2 and 4.3), standard metrics were applied. Sensitivity (SE) or true positive rate, measures the proportion of melanomas correctly identified as such; Specificity (SP) or true negative rate, measures the proportion of non-melanomas that are identified correctly as such, and Accuracy (Acc) is defined as the number of correct classifications obtained by the algorithm. As a first attempt to look at the data obtained from the standard metrics for each classifier and feature grouping, a boxplot representation is presented in Fig. 2.

As is possible to observe, these plots show a slight degree of dispersion, thus not pointing to either a prominent arrangement of features with higher scores in terms of SE or Acc, or to a classifier with superior performance.

Nevertheless, this simple graphical analysis allows us to discard some feature arrangements or feature subset cardinalities. In this case, it can be seen in Fig. 2 the reduced representativeness of the TSL10 feature subset for the three ML algorithms in terms of SP and Acc. These poor results soften the relatively high SE values achieved by TSL10 J48 and TSL10 SVM.

Furthermore, the SVM algorithm achieved lower average scores than J48 and NN for all evaluation measures. Another finding is the lowest dispersion achieved by J48 and NN in terms of Acc and SE, respectively, which suggests that they obtained higher stability during the experimental evaluation.

Despite the initial hints provided by this preliminary analysis, it is important to establish a measurable parameter for comparison purposes. Thus, for each learning algorithm, the models built from the 16 feature subsets evaluated in this work were submitted to an unpaired Kruskal–Wallis test (significance level = 1%, p-values in Fig. 2 caption, software GraphPad InStat, GraphPad Software, Inc., US). As a result, a statistical difference was found in just one group (for Acc in J48, p-value= 0.00053) and thus, for this group, a post-hoc Tukey test was conducted. This test pointed to significant statistical differences between TSL10 group and other feature groups (See Fig. 2 caption for details).

The strategy chosen to find the best feature combination enlightens trends in the establishment of the best feature set. It was then found that the problem to address is wider than identifying differences in all pairs of evaluated feature subsets, because no statistical difference was found between best-scored groups. Actually, what is needed is to identify the best among the bests. Accordingly, a strategy to find possible differences among the highest values of SE and Acc was implemented, aiming to discharge differences between groups located above the 3rd quartile. This procedure was done by applying a Harrel-Davis estimator joint with a bootstrap strategy, according to the methodology presented in [57]. Also, in this case, no relevant difference between groups was found. Thus, this approach did not prove to be suitable as a tool to identify the best combination of features and/or algorithm.

Classically, another complementary methodology in ML is the

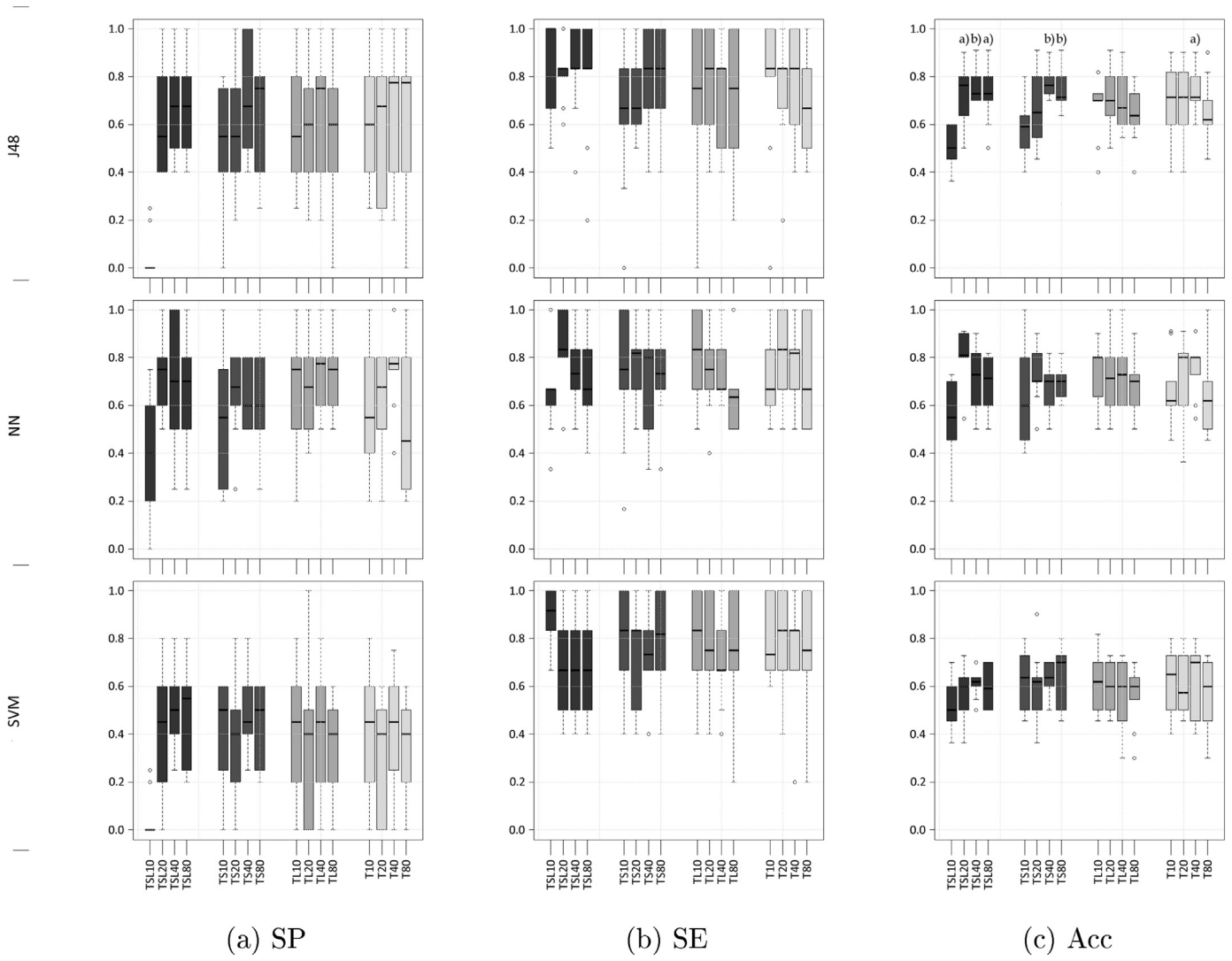

Fig. 2. Results of SP, SE and Acc for the classifiers (SP p-values: J48 = 0.011, NN = 0.101, SVM = 0.036; SE p-values: J48 = 0.683, NN = 0.334, SVM = 0.756; Acc p-values: J48 = 0.00053, NN = 0.021, SVM = 0.762). In J48 Acc plots a) is for a p-value < 0.01 and b) for a p-value < 0.001.

Receiving Operation Characteristic (ROC) curve, used as a procedure to assess the classification performance of a binary learner. Actually, the classifier performance is measured after the calculation of the area under the ROC curve (AUC). This approach was also used in this dataset, generating 16 curves for each ML algorithm. An example is shown in Fig. 3 for the J48 classifier. The Decision Tree was chosen in this illustration due to the low dispersion presented in terms of Acc, a measure not directly used to plot ROC curves.

