ARCHIVIO ISTITUZIONALE DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

A Feasibility Study for a Persistent Homology-Based k-Nearest Neighbor Search Algorithm in Melanoma Detection

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Ferri, M., Tomba, I., Visotti, A., Stanganelli, I. (2017). A Feasibility Study for a Persistent Homology-Based k-Nearest Neighbor Search Algorithm in Melanoma Detection. JOURNAL OF MATHEMATICAL IMAGING AND VISION, 57(3), 324-339 [10.1007/s10851-016-0680-6].

Availability:
This version is available at: https://hdl.handle.net/11585/559775 since: 2016-08-29

Published:

DOI: http://doi.org/10.1007/s10851-016-0680-6

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

15 September 2026

This is the final peer-reviewed accepted manuscript of:

Ferri, M., Tomba, I., Visotti, A. *et al.* A Feasibility Study for a Persistent Homology-Based *k*-Nearest Neighbor Search Algorithm in Melanoma Detection. *J Math Imaging Vis* 57, 324–339 (2017).

The final published version is available online at: **https://doi.org/10.1007/s10851-016-0680-6**

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

A feasibility study for a persistent homology based **k**-Nearest Neighbor search algorithm in melanoma detection

Massimo Ferri
Università di Bologna - Bologna, Italy, massimo.ferri@unibo.it

Ivan Tomba
CA-MI S.r.l., Pilastro - Parma, Italy, tomba.ivan@gmail.com

Andrea Visotti
CA-MI S.r.l., Pilastro - Parma, Italy, andrea.visotti@gmail.com

Ignazio Stanganelli
IRCCS-IRST, Meldola - Forlì, Italy and
Università di Parma - Parma, Italy, igstanga@tin.it

Abstract

Persistent Homology is a fairly new branch of Computational Topology which combines geometry and topology for an effective shape description of use in Pattern Recognition. In particular it registers through "Betti Numbers" the presence of holes and their persistence while a parameter ("filtering function") is varied. In this paper, some recent developments in this field are integrated in a k-Nearest Neighbor search algorithm suited for an automatic retrieval of melanocytic lesions. Since long, dermatologists use five morphological parameters (A = Asymmetry, B = Boundary, C = Color, D = Diameter, E = Elevation or Evolution) for assessing the malignancy of a lesion. The algorithm is based on a qualitative assessment of the segmented images by computing both 1 and 2-dimensional Persistent Betti Numbers functions related to the ABCDE parameters and to the internal texture of the lesion. The results of a feasibility test on a set of 107 melanocytic lesions are reported in the section dedicated to the numerical experiments.

Biographies

Massimo Ferri is full professor of Geometry at the University of Bologna. Coming from topology of low-dimensional manifolds, he grew interested in the applications of geometry and topology to robotics and to pattern recognition. His present research is in persistent topology and applications.

Ivan Tomba received his Ph.D. degree in Mathematics from the University of Bologna in 2013. He is currently software developer for CA-MI S.r.l. and inventor (with Andrea Visotti) of a patent pending video-dermatoscope with a software for an automatic image retrieval of melanocytic lesions. His research interests include Inverse Problems and geometrical-topological methods for shape comparison and applications in Pattern Recognition.
Andrea Visotti has been the R&D Coordinator of CA-MI S.r.l. for two years, during which he filed two patent applications and a model utility as inventor. He is also inventor (with Ivan Tomba) of a patent pending video-dermatoscope with a software for an automatic image retrieval of melanocytic lesions. Currently he is consultant for CA-MI S.r.l. and co-founder and R&D Coordinator of IBD S.r.l. to reverse engineer and innovate the total spectrum of biomedical machines with most of the functionalities but at a much lower cost.
Ignazio Stanganelli is director of the Skin Cancer Unit IRCCS Istituto Scientifico Romagnolo per la cura e lo studio dei tumori (IRST) and associate professor of Dermatology Department at University of Parma. His research is mainly focused on the development of new non-invasive technologies such as digital dermoscopy, computer aided diagnosis and confocal laser microscopy in the skin tumors diagnosis, especially in the secondary prevention of melanoma.

1 Introduction

The incidence of malignant melanoma in fair-skinned patients has increased dramatically in most parts of the world over the past few decades.
According to the World Health Organization (2006), 132 thousands melanomas occur globally each year and more than 65 thousands people a year worldwide die for melanoma [33].
According to the American Cancer Society, 6.640 males and 3.300 females are expected to die of melanoma in 2015 and the incidence of melanoma has increased 15 times in the last 40 years [50]. In the United Kingdom and Europe, a similar increase in the incidence has been seen. In the European Union, the incidence of invasive melanoma is of 9 cases over 100.000 every year [17]. This incidence tends to increase with the latitude, with a larger prevalence among the populations with a lower pigmentation (from 12 to 17 cases over 100.000 every year) than among the Mediterranean countries (from 3 to 5 cases over 100.000 every year). Moreover, melanoma is the second most frequent neoplasm among people under 40 years old, underlining that the problem is often related to the patients' young age [43].
Since the prognosis of melanoma depends almost entirely on tumor thickness, detection of early melanoma is crucial for the survival of patients [28, 46]. The widely used acronym ABCDE (asymmetry, irregular borders, multiple colors, diameter > 6 mm, enlarging lesion) contains the primary clinical criteria for diagnosing suspected skin melanoma.
However, the early phase of malignant melanoma is difficult to identify by naked eye because cutaneous malignant melanoma can share many clinical features with an atypical naevus. Several studies have described diagnostic accuracy rates ranging from 50

to 75%, indicating a need for additional diagnostic tools [12, 36, 34].
Dermoscopy (or epiluminescence microscopy) is the examination of skin lesions with a dermatoscope. This traditionally consists of a magnifier, a non-polarised light source, a transparent plate and a liquid medium between the instrument and the skin. It is a noninvasive method that allows the in vivo evaluation of colors and microstructures of the epidermis, the dermoepidermal junction and the papillary dermis not visible to the naked eye. These structures are specifically correlated to histologic features. The identification of specific diagnostic patterns related to the distribution of colors and dermoscopy structures can better suggest a malignant or benign pigmented skin lesion. The use of this technique provides a valuable aid in diagnosing pigmented skin lesions [35, 39, 38].
Because of the complexity involved, this methodology is reserved for experienced clinicians [48, 47, 45]. In [49] is reported that dermoscopy assessment is more accurate than clinical evaluation by naked eye for the diagnosis of cutaneous melanoma ($OR = 15.6$, $p = 0.016$). In this study, the mean sensitivity in the diagnosis of melanoma was 74% for the examination by naked eye and 90% for dermoscopy.
The acquisition of dermoscopic images has also stimulated the automatization of the diagnostic process: many rather successful computer programs have been implemented to automatically analyze melanocytic lesions, providing the dermatologist with a support in the diagnosis phase [4, 9, 34, 41, 44]. These programs are generally based on a software that given an image of a skin lesion from a test set (query image) and a training set of already classified images performs the following four steps on the query:

1) removal of artifacts (hair, bubbles, ...);

2) segmentation: isolation of the skin lesion from its background;

3) analysis: feature extraction;

4) classification: assignment of the query to one of the classes in which the training set is divided.

Many different techniques such as decision trees, support vector machine, extreme learning machine, statistical methods, the k-nearest neighbor (k-NN) and many others have been developed to perform the classification step, with a very wide range of results in terms of sensitivity, specificity and diagnostic accuracy: see [34] for a summary of the results obtained by many different algorithms in the last two decades; see also [14] and [19] for a comparison between automatic and human performance.
In the analysis step, most of the programs keep into account the traditional ABCDE parameters used by dermatologists: Asymmetry (of boundary, texture, and color), Boundary (irregularity and dishomogeneity), Color (presence of several colors), Dimension, and Evolution. An original method for comparing in a qualitative, yet precise way two skin lesions is based on the mathematical theory of Persistent Homology (originated as Size Theory). This method was introduced in the early 90's to compare homeomorphic topological spaces and has considerably grown in popularity since it has been proven to provide both theoretical and computational tools for shape comparison [2, 3]. The main idea is to take into account topological shape features with respect to some geometric properties conveyed by real functions defined on the shape itself. Typically, a