One can observe in this figure that using the TSL10 attribute subset led to the poorest results, as was also the case for NN and SVM. This finding is consistent with the one obtained in the global methodology previously shown. The AUC approach for the three classification algorithms used in this work presents similar results concerning the selection of the best performance, yielding ROC curves with identical shapes. Accordingly, a Kruskal–Wallis test at 1% of significance was conducted revealing statistical difference for the J48 classifier (p-value $p = 0.009$). Again, a pos-hoc Tukey test showed similar conclusions of the AUC values pointing statistical differences in all feature combinations except for TSL10.

The analysis of the results presented does not point to a superior SE, Acc or AUC performance, either to a classification algorithm or model built from a specific combination of features. In spite of this, the obtained values can be further regarded as promising in the use of ML in daily routine, once the observed performance is suitable regardless the chosen algorithm and the feature subset combination. In fact, after the first optimization step consisting in the choice of the feature subset (type and percentage), the obtained scores are quite in line with standard values in the literature regarding melanoma automatic classification [20,26,28–30].

4.2. Multi-criteria decision analysis

Achieving the objective of identifying the ideal classifier will depend on the type, grouping and amount of selected attributes. Thus, the optimized solution may be dependent on these choices, and therefore a unique optimal solution probably does not exist. To find the best classifier among the best ones is even harder, once it is important to establish a method encompassing multiple relevant variables. Accordingly, the Multi-Criteria Decision Analysis (MCDA) method proposed in [58] can be used.

As the first step to apply this method, it is necessary to choose the performance measures to be simultaneously considered. Besides the AUC, Acc and SE measures averaged across 10 folds in Section 4.1, the Coefficient of Variation (CV) was included to reflect objectively the dispersion inherent to these measures. Different from the averages, which yield values that must be maximized, CV must be minimized.

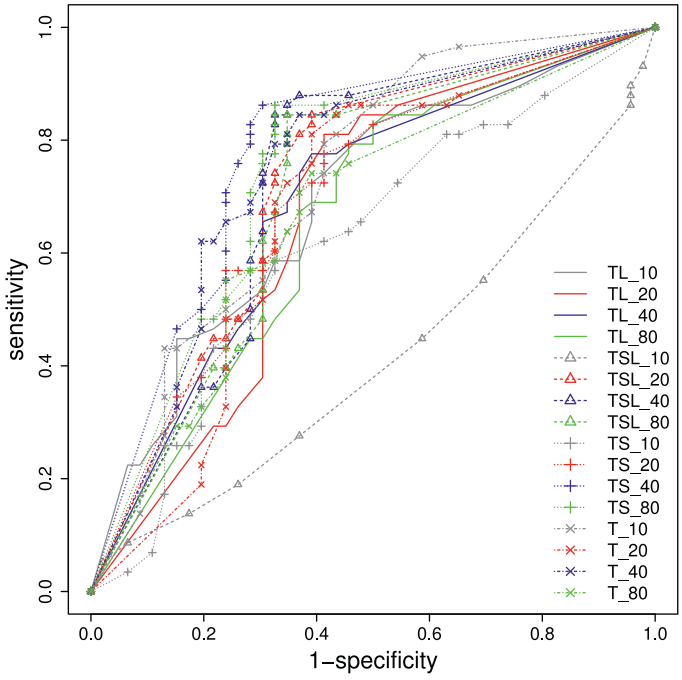

Fig. 3. ROC curves corresponding to the J48 classifiers built from 16 feature subsets.

Thus, we employ in this work the CV complement, CVc = 1 - CV. In the end, this work conducts MCDA by taking into account six variables: mean AUC, mean Acc, mean SE, CVc AUC, CVc Acc and CVc SE.

After configuring the MCDA method proposed in [58] with these variables, it is possible to obtain a single measure to estimate the predictive performance for each evaluated classifier. Thus, the method considers the total area of the six dimension polygon, which in turn is defined from the values achieved by the classifier in the six variables. The higher the Total Area (A_t), the better the corresponding classifier is. Afterwards, all the total values are ranked to find the best ones. To illustrate this idea, Fig. 4 shows the polygons regarding the NN model built from three feature subsets: TSL10 (A_t = 2.67), TSL20 (A_t = 4.55), and T80 (A_t = 3.09).

4.2.1. Overall comparison

Table 5presents the multi-criteria total area values considering the six previously mentioned axes, grouped by the learning algorithm and ranked in decreasing order.

The best subset of features found for J48 and SVM was TS40, while TSL20 stands out as the best-ranked group of features for the NN algorithm. These findings strengthen the choice of using ReliefF, a representative of the filter approach to select features [32–34]. In particular, filter methods stand out due to the independence of any learning algorithm. This issue might be related to the achievement of several good classifiers built from the same feature subset, such as TS40, TSL40 and TSL20 – Table 5. Due to the relative computational efficiency, filters highlight as they are based on properties inherent to data and

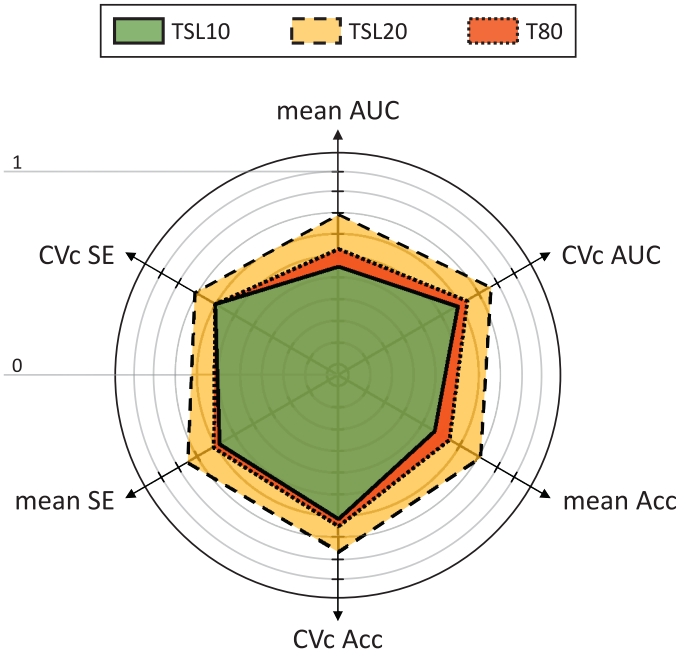

Fig. 4. Polygons summarizing NN predictive performance over TSL10, TSL20, and T80.

Table 5
Total area values for each combination of attribute subset and learning algorithm (grouped by algorithm and ranked in descending order).