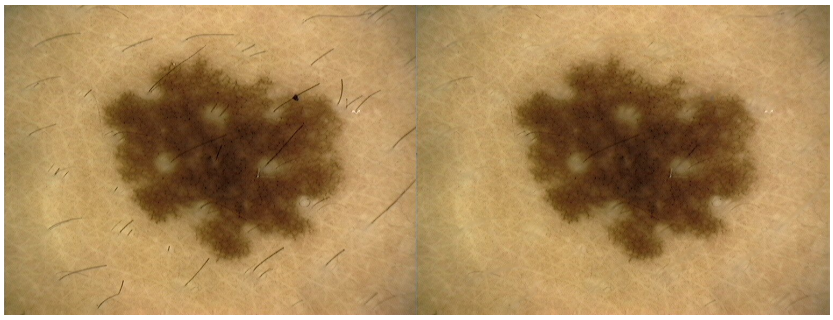
shape is represented by a pair $(\mathscr{X}, \varphi)$, where $\mathscr{X}$ is a topological space and $\varphi : \mathscr{X} \to \mathbb{R}$ is a continuous real-valued function called filtering (or measuring) function. Persistent Homology allows to associate to the pair $(\mathscr{X}, \varphi)$ some shape descriptors (Persistent Betti Number Functions) which register quantitatively the behaviour of the filtering function. This allows to compare pairs of the type $(\mathscr{X}, \varphi)$ and $(\mathscr{Y}, \psi)$, with $\mathscr{X}$ and $\mathscr{Y}$ homeomorphic, by computing a distance between their Persistent Betti Number (PBN) functions. The Research Group of Vision Mathematics of the University of Bologna has obtained interesting results using these techniques, as proved in a series of papers [9, 18, 44].
Recent developments in the field of Persistent Homology led the authors to the implementation of a *k*-NN algorithm that integrates some recent discoveries for the retrieval of skin lesions.
The choice of a *k*-NN algorithm permits retrieval and visualization of the "most similar" cases to those at hand. This aspect partly resembles the medical reasoning and allows a dermatologist to directly compare unknown lesions with other known skin lesions. This can provide an advantage in areas where black-box models are inadequate, which is exactly the case of many borderline skin lesions. Moreover, retrieval provides a valid support in the diagnosis decision, without necessarily producing an automatic classification, which tends to influence the user's judgement. On the other hand, as stated in [34], the major difficulty of *k*-NN algorithms lies in the definition of a metric that measures the distance between data items. Another problem is that whereas classification can focus on the search for some characteristic features, retrieval needs a formalization of the concept of similarity which has to be much more adherent to the intuitive concept with this name: a poor performance in retrieval is as risky as in classification, but much more evident just by sight.
The authors believe that Persistent Homology may be an adequate technical tool in the transition from a risk-assessment output based on classification to image retrieval. However, since this transition is delicate, as a first step the new algorithm has been evaluated as a classifier. Thus, the aim of this work is to:

- describe the new algorithm;
- present the numerical results obtained in a preliminary feasibility test made on 107 dermoscopic melanocytic lesions;
- evaluate which of the new technical tools may be worth of further investigation.

The paper is organized as follows: section 2 describes the pre-processing steps (hair removal and segmentation); section 3 presents the analysis step; section 4 discusses the problems of the choice of the metrics and classification; section 5 is dedicated to the numerical experiments and their discussion and section 6 to the conclusions.

2 Pre-processing

The pre-processing phase consists of three distinct steps.
As a first pre-processing optional step, the software performs the removal of the hair present on the lesion. The areas that contain the hair are identified by means of an



(a) Original image. (b) Output of the hair-removal algorithm.

Figure 1: Example of the hair-removal algorithm.

erosion/dilation with straight-line segments in a similar way to that proposed in [29]. This process terminates with the creation of a Boolean mask containing the hair zones and some noise. For removing the noise, the mask is then treated in a similar way to that described in [26], obtaining a new Boolean mask. At last, the pixels inside this mask are replaced with the average over the neighboring pixels not belonging to the mask. Figure 1 shows the result obtained in a practical example, where the picture on the left is the original image acquired by the dermatoscope and the picture on the left is the output of the hair removal algorithm.

In the second step, the aim of the algorithm is to identify the discrete (digital) versions of the two topological spaces that will be associated to the skin lesions: the boundary of the lesion and the lesion itself. This is achieved by a segmentation algorithm. The segmentation determines the edge of the lesion, identified by a closed curve that separates the lesion's area from the background. It is performed as follows.

The image is cropped by the user, for framing the skin lesion in an optimal way. Then, a level set evolution type algorithm is applied.