J48		NN		SVM	
TS40	4.21	TSL20	4.55	TS40	3.70
TSL40	4.17	T40	4.22	TS80	3.42
TSL20	4.01	TL10	4.07	TSL40	3.38
TS80	4.01	TL40	3.95	TL10	3.35
T40	3.92	TS20	3.94	TSL20	3.34
TSL80	3.69	TS80	3.83	T20	3.32
TL40	3.62	TSL40	3.81	TS10	3.32
TS20	3.46	T20	3.76	TSL10	3.32
T10	3.41	TL20	3.67	T10	3.31
TL20	3.35	TSL80	3.61	TL20	3.26
TSL10	3.34	TS40	3.58	TSL80	3.22
T20	3.31	TL80	3.57	T40	3.16
T80	3.30	T10	3.35	T80	2.97
TL80	3.22	T80	3.09	TS20	2.96
TL10	3.20	TS10	2.84	TL80	2.94
TS10	2.74	TSL10	2.67	TL40	2.94

require only one application before classification by one or more learning algorithms.

ReliefF was applied to rank features according to their relevance, *i.e.*, ability to differentiate non-melanoma from melanoma images. The ranking strategy, in particular, is flexible, as it allows one to apply distinct threshold settings to compose feature subsets with different cardinalities. In this work, four settings were considered – Section 3.4. It should be emphasized that, although ReliefF disregards the removal of redundant features, it outperformed the Correlation-based Feature Selection (CFS) method in preliminary evaluations. Different from ReliefF, CFS evaluates several candidate feature subsets to yield a small subset with relevant and non-redundant features [33].

Feature Selection is welcome in these scenarios, in which a set of feature categories, representing different views of an image, was combined in distinct subsets. In this case, FS was able to remove irrelevant features and reduce data dimensionality while keeping the rich complementarity inherent to the attribute combinations. By analyzing the grouping of features evaluated in Table 5, one finds that the TSL10 subset still presents the worst results for NN, which is consistent with the results of the previous subsection. In fact, TSL10 led to the worst NN classifiers in terms of Acc and AUC, as well as the second worst classifier according to SP. As the average values of these measures represent 3 out of the 6 axes of the MCDA polygon, it did not perform well in the total area measure. Recall that, as indicated in Section 4.1, TSL10 J48 and TSL10 SVM achieved low Acc and AUC values, but high SE values. As a result, the corresponding total area is higher than the TSL10 NN one but lower than the best models based on TSL. On one hand, the TSL group has the highest heterogeneity, as it combines Texture (T), Shape (S) and Local Binary Patterns (L) features. On the other hand, the relatively low number of S and L features in TSL10 – 3 and 1 attributes, respectively, as indicated in Table 6 – might have hindered the corresponding learning performance.

TSL20, in turn, contains a slightly more balanced distribution, with more S and L features – Table 6 – and leads to an NN model with the highest total area – Table 5. This suggests that TSL20 has enough information from T, S, and L to differentiate well non-melanoma from melanoma images.

Although TSL40 and TSL80 also show reasonable distributions, they have higher cardinality and likely more redundant features. In fact, the feature subsets were defined by ReliefF, an algorithm that disregards the removal of redundant features. In this scenario, as the number of features in a subset increases, the probability that redundant features are included also increases.

Texture features are considered in the literature as central to an accurate diagnosis [59–61].This motivated us to concentrate on texture measures – from the 166 implemented features, 96 are considered texture ones – and keep them in all feature subsets. In particular, these features can support the finding of different skin structures, such as the atypical pigment network, with histological correlates and frequent association with melanoma [62–64]. In this work, the analysis of the subsets constituting only with texture features highlighted T40 as the best-scored group regardless of the learning algorithm. By comparing T40 based classifiers, NN achieved the highest score.

Despite the relevance of this feature type, Table 5 shows improved results when combining texture with shape and geometric features, as illustrated by TS40 J48 and TS40 SVM classifiers. In fact, the combination of different feature types, such as shape, color, and texture, is inherent to several dermoscopic approaches, such as the ABCD rule and literature methods [53]. The shape and geometric features, in particular, have played an important role in image classification in different domains, as they mimic visual patterns searched for medical experts [26,40,42]. Table 7 shows the frequency of Texture (T) and Shape (S) features in the TS group.

By taking Table 6 as a reference, one can note in Table 7 that TS exhibits different behavior, obtaining similar distributions for all threshold settings evaluated. When comparing TSL20 NN and TS40 J48, highlighted in terms of total area in Table 5, it is found that the second model needed higher dimensionality, and especially more texture features, to have enough information to build a competitive classifier. Increasing the number of features did not imply in a better J48 learner. Specific J48 properties discussed in Section 4.2.2, such as an additional procedure to select features, may have some influence on these results. Among the classification algorithms, the optimal performance was achieved by building an NN classifier from TSL20. In this case, the NN classifier uses the most heterogeneous group of features (TSL), but with reduced cardinality (only 20 percent). Thus, NN produced the best results when using a heterogeneous feature set of reduced dimension. It should also be emphasized that, different from TS subsets, TSL includes Local Binary Patterns, a texture descriptor already highlighted in the classification of skin lesions [59].

After ordering all the total areas, regardless of the learning algorithm, Table 8 presents the nine best results. The remaining total area values, calculated as illustrated in Fig. 4, can be found in the supplementary material. It should be emphasized that Table 5 gives one a different view, as the total areas are grouped by learning algorithm before ranking.

By picking up the three best combinations in this table – TSL20 NN, T40 NN and TS40 J48 – , it is possible to represent the two best classification algorithms found in the classical and MCDA analyses. Furthermore, the corresponding feature subsets illustrate 3 out of the 4 feature groups investigated in this work: TSL, T, and TS. Thus, in what follows, TSL20 NN, T40 NN and TS40 J48 are analyzed in more detail. To find a brief description of the features present in these subsets,

Table 6
Frequency (and percentage) of each feature type inherent to TSL subsets.

	T	S	L
TSL10	12 (75%)	3 (19%)	1 (6%)
TSL20	23 (70%)	7 (21%)	3 (9%)
TSL40	39 (59%)	12 (18%)	15 (23%)
TSL80	66 (50%)	14 (11%)	52 (39%)

Table 7
Frequency (and percentage) of each feature type inherent to TS subsets.

	T	S
TS10	8 (80%)	2 (20%)
TS20	19 (90%)	2 (10%)
TS40	38 (90%)	4 (10%)
TS80	72 (85%)	13 (15%)

Table 8
Nine best total area values regarding some combinations of attribute subset and learning algorithm (ranked in descending order).

Attribute subset	Learning algorithm	Total area value
TSL20	NN	4.55
T40	NN	4.22
TS40	J48	4.21
TSL40	J48	4.17
TL10	NN	4.07
TSL20	J48	4.01
TS80	J48	4.01
TL40	NN	3.95
TS20	NN	3.94

please refer to the supplementary material.

4.2.2. Selected models (TSL20 NN × T40 NN × TS40 J48)

First, the one-way ANOVA hypothesis test (significance level = 1%) was conducted to compare only the three models in terms of AUC, Acc, and SE. A parametric method was chosen in this case because input data passed by a normality test. However, no statistical difference was found.

Second, the ROC curves associated with the three models were analyzed – Fig. 5. The need to carry out this study arises from the fact that curves with different shapes can be associated with similar AUC values.

In accordance with findings from the AUC analysis, TSL20 NN, T40 NN, and TS40 J48 exhibit similar curves. The curves related to NN are especially close in terms of shape and number of points, *i.e.*, amount of compromises between sensitivity and the rate of false positives (1-specificity). Although the J48 curve consists of more points, further analysis is still required to objectively identify the best combination of attribute subset and learning algorithm.

Regardless the good performance achieved by the three pairs of feature groups and threshold settings focused – TSL20, T40, and TS40–, there are some characteristics specific to each ML algorithm to be analyzed. These specifications can simplify the process, decrease the complexity or even elucidate the procedure to decide the image label

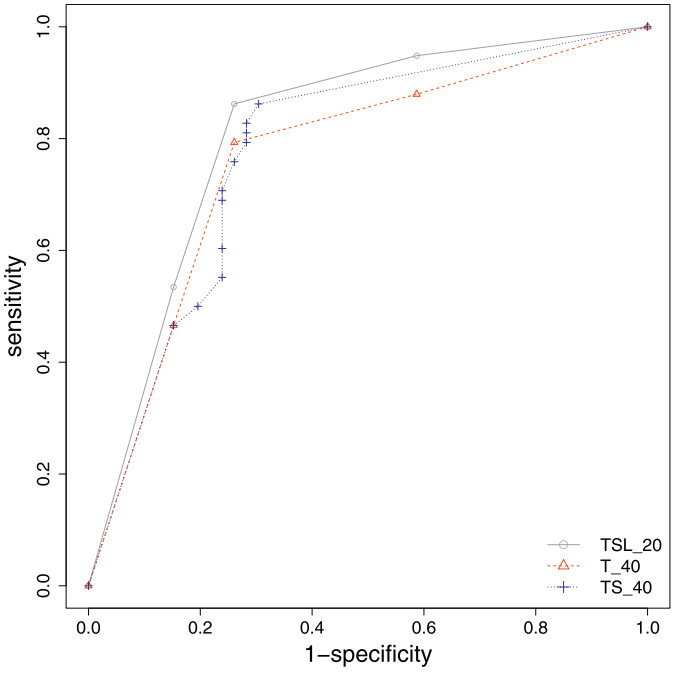

Fig. 5. ROC curves corresponding to the three best combinations (TSL20 NN, T40 NN, and TS40 J48).

(non-melanoma or melanoma). They also can be determinant to choose the learning algorithm to be used, acting as a deciding factor. In fact, the ability to understand the process of decision is relevant, once these methodologies are meant to be used to reinforce the human decision.

A characteristic differentiating NN from J48 consists in the approach used to deal with instances (training examples representing dermoscopic images) [56]. The former algorithm follows the lazy approach, postponing the learning from instances until the last moment before classifying a new example. Although NN saves computational time initially, it spends reasonable time later, during the actual decision. In addition, the lazy algorithm demands efficient techniques for instance storage.

On the contrary, J48 performs eager learning, an alternative approach common to SVM and other traditional ML algorithms. Eager learners construct a generalization model (classifier) as soon as training examples are available. Although the training step is usually expensive, the corresponding model is built only once. Then it is ready to efficiently classify one or more new examples. In a practical scenario, eager J48 is more attractive than lazy NN due to the potential to give a quicker support for medical experts on new images.

Another important specification consists of the algorithm parameters. The parameter simplicity comes in favor of NN, as its two main parameters are relatively intuitive: the number of neighbors k and the dissimilarity measure d. Even if $k = 1$ instances are used, as is the case in this work, satisfactory performance can be achieved. In addition, the Euclidean distance, used by default in the Weka library, is well-known and usually effective to differentiate numerical examples. Finally, although NN and ReliefF conduct distinct tasks, both search for neighbors and use a dissimilarity measure.

J48, in turn, has less intuitive parameters, such as the ones related to tree pruning and the minimum number of instances per leaf [37].

The specifications considered suggest an overall equilibrium between NN and J48. Thus, a specification regarding the number of features used by the model is regarded to finally identify the best learning algorithm in this work. According to Table 9, TS40 J48 appears to be the combination with the highest number of attributes used in the model. However, it is interesting to note that the J48 algorithm is the unique to generate decision trees. Therefore, it performs an internal choice of the attributes enrolled in its construction, supplementing ReliefF filter Feature Selection. This internal process is also known as Embedded Attribute Selection [37]. Thus, despite receiving 42 attributes as input, this additional procedure allows us to use a smaller number for the final model. This is exactly the case in this work, as only 10 attributes out of the 42 initially received as input were kept in the final TS40 J48 model. Thereby, among the three classifiers chosen in this section, the J48 model based on texture and shape features proved to be the one with the lowest dimensionality in the end, while achieving performance similar to the remaining models.

It should be emphasized that, from the 10 features present in the TS40 decision tree, only one represents Shape (S), while the nine remaining attributes consider image Texture (T). As the feature subset originally submitted to this model contains only four S features against 38 T features – Table 7 – , the distribution of these feature types after embedded attribute selection would be also imbalanced. Feeding the tree with more relevant shape and geometric descriptors would yield more balanced distributions. To find a brief description of the 10

Table 9
Percentage and number of features selected for the three models.

Combination of features and learning algorithm	Percentage and number of features selected
TSL20 NN	20% TSL and 33 attrib.
T40 NN	40% T and 36 attrib.
TS40 J48	40% TS and 42 attrib.