Level set methods were introduced by Osher and Sethian in [37] to study front propagations. Standard formulations consider a region $\Omega \subset \mathbb{R}^2$, a closed, continuous curve $\mathbf{\Gamma}(-,t) : [0,1] \to \Omega$ evolving in time $t \in [0,+\infty)$ and a real-valued function $\Phi : \Omega \times [0,+\infty)$, usually called the level set function, such that $\Phi(\mathbf{\Gamma}(q,t),t) = 0$ for every $q \in [0,1]$ and for every $t \geq 0$. In these formulations, the level set function typically develops irregularities during its evolution, which may cause numerical errors and eventually destroy the stability of the evolution [42]. Therefore, a numerical remedy, called reinitialization, is typically applied to periodically replace the degraded level set function with a signed distance function [42]. However, the practice of reinitialization affects numerical accuracy and efficiency in an undesirable way [42, 23]. For this reason, in a series of papers [31, 32, 30], C. Li et al. proposed a new variational level set formulation (known as DRLSE - Distance Regularized Level Set Evolution) in which the regularity of the level set function is intrinsically maintained during the level set evolution. The level set evolution is derived as the gradient flow that minimizes an energy functional with a distance regularization term and an external energy that drives the

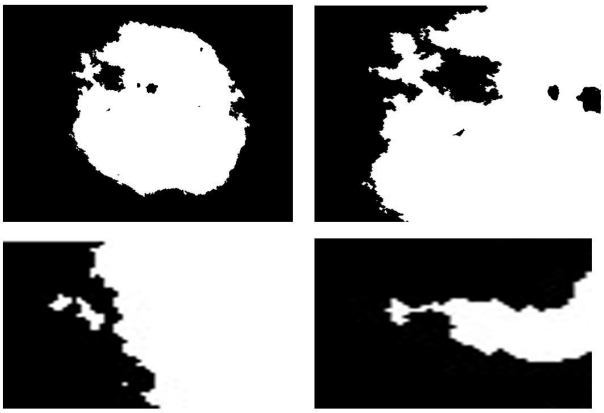


Figure 2: An output mask of the DRLSE algorithm (top left). The top-right figure enlights some holes in the main connected component of the mask; the bottom-left figure shows some smaller connected components of the mask; the bottom-right figure shows that the contour curve is not simple, due to the presence of an isthmus of boundary pixels which connects two distinct connected components of the inside.

motion of the zero level set toward desired locations. The distance regularization effect eliminates the need for reinitialization and thereby avoids its induced numerical errors. This led the authors to the implementation of the DRLSE method described in [30], with satisfying results, in the sense that the algorithm computed the segmentation in real-time and the outcome was in line with the clinicians' expectations for the images considered in the numerical experiments.

The output of the DRLSE algorithm is a Boolean mask containing all those pixels on which the level set function computed by the DRLSE is negative. This mask will represent the digital space associated to the skin lesion and the border of the mask will be the digital space associated to its border.

In general, in skin lesion images, the DRLSE algorithm produces masks constituted by a large connected component containing the most of the skin lesion and by several very small connected components near the boundary of the main component: a typical example is shown in Figure 2. Unfortunately, dealing with objects with a different number of connected components does not fit one of the main assumptions one usually makes when comparing two sets using persistent homology, i.e. that the underlying topological spaces are homeomorphic (cf. [3, 2]).

In the continuous 2-dimensional model, the simplest way to deal with homeomorphic spaces with homeomorphic boundaries is to consider the standard situation in which the sets to be compared are compact topological spaces whose boundary is a Jordan curve in $\mathbb{R}^2$. The Jordan Curve Theorem ensures that the spaces are homeomorphic.

In the discrete setting, the analogous of the Jordan Curve Theorem is a classic result of digital topology usually called the Discrete Jordan Curve Theorem [15, 27]. The main point behind the Discrete Curve Theorem is to use different connectivity notions for the boundary and the inside, in order to avoid the so-called connectivity paradoxa [40].

The digital version of the continuous properties of the Jordan Curve Theorem can be summarized as follows.

Definition 1 *Let I and J be respectively the number of rows and columns of an image* I*. A function*

$$\mathsf{M} : \{0,...,I-1\} \times \{0,...,J-1\} \longrightarrow \{-1,0,1\}$$

is said to be a mask of I *suitable for comparisons if it satisfies the following properties:*

- $P1$ M *is surjective. In this case, a pixel* (i,j) *will be said to belong to the background if* $\mathsf{M}(i,j) = 1$*, to the inside if* $\mathsf{M}(i,j) = -1$ *and to the boundary if* $\mathsf{M}(i,j) = 0$*.*
- $P2$ *The background is* 4*-connected and contains the border of the window (i.e. all the pixels such that* $i = 0$*,* $i = I-1$*,* $j = 0$*, or* $j = J-1$*).*
- $P3$ *The inside is* 4*-connected and is* 4*-disconnected from the background.*
- $P4$ *The boundary is* 8*-connected and if a pixel belongs to the boundary, it has two and only two* 8*-neighbors in the boundary.*

In general, the output mask of the DRLSE algorithm is not suitable for comparisons (cf. Figure 2).
This motivates the introduction of a third step in the pre-processing phase, which transforms the DRLSE output mask into a suitable for comparisons mask as follows.