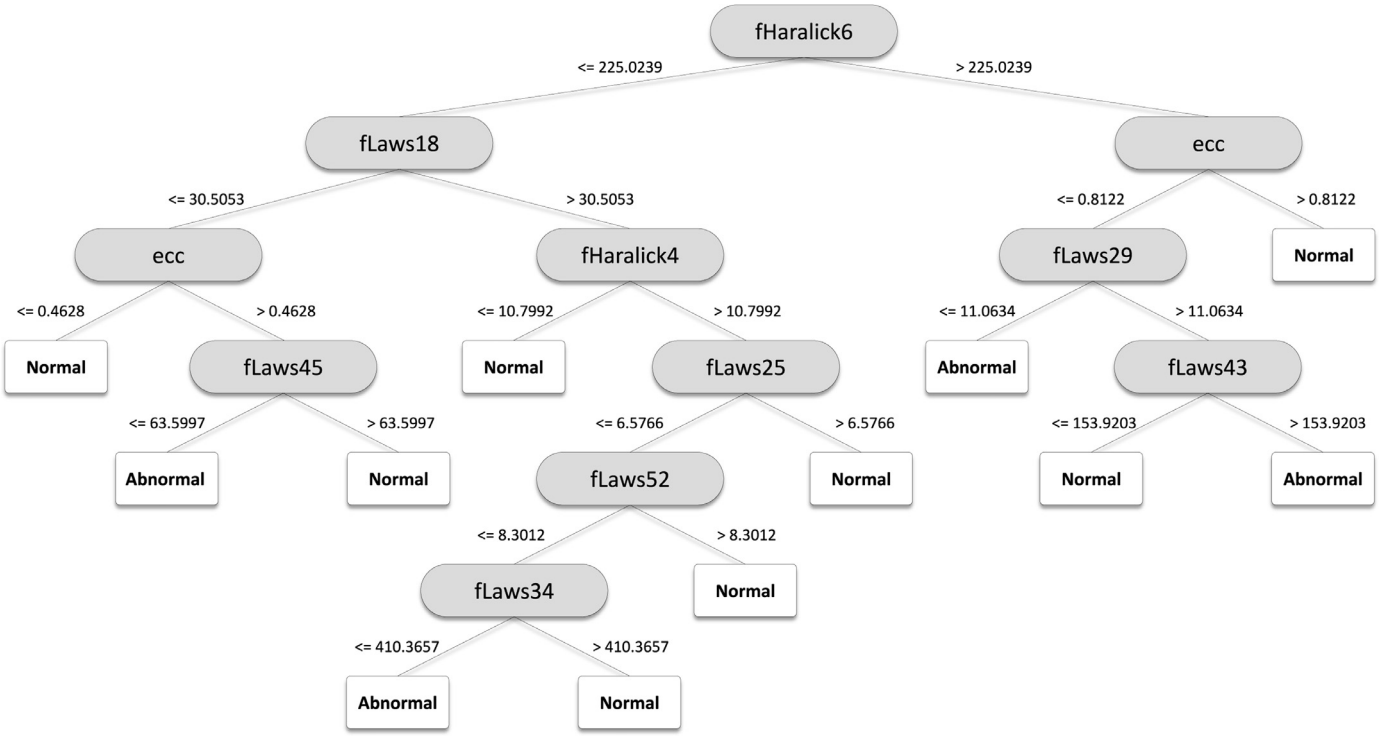

Fig. 6. Decision tree for TS40 J48.

features used by this tree and of the features considered by the TSL20 NN, T40 NN and TS40 J48 models, please refer to the supplementary material.

Another advantage of TS40 J48 is the possibility to verify each of the decision rules generated by the path from the root to each of its leaves. This benefit is especially useful to understand how the learned rules relate the Texture (T) and Shape (S) features inherent to TS40. As a result, an intuitive reasoning could be provided for medical experts interested in finding support to explore combinations of prominent indicators for dermoscopic images. The same achievement is not possible for the NN models, which offer little explanation or insight into the data structure [56]. Fig. 6 shows the decision tree corresponding to TS40 J48.

It is worth emphasizing the variety of attributes employed in this model, which includes attributes of Haralick, Laws masks and also Shape features. By using eccentricity (a shape feature) the model incorporates the symmetry marker of the classical ABCD rule and the inclusion of the Haralick descriptor is coherent with the results of the literature that use these type of descriptors to distinguish typical from an atypical network of a reticular pattern. Laws energy masks, being a filtering derived methodology, are able to capture some clues in which the scale can be determinant. Hence, the choice of the J48 algorithm and TS40 subset reveals to be coherent and suitable to the proposed objectives in this work.

4.3. Numerical comparison with the literature

As previously shown in the state-of-art review presented in Section 2, a large part of the related work seeks to determine a set of features (image descriptors) that enable ML methods to properly classify lesions in dermoscopy images into the malign or benign classes. As various learning algorithms do not present a good performance in the presence of a large number of features [33], a few recent studies already include in their proposals some methods to reduce the number of attributes [26,28].

In this section, we provide a numerical comparison of our best result – represented by the configuration TSL20 using the NN algorithm – concerning those achieved in related works. A rigorous assessment of such outcomes is difficult, as the datasets, the extracted features, and also the enrolled classifiers are distinct among the literature. Moreover, some of the classical metrics are absent in some works, and thus, the results comparison is a demanding task. Nevertheless, the presented outcomes (even though the referred lack of uniformity), allows to give us an idea of how promising the results are. In this context, Table 10 compares the performance of our proposal, which includes Feature Selection, with other models reported in the literature. In this table, empty cells correspond to not reported results.

As can be seen in Table 10, this study achieved results better than or similar to the ones obtained in [20] and [26] with respect to SE and SP respectively, and [28–30] in terms of Acc. It should be taken into account that this work and [31] are the unique ones in which the considered dataset has more melanoma than non-melanoma images – Table 1 – , which can have some influence in the SE and SP results.

One can also note that the numerical results found in this

Table 10
Numerical comparison between our proposal and related methods.

Paper	SE(%)	SP(%)	Acc(%)
[19]	98.00	86.00	–
[20]	85.00	87.00	87.00
[21]	93.00	85.00	–
[22]	91.70	74.50	81.70
[23]	96.00	80.00	–
[24]	92.50	76.30	84.30
[25]	–	–	87.90
[26]	98.46	70.00	–
[27]	97.60	90.50	96.50
[28]	–	–	77.60
[29]	–	–	68.51
[30]	–	–	66.90
[31]	100	99.10	99.70
This paper	86.00	73.00	80.91

paper, [19–24] were estimated according to the 10-fold stratified cross-validation strategy – Table 1. In general, this strategy is a better choice than the holdout approach, used in [25–27,31], and the non-stratified version of the same strategy applied in [28–30]. In fact, *k*-fold stratified cross-validation approximates the dataset class distribution in each fold and performs *k* evaluations, one per fold. As a result, it usually promotes relatively low bias and variance, two important aspects of classifier performance assessment [56]. Finally, comparing estimated performances by using statistical tests, as illustrated by this paper, [25] and [30], can strengthen empirical findings.

In the context of this work, obtaining a classifier with high accuracy is a great challenge, given that the experimental protocol outlined is quite robust, as the analysis of the results were more rigorous than that commonly adopted in the related work. It is also important to point out that the effort performed in the image preprocessing and feature extraction phases, although relatively lower than the ones presented in the literature, was sufficient as our results have shown to be superior or comparable to most of the related work, where the time spent in the aforementioned phases is much higher. Although our goal has focused on selecting important features, our classifiers can still be optimized in terms of parameter tuning.

During the comprehensive review, we also found papers that use a greater set of features or features of greater complexity than those presented in Table 1. Even so, these works provide results very close to the ones reported in the present work, which uses the combination of common and low complexity characteristics allied to widely disseminated ML methods.

5. Conclusion

This paper applies computational methods to assist the dermoscopy domain, illustrating a fruitful interaction between Medicine and Technology. In particular, Machine Learning approaches are used to differentiate normal from melanoma images. In the end, promising experimental results in terms of classical evaluation measures were achieved.

In last decades the number of works related to CAD has increased. Either by the proposal of new algorithms in feature extraction or by the advances regarding new paradigms in Machine Learning (e.g. deep learning), today there exists an extremely large amount of choices to deal with when one intents to use these proposals in a particular case. In addition, these are rather complex algorithms, mostly requiring high computational costs and ideally supported by Information Technology (IT) professionals in the field. Once these methodologies are meant to act as an aid to diagnosis, it is worth optimizing parameters of classical methodologies that proved to be competitive, rather than using methods that perform better but do not explain their reasoning to the medical user. This optimization process guided the objective of this work and will approach medical community to the daily use of Machine Learning in dermoscopy.

The results put forward in this work highlighted the J48 decision tree algorithm, using a multi-criteria performance analysis based on standard measures. Moreover, this model achieved the lowest number of features and the highest comprehensibility. The obtained results motivate further applications of Machine Learning and Feature Selection algorithms to assist CAD, leading this optimization strategy to other medical imaging modalities in which decision appears as a demanding issue.

It is interesting to notice that three algorithms from different learning paradigms were chosen in this work. The idea was to find, with standard parameter values, which method would perform better together with a combination of attributes to describe the problem. This decision allowed us to verify the symbolic method competitiveness in relation to the other algorithms. In particular, it shows benefits in the explicit knowledge of the process, which is an advantage regarding the task of helping the human decision.

In contrast to the literature methods, this proposal invested little effort on image processing tasks. Besides, default parameter values were used without an optimization strategy, and did not concern the imbalance class problem. Even so, the inclusion of FS led to some results similar to or even better than the outcomes reported in related works. It is possible to improve the achieved results by associating FS algorithms with developments in parameter optimization, by joining clinical information (e.g. anatomopathological) and by analyzing micro-texture from the input images [65]. Given the huge range of CAD proposals in dermoscopy (all of them with good and very good ratings), we are now in a position to go beyond the numbers. The good performance in terms of evaluation metrics is taken for granted, the hint now is looking at the process, optimizing the choices and using CAD as a tool to aid the clinical diagnostic task. This was the main concern behind this research paper.

Acknowledgments

We would like to acknowledge EurekaSD: Enhancing University Research and Education in Areas Useful for Sustainable Development - grants EK14AC0037 and EK15AC0264. We would like to thank Araucária Foundation for the Support of the Scientific and Technological Development of Paraná through a Research and Technological Productivity Scholarship for H. D. Lee (grant 534/2014). The Portuguese team was partially supported by Fundação para a Ciência e a Tecnologia (FCT), Portugal, project DERMOPLENO, in the scope of R&D Unit (UID/EEA/50008/2013) through national funds and where applicable co-funded by FEDER - PT2020 partnership agreement. We also would like to thank PGEEC/UNIOESTE through a postdoctoral scholarship for N. Spolaôr, the Brazilian National Council for Scientific and Technological Development (CNPq) through the grant 140159/2017-7 for A. R. S. Parmezan and the Coordination for the Improvement of Higher Education Personnel (CAPES) through a Ph.D. scholarship for J. T. Oliva. These agencies did not have any further involvement in this paper. The authors thank J. G. Martins for his help in LBP implementation.

Conflict of interest This study has been conducted with funding support of EurekaSD (FCT and FEDER-PT2020), Araucária Foundation, PGEEC/UNIOESTE, CNPq, and CAPES.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.knosys.2018.05.016.

References

[1] W.C.S. Cho, Latest discoveries and trends in translational cancer research: highlights of the 2008 annual meeting of the american association for cancer research, Technol. Cancer Res. Treat. 4 (7) (2008) 269–277.
[2] C.A.M.L. Porta, Skin Cancers - Risk Factors, Prevention and Therapy, InTech, Rijeka, 2011.
[3] A.C. Society, Cancer facts and figures, 2016, URL: https://goo.gl/anZaLQ(accessed02.05.17).
[4] R.S. Stern, Prevalence of a history of skin cancer in 2007: results of an incidence-based model, Arch. Dermatol. 146 (3) (2010) 279–282, http://dx.doi.org/10.1001/archdermatol.2010.4.
[5] S.B. Edge, C.C. Compton, The american joint committee on cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM, Ann. Surg. Oncol. 17 (6) (2010) 1471–1474, http://dx.doi.org/10.1245/s10434-010-0985-4.
[6] R.H. Johr, Dermoscopy: alternative melanocytic algorithms - the ABCD rule of dermatoscopy, menzies scoring method, and 7-point checklist, Clin. Dermatol. 20 (3) (2002) 240–247, http://dx.doi.org/10.1016/S0738-081X(02)00236-5.
[7] P. Carli, V. De Giorgi, E. Crocetti, F. Mannone, D. Massi, A. Chiarugi, B. Giannotti, Improvement of malignant/benign ratio in excised melanocytic lesions in the 'dermoscopy era': a retrospective study 1997–2001, Br. J. Dermatol. 150 (4) (2004) 687–692, http://dx.doi.org/10.1111/j.0007-0963.2004.05860.x.
[8] C. Massone, A.D. Stefani, H.P. Soyer, Dermoscopy for skin cancer detection, Curr. Opin. Oncol. 17 (2) (2005) 147–153.
[9] M.E. Vestergaard, P. Macaskill, P.E. Holt, S.W. Menzies, Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-