2.1 Cleanmask algorithm

Let

$$\mathsf{M}_0 : \{0,...,I-1\} \times \{0,...,J-1\} \longrightarrow \{-1,1\}$$

be the mask obtained as the output of the DRLSE algorithm. The Cleanmask Algorithm will be defined as the consecutive execution of the following actions:

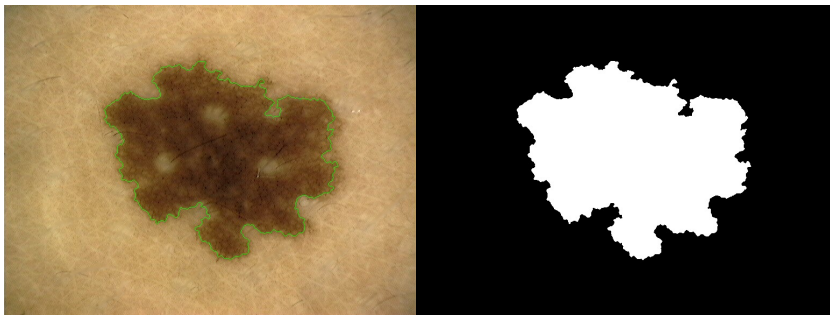
- **A1** copy the mask M_0 on a mask M setting to the background all the pixels at the border of the window;
- **A2** identify a maximal connected component (in the sense of 4-connectivity) of $\mathsf{M}^{-1}(\{-1\})$ by means of a Depth First Search (DFS) and set every pixel not belonging to this maximal connected component to the background ($\mathsf{M}(i,j) = 1$);
- **A3** identify the connected components (in the sense of 4-connectivity) of the resulting background by means of a DFS and in particular the component containing the boundary of the image. Set all the pixels not belonging to this component to the inside ($\mathsf{M}(i,j) = -1$);
- **A4** for every pixel (i,j) s.t. $\mathsf{M}(i,j) = -1$, if (i,j) has at least a 4-neighbor belonging to the background, set $\mathsf{M}(i,j) = 0$;
- **A5** identify a maximal connected component of $\mathsf{M}^{-1}(\{-1\})$ and set all the pixels of $\mathsf{M}^{-1}(\{-1\})$ belonging to the other components to the background;

A6 for every (i, j) s.t. $\mathsf{M}(i, j) = 0$, if (i, j) has no neighbors belonging to the inside, set $\mathsf{M}(i, j) = 1$.

The output of the Cleanmask Algorithm satisfies the properties $P1$-$P4$ thanks to the Discrete Jordan Curve Theorem.
The connectivity of the mask and of the boundary (achieved by actions **A2** and **A3** respectively) are essential assumptions of the persistent homology based algorithm described in the following sections, since they ensure that the distances between pairs of segmented skin lesions remain finite (see also section 3 and [3, 2]).
In addiction, filling the holes of the mask (action **A3**) is important from the clinical point of view: it allows to restore those parts of the skin lesion that were eliminated by the DRLSE algorithm because the skin was not dark enough there, but that might be of some significance for the diagnosis. For example, a light bluish mark on a dark pigmented skin lesion may be a sign of melanoma.
Actions **A5** and **A6** ensure that the boundary is a simple, closed curve in the digital plane. These operations are not actually necessary for the persistent-homology algorithm to be well defined, since only 0-homology will be considered. Moreover, the fact that some parts of the skin lesion segmented by the DRLSE algorithm may be deleted due to these actions is a possible cause of concern (for example, in the bottom-right picture of Figure 2 the western region of that portion of the mask will be eliminated). However, according to the experiments, in practice only very small parts of the lesion are deleted in this phase. On the other hand, this provides contour curves that are topologically equivalent in the discrete setting (indeed, they are Jordan curves in the digital plane in the sense of [25]), resembling the continuous setting in a better way and allowing potential extensions of the comparisons to homology groups of degree > 0.
Summing up, the software performs hair-removal and computes the segmentation of the skin lesion by means of the DRLSE algorithm corrected by the Cleanmask Algorithm, to obtain a mask which is suitable for comparisons. The final result of the pre-processing phase for the image of Figure 1 is shown in Figure 3, which shows the segmented image (left picture) and the corresponding suitable for comparisons mask (right picture). The green line in the left picture shows the simple closed boundary contour of the mask.