analysis of studies performed in a clinical setting, Br. J. Dermatol. 159 (3) (2008) 669–676, http://dx.doi.org/10.1111/j.1365-2133.2008.08713.x.
[10] K. Doi, Computer-aided diagnosis in medical imaging: historical review, current status and future potential, Comput. Med. Imaging Graph. 31 (4-5) (2007) 198–211, http://dx.doi.org/10.1016/j.compmedimag.2007.02.002.
[11] R.A. Castellino, Computer aided detection (CAD): an overview, Cancer Imaging 5 (1) (2005) 17–19, http://dx.doi.org/10.1102/1470-7330.2005.0018.
[12] R. Goldenberg, D. Eilot, G. Begelman, E. Walach, E. Ben-Ishai, N. Peled, Computer-aided simple triage (CAST) for coronary ct angiography (CCTA), Int. J. Comput. Assist. Radiol. Surg. 7 (6) (2012) 819–827, http://dx.doi.org/10.1007/s11548-012-0684-7.
[13] H. Aihara, S. Saito, H. Inomata, D. Ide, N. Tamai, T. Ohya, T. Kato, S. Amitani, H. Tajiri, Computer-aided diagnosis of neoplastic colorectal lesions using real-time numerical color analysis during autofluorescence endoscopy, Eur. J. Gastroenterol. Hepatol. 25 (4) (2013) 488–494, http://dx.doi.org/10.1097/MEG.0b013e32835c6d9a.
[14] Y. Kominami, S. Yoshida, S. Yoshida, S. Tanaka, Y. Sanomura, T. Hirakawa, B. Raytchev, T. Tamaki, T. Koide, K. Kaneda, K. Chayama, Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy, Gastrointest. Endosc. 83 (3) (2016) 643–649, http://dx.doi.org/10.1016/j.gie.2015.08.004.
[15] J.T. Oliva, H.D. Lee, N. Spolaôr, C.S.R. Coy, F.C. Wu, Prototype system for feature extraction, classification and study of medical images, Expert Syst. Appl. 63 (C) (2016) 267–283, http://dx.doi.org/10.1016/j.eswa.2016.07.008.
[16] A.M. Scaranelo, R. Eiada, K. Bukhanov, P. Crystal, Evaluation of breast amorphous calcifications by a computer-aided detection system in full-field digital mammography, Br. J. Radiol. 85 (1013) (2012) 517–522, http://dx.doi.org/10.1259/bjr/31850970.
[17] M. Breuninger, B. van Ginneken, R.H.H.M. Philipsen, F. Mhimbira, J.J. Hella, F. Lwilla, J. van den Hombergh, A. Ross, L. Jugheli, D. Wagner, K. Reither, Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: a validation study from sub-saharan africa, PLoS ONE 9 (9) (2014) 1–10, http://dx.doi.org/10.1371/journal.pone.0106381.
[18] M. Machado, J. Pereira, R. Fonseca-Pinto, Classification of reticular pattern and streaks in dermoscopic images based on texture analysis, J. Med. Imaging 2 (4) (2015) 044503, http://dx.doi.org/10.1117/1.JMI.2.4.044503.
[19] C. Barata, J.S. Marques, J. Rozeira, The role of keypoint sampling on the classification of melanomas in dermoscopy images using bag-of-features, in: J.M. Sanches, L. Micó, J.S. Cardoso (Eds.), Proceedings of the 6th Iberian Conference on Pattern Recognition and Image Analysis, IbPRIA, Springer Berlin Heidelberg, Funchal, Madeira, Portugal, 2013, pp. 715–723, , http://dx.doi.org/10.1007/978-3-642-38628-2_85.
[20] C. Barata, J.S. Marques, J. Rozeira, Evaluation of color based keypoints and features for the classification of melanomas using the bag-of-features model, in: G. Bebis, R. Boyle, B. Parvin, D. Koracin, B. Li, F. Porikli, V. Zordan, J. Klosowski, S. Coquillart, X. Luo, M. Chen, D. Gotz (Eds.), Proceedings of the 9th International Symposium on Advances in Visual Computing (ISVC), Springer Berlin Heidelberg, Rethymnon, Crete, Greece, 2013, pp. 40–49, , http://dx.doi.org/10.1007/978-3-642-41914-0_5.
[21] C. Barata, J.S. Marques, T. Mendonça, Bag-of-features classification model for the diagnose of melanoma in dermoscopy images using color and texture descriptors, Proceedings of the International Conference Image Analysis and Recognition, (2013), pp. 547–555.
[22] C. Barata, J.S. Marques, M.E. Celebi, Improving dermoscopy image analysis using color constancy, Proceedings of the IEEE International Conference on Image Processing, (2014), pp. 3527–3531, http://dx.doi.org/10.1109/ICIP.2014.7025716.
[23] C. Barata, M. Ruela, M. Francisco, T. Mendonça, J.S. Marques, Two systems for the detection of melanomas in dermoscopy images using texture and color features, IEEE Syst. J. 8 (3) (2014) 965–979, http://dx.doi.org/10.1109/JSYST.2013.2271540.
[24] C. Barata, M.E. Celebi, J.S. Marques, Improving dermoscopy image classification using color constancy, IEEE J. Biomed. Health Inform. 19 (3) (2015) 1146–1152, http://dx.doi.org/10.1109/JBHI.2014.2336473.
[25] R. Kaur, P.P. Albano, J.G. Cole, J. Hagerty, R.W. LeAnder, R.H. Moss, W.V. Stoecker, Real-time supervised detection of pink areas in dermoscopic images of melanoma: importance of color shades, texture and location, Skin Res. Technol. 21 (4) (2015) 466–473, http://dx.doi.org/10.1111/srt.12216.
[26] M. Rastgoo, R. Garcia, O. Morel, F. Marzani, Automatic differentiation of melanoma from dysplastic nevi, Comput. Med. Imaging Graph. 43 (2015) 44–52, http://dx.doi.org/10.1016/j.compmedimag.2015.02.011.
[27] O. Abuzaghleh, M. Faezipour, B.D. Barkana, A comparison of feature sets for an automated skin lesion analysis system for melanoma early detection and prevention, Long Island Systems, Applications and Technology, (2015), pp. 1–6, http://dx.doi.org/10.1109/LISAT.2015.7160183.
[28] 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, http://dx.doi.org/10.1109/TMI.2015.2506270.
[29] 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 on Hybrid Artificial Intelligent Systems, HAIS, Springer International Publishing, Seville, Spain, 2016, pp. 427–438, , http://dx.doi.org/10.1007/978-3-319-32034-2_36.
[30] M. Pérez-Ortiz, A. Sáez, J. Sánchez-Monedero, P.A. Gutiérrez, C. Hervás-Martínez, Tackling the ordinal and imbalance nature of a melanoma image classification problem, Proceedings of the International Joint Conference on Neural Networks, (2016), pp. 2156–2163, http://dx.doi.org/10.1109/IJCNN.2016.7727466.
[31] S. Yang, B. Oh, S. Hahm, K.-Y. Chung, B.-U. Lee, Ridge and furrow pattern classification for acral lentiginous melanoma using dermoscopic images, Biomed. Signal Process. Control 32 (2017) 90–96, http://dx.doi.org/10.1016/j.bspc.2016.09.019.
[32] J. Demšar, Algorithms for subsetting attribute values with relief, Mach. Learn. 78 (3) (2010) 421–428, http://dx.doi.org/10.1007/s10994-009-5164-0.
[33] H. Liu, H. Motoda, Computational methods of feature selection, Chapman & Hall/CRC, Boca Raton, 2007.
[34] I. Kononenko, Estimating attributes: Analysis and extensions of RELIEF, European Conference on Machine Learning, (1994), pp. 171–182.
[35] A. Boer, K. Nischal, A growing online resource for learning dermatology and dermatopathology, Indian J. Dermatol. Venereol. Leprol. 73 (2) (2007) 138–140, http://dx.doi.org/10.4103/0378-6323.31909.
[36] G. Argenziano, I. Zalaudek, Dermoscopy: a new perspective, Dermatol. Pract. Concept. 1 (1) (2011) 57–58.
[37] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, third ed., Morgan Kaufmann, Burlington, 2011.
[38] J. Pereira, R. Fonseca-Pinto, Segmentation strategies in dermoscopy to follow-up melanoma: combined segmentation scheme, Online J. Sci. Technol. 5 (3) (2015) 56–61.
[39] J. Pereira, A. Mendes, C. Nogueira, D. Baptista, R. Fonseca-Pinto, An adaptive approach for skin lesion segmentation in dermoscopy images using a multiscale local normalization, in: J.-P. Bourguignon, R. Jeltsch, A.A. Pinto, M. Viana (Eds.), Proceedings of the International Conference on Dynamics, Games and Science and Advanced School Planet Earth, Springer International Publishing Switzerland, 2013, pp. 537–545, , http://dx.doi.org/10.1007/978-3-319-16118-1_29.
[40] M. Elter, A. Horsch, CADx of mammographic masses and clustered microcalcifications: a review, Med. Phys. 36 (6) (2009) 2052–2068, http://dx.doi.org/10.1118/1.3121511.
[41] S.M. Pincus, W.-M. Huang, Approximate entropy: Statistical properties and applications, Commun. Stat. Theory Methods 21 (11) (1992) 3061–3077, http://dx.doi.org/10.1080/03610929208830963.
[42] K. Najarian, R. Splinter, Biomedical Signal and Image Processing, CRC Press, Boca Raton, 2006.
[43] J.S. Noh, K.H. Rhee, Palmprint identification algorithm using Hu invariant moments and Otsu binarization, Proceedings of the ACIS International Conference on Computer and Information Science, (2005), pp. 94–99, http://dx.doi.org/10.1109/ICIS.2005.97.
[44] M. Amadasun, R. King, Textural features corresponding to textural properties, IEEE Trans. Syst. Man. Cybern. 19 (5) (1989) 1264–1274, http://dx.doi.org/10.1109/21.44046.
[45] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. SMC-3 (6) (1973) 610–621, http://dx.doi.org/10.1109/TSMC.1973.4309314.
[46] S. Gupta, Haralick texture features Matlab toolbox version 0.1b, 2007, URL: https://goo.gl/DCIxi4(accessed02.05.17).
[47] A.R. Dominguez, A.K. Nandi, Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection, Comput. Med. Imaging Graph. 32 (4) (2008) 304–315, http://dx.doi.org/10.1016/j.compmedimag.2008.01.006.
[48] C.B. Caldwell, S.J. Stapleton, D.W. Holdsworth, R.A. Jong, W.J. Weiser, G. Cooke, M.J. Yaffe, Characterisation of mammographic parenchymal pattern by fractal dimension. Phys. Med. Biol. 35 (2) (1990) 235–247.
[49] N. Spolaôr, M.Z. Nascimento, A.C. Lorena, Evaluating feature selection techniques to differentiate brest tissues using computational systems (avaliando técnicas de seleção de características para a diferenciação de nódulos mamários por sistemas computacionais), Brazilian Congress of Health Informatics, (2010), pp. 1–6.
[50] K.I. Laws, Texture energy measures, DARPA Image Understanding Workshop, (1979), pp. 47–51.
[51] A.N. Karahaliou, I.S. Boniatis, S.G. Skiadopoulos, F.N. Sakellaropoulos, N.S. Arikidis, E.A. Likaki, G.S. Panayiotakis, L.I. Costaridou, Breast cancer diagnosis: analyzing texture of tissue surrounding microcalcifications, IEEE Trans. Inf. Technol. Biomed. 12 (6) (2008) 731–738, http://dx.doi.org/10.1109/TITB.2008.920634.
[52] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987, http://dx.doi.org/10.1109/TPAMI.2002.1017623.
[53] W.V. Stoecker, K. Gupta, B. Shrestha, M. Wronkiewiecz, R. Chowdhury, R.J. Stanley, J. Xu, R.H. Moss, M.E. Celebi, H.S. Rabinovitz, M. Oliviero, J.M. Malters, I. Kolm, Detection of basal cell carcinoma using color and histogram measures of semitranslucent areas, Skin Res. Technol. 15 (3) (2009) 283–287, http://dx.doi.org/10.1111/j.1600-0846.2009.00354.x.
[54] J.R. Quinlan, Simplifying decision trees, Int. J. Man-Mach. Stud. 27 (3) (1987) 221–234, http://dx.doi.org/10.1016/S0020-7373(87)80053-6.
[55] E. Alpaydin, Introduction to Machine Learning, third ed., MIT Press, 2010.
[56] J. Han, Data Mining: Concepts and Techniques, fifth ed., Morgan Kaufmann Publishers, San Francisco, 2011.
[57] J.C. Wilcox, A. Barbottin, D. Durant, M. Tichit, D. Makowski, Farmland birds and arable farming, a meta-analysis, in: E. Lichtfouse (Ed.), Sustainable Agriculture Reviews: Volume 13, Springer International Publishing, Cham, 2014, pp. 35–63, , http://dx.doi.org/10.1007/978-3-319-00915-5_3.
[58] A.R.S. Parmezan, H.D. Lee, F.C. Wu, Metalearning for choosing feature selection algorithms in data mining: proposal of a new framework, Expert Syst. Appl. 75