3 Analysis

To the aim of comparing images, the main tool the algorithm relies on is the theory of Size Functions (SFs) [21], extended in Persistent Homology [7, 16] as Persistent Betti Number functions, also in the form of Persistence Diagrams (PDs).
Essentially, the PBN registers the behavior of the filtering function by using Morse theory [20]. The main idea is to take into account topological shape features with respect to some geometric properties conveyed by real functions defined on the shape itself. In formal settings, that means a shape is represented by a size pair $(\mathscr{X}, \phi)$, where $\mathscr{X}$ is a topological space and $\phi : \mathscr{X} \longrightarrow \mathbb{R}$ is a continuous real-valued function called filtering (or measuring) function.
Size Theory was introduced in the early 90's to store quantitatively some qualitative



(a) Pre-processing output. (b) Its mask.

Figure 3: The final result of the pre-processing algorithm for the image of Figure 1 and the corresponding mask, which satisfies properties $P1$-$P4$.

information about shapes. In particular, the (1-dimensional) 0-PBN (or SF) of the size pair $(\mathscr{X}, \phi)$ is the function $\ell_{(\mathscr{X},\phi)}$ defined on the half-plane

$$\Delta^+ := \{(u,v) \in \mathbb{R}^2 \mid v > u\}$$

s.t. for every point $(u,v) \in \Delta^+$ $\ell_{(\mathscr{X},\phi)}(u,v)$ is the number of path-connected components of the sub-level set

$$\mathscr{M}_v := \{x \in \mathscr{X} \parallel \phi(x) \leq v\}$$

which contain at least one point of the corresponding sub-level set $\mathscr{M}_u$.
In Persistent Homology theory, k-PBNs with generic integers $k \geq 0$ have also been studied in detail (see e.g. [16] for a formal definition of the k-PBNs and [5] for a simple example in which 1-PBNs play an essential role). However, since in practical applications k-PBNs with $k > 0$ are rarely used as they are not easy to calculate, in this paper only 0-PBNs (later, for simplicity, PBNs) will be considered.
Figure 4 shows the computation of a PBN function: in this situation, $\mathscr{X}$ is the (black) Jordan curve in the left picture and the function ϕ is the Euclidean distance from the point C (center of the dotted circles). The picture on the right represents the PBN of this size pair: the numbers on the various coloured areas of Δ^+ indicate the values that the PBN function assumes on these zones.
In general, a PBN function can be seen as a linear combination (with natural numbers as coefficients) of characteristic functions associated to the (possibly unbounded) triangles laying on Δ^+ (cf. [22]). The bounded triangles are of the form $\{(u,v) \in \Delta^+ : \alpha \leq u < v < \beta\}$, while the unbounded ones are of the form $\{(u,v) \in \Delta^+ : \eta \leq u < v\}$. Hence, a simple and compact representation is obtained by associating the set $\{(u,v) \in \Delta^+ : \alpha \leq u < v < \beta\}$ to the point (α,β), and the set $\{(u,v) \in \Delta^+ : \eta \leq u < v\}$ to the point at infinity (η,∞). The points of a formal series having finite coordinates are called proper cornerpoints, while the ones with a coordinate at infinity are said to be cornerpoints at infinity or cornerlines.
Persistent Betti Number functions can be compared by using a suitable matching (also

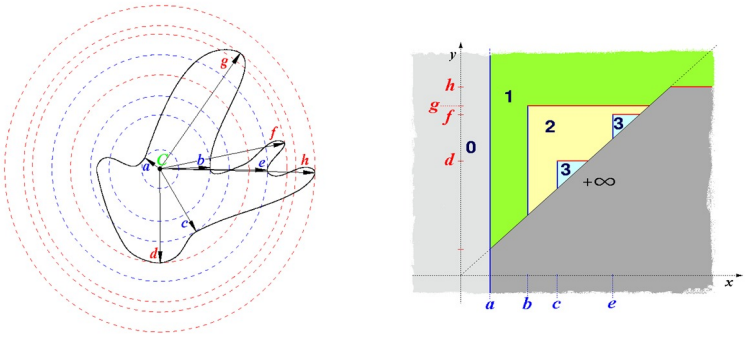


Figure 4: The computation of the 0-PBN function of the size pair represented by a Jordan Curve (black line in the left picture) with the Euclidean distance from the point C as filtering function. The dotted circles show the radial values that influence the generation of the cornerpoints of the PBN function (right picture).

called bottleneck) distance [10, 11], and have been widely applied to Pattern Recognition and dermoscopy-related problems (see [3, 6, 13, 9, 18, 44]). The matching distance d_{match} can be seen as a measure of the cost of transporting the cornerpoints of a PBN function into the cornerpoints of another one. Formally, let ℓ_1 and ℓ_2 be two PBN functions, and let C_1 and C_2 be the multisets of their cornerpoints counted with their multiplicities and augmented by adding the points of the diagonal $\Delta := \{(u,v) \in \mathbb{R}^2 : \ u = v\}$ counted with infinite multiplicity. Denoting by $\overline{\Delta^*}$ the set $\overline{\Delta^+}$ extended by the points at infinity of the kind (a,∞), $a \in \mathbb{R}$, i.e. $\overline{\Delta^*} := \overline{\Delta^+} \cup \{(a,\infty), a \in \mathbb{R}\}$, the matching distance is defined as

$$d_{match}(\ell_1,\ell_2) := \min_{\sigma} \max_{P \in C_1} \delta(P, \sigma(P)), \tag{1}$$

where σ varies among all the bijections from C_1 to C_2 and the distance $\delta((u,v),(u'v'))$ between two points (u,v) and (u',v') in $\overline{\Delta^*}$ is defined by

$$\min\{\max\{|u-u'|,|v-v'|\}, \max\{\frac{v-u}{2}, \frac{v'-u'}{2}\}\},$$

with the convention about ∞ that $\infty - v = v - \infty = \infty$ when $v \neq \infty$, $\infty - \infty = 0$, $\infty/2 = \infty$, $|\infty| = \infty$, $\min\{c,\infty\} = c$, $\max\{c,\infty\} = \infty$. A consequence of this convention is that if ℓ_1 and ℓ_2 have a different number of cornerlines, their matching distance is automatically ∞. Indeed, it is easy to see that the number of cornerlines is equal to the number of connected components of the space $\mathscr{X}$, meaning that the distance between two spaces with a different number of connected components would be equal to ∞.
The pseudometric δ measures the pseudo-distance between two points (u,v) and (u',v') as the minimum between the cost of moving one point onto the other and the cost of moving both points onto the diagonal, with respect to the max-norm and under the assumption that any two points of the diagonal have vanishing pseudo-distance.

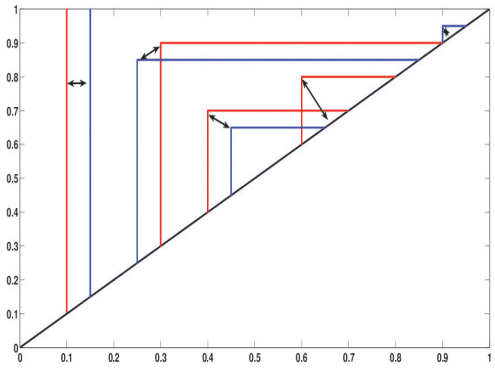


Figure 5: The matching distance between two PBNs in the half-plane is obtained by searching the best matching between the two sets of cornerpoints.

As different PBN functions may in general have a different number of cornerpoints, d_{match} allows a proper cornerpoint to be matched to a point of the diagonal: this matching can be interpreted as the deletion of a proper cornerpoint.
Figure 5 visualizes the computation of a matching distance between two very simple PBN functions, both constituted by a cornerline and three cornerpoints (the first PBN is represented in red, the second in blue): after matching the cornerlines, the minimum cost is achieved by matching together two couples of points that are sufficiently close to each other and by matching the remaining cornerpoints to the diagonal (i.e. deleting them).
The success of the matching distance is mainly due to its stability properties. In order to understand these properties, it is common to introduce the natural pseudo-distance between two size pairs $(\mathscr{X},\phi)$ and $(\mathscr{Y},\psi)$, with $\mathscr{X}$ and $\mathscr{Y}$ homeomorphic, as the quantity:

$$d((\mathscr{X},\phi),(\mathscr{Y},\psi)) := \min_{f} \max_{x\in\mathscr{X}} |\phi(x) - \psi(f(x))|, \tag{2}$$

where f varies among all possible homeomorphisms between $\mathscr{X}$ and $\mathscr{Y}$. The natural pseudo-distance is a measure of the dissimilarity between the two size pairs, but it is computable only in a very few cases. It turns out that the matching distance is stable with respect to perturbations of the measuring function in the following sense:

$$d_{match}(\ell_1,\ell_2) \leq d((\mathscr{X},\phi),(\mathscr{Y},\psi)). \tag{3}$$