(2017) 1–24, http://dx.doi.org/10.1016/j.eswa.2017.01.013.
[59] V. González-Castro, J. Debayle, Y. Wazaefi, M. Rahim, C. Gaudy-Marqueste, J.-J. Grob, B. Fertil, Texture descriptors based on adaptive neighborhoods for classification of pigmented skin lesions, J. Eletron. Imaging 24 (6) (2015) 061104–1–061104–8, , http://dx.doi.org/10.1117/1.JEI.24.6.061104.
[60] B. Shrestha, J. Bishop, K. Kam, X. Chen, R.H. Moss, W.V. Stoecker, S. Umbaugh, R.J. Stanley, M.E. Celebi, A.A. Marghoob, G. Argenziano, H.P. Soyer, Detection of atypical texture features in early malignant melanoma, Skin Res. Technol. 16 (1) (2010) 60–65, http://dx.doi.org/10.1111/j.1600-0846.2009.00402.x.
[61] T. Tanaka, S. Torii, I. Kabuta, K. Shimizu, M. Tanaka, H. Oka, Pattern classification of nevus with texture analysis, Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society, (2004), pp. 1459–1462, http://dx.doi.org/10.1109/IEMBS.2004.1403450.
[62] G. Argenziano, H.P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G.D. Rosa, G. Ferrara, R. Hofmann-Wellenhof, M. Landthaler, S.W. Menzies, H. Pehamberger, D. Piccolo, H.S. Rabinovitz, R. Schiffner, S. Staibano, W. Stolz, I. Bartenjev, A. Blum, R. Braun, H. Cabo, P. Carli, V.D. Giorgi, M.G. Fleming, J.M. Grichnik, C.M. Grin, A.C. Halpern, R. Johr, B. Katz, R.O. Kenet, H. Kittler, J. Kreusch, J. Malvehy, G. Mazzocchetti, M. Oliviero, F. Ozdemir, K. Peris, R. Perotti, A. Perusquia, M.A. Pizzichetta, S. Puig, B. Rao, P. Rubegni, T. Saida, M. Scalvenzi, S. Seidenari, I. Stanganelli, M. Tanaka, K. Westerhoff, I.H. Wolf, O. Braun-Falco, H. Kerl, T. Nishikawa, K. Wolff, A.W. Kopf, Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet, J. Am. Acad. Dermatol. 48 (5) (2003) 679–693, http://dx.doi.org/10.1067/mjd.2003.281.
[63] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, M. Delfino, Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis, Arch. Dermatol. 134 (12) (1998) 1563–1570, http://dx.doi.org/10.1001/archderm.134.12.1563.
[64] S. Yadav, K.A. Vossaert, A.W. Kopf, M. Silverman, C. Grin-Jorgensen, Histopathologic correlates of structures seen on dermoscopy (epiluminescence microscopy), Am. J. Dermatopath. 15 (4) (1993) 297–305.
[65] K. Seetharaman, Image retrieval based on micro-level spatial structure features and content analysis using full range gaussian markov random field model, Eng. Appl. of Artif. Intell. 40 (Supplement C) (2015) 103–116, http://dx.doi.org/10.1016/j.engappai.2015.01.008.