Moreover, it has been proved that d_{match} provides the best lower bound for the natural pseudo-distance, in the sense that any other distance between size functions would furnish a worse bound. For details about these results, the interested reader is addressed to [3, 2, 10, 11] and the references therein.
In the last decade, the theory of PBNs has developed extensively: in particular, a multi-dimensional PBNs theory is now available. In this context, filtering functions with values in $\mathbb{R}^k$ ($k \geq 1$) defined on the space $\mathscr{X}$ generate k-dimensional PBN functions

and the generalization of the domain Δ^+ lies in $\mathbb{R}^{2k}$. The multidimensional matching distance has been proved to improve the lower bounds achieved by the 1-dimensional PBNs separately (Proposition 4 in [3]) and in the case $k = 2$ a stable algorithm to compute the 2-dimensional matching distance between 2-dimensional PBNs is now available [2, 5].

To the authors' knowledge, no attempts to apply these results in the dermatology field have been made before this work. Here, the computation of 1 and 2-dimensional PBNs given an *RGB* skin lesion's image is performed following the approach described below.

Since the algorithm deals with digital images, it is important to discuss Persistent Homology and Size Theory in the digital world. The discrete version of the theory substitutes the topological space $\mathscr{X}$ with a graph $G = (V, E)$, the function $\phi : \mathscr{X} \longrightarrow \mathbb{R}$ with a function $\phi : V \longrightarrow \mathbb{R}$ and the concept of topological connectedness with the usual connectedness notion for graphs [8]. Indeed, the size pair $(\mathscr{X}, \phi)$ can be obtained from the discrete data by taking a geometric realization $|V|$ of the graph that encodes the non-zero elements of the image as nodes and the neighborhood adjacency among the digital points as edges. Depending on the number of neighbors that may be adjacent to a point, the connectivity (i.e. the number of edges) of the graph depends on the number of neighbors that are admitted to be adjacent to a point in the 2D image (i.e., 8-, 6- or 4-neighborhoods). Then, the function $\phi : |V| \to \mathbb{R}$ is a piecewise linear function first defined on the nodes of the graph and then linearly extended to the edges. The pair (G, ϕ) is usually called the Size Graph. The PBN functions are obtained from the Size Graphs by means of the algorithm described in [8]. This algorithm directly computes the multiset of cornerpoints and cornerlines that completely determines a PBN function in $O(n \log n + m\alpha(2m+n, n))$ operations, where n and m are the number of vertices and edges in the Size Graph, respectively, and α is the inverse of the Ackermann function [1].

The matching distance in a single half-plane is computed by means of a Hopcroft-Karp type algorithm, whose complexity is $O(p^{2.5})$, where p is the number of cornerpoints taken into account for the comparison.

The evaluation of the 2-dimensional matching distance between 2-dimensional PBN functions requires the computation of infinitely many 1-dimensional matching distances, one for each point of a certain rectangle of $\mathbb{R}^2$, whose dimensions depend on the filtering function [2, 3, 5]. Since in practice it is impossible to consider every point of this rectangle, the algorithm proposed in [2] computes a stable approximation of this 2-dimensional matching distance on a grid defined on the rectangle: the denser the grid, the smaller the error. Indeed, the bound for this error is an input of this algorithm: in order to evaluate the performances of the 2-dimensional algorithm in this dermatology application, in the experiments considered in this paper this bound has always been chosen small enough so that this error could be considered negligible.

Focusing on the case of digital images of pigmented skin lesions, the mask obtained by the DRLSE algorithm is turned into a suitable for comparisons mask as described in section 2. Then, 2 different graphs are computed: one (hereafter called the Boundary Graph) contains only the boundary pixels and the other one (hereafter called the Global Graph) contains all pixels of the mask. The connectivity notion on these graphs is given by the 8-connectivity for the case of the Boundary Graph and by the 6-connectivity for

the case of the Global Graph. It should be noted that, in line with the continuous setting case, the number of cornerlines computed by the algorithm in [8] is equal to the number of connected components of the Size Graph. This explains the necessity to consider Size Graphs with only one connected component, in order to avoid cases in which the matching distance is equal to ∞: actions **A2** and **A3** in the Cleanmask Algorithm of section 2 are necessary to achieve this goal.
In order to compute PBN functions, the following filtering functions have been defined:

1 on the Boundary Graph, the Euclidean distance from the barycenter of the mask;

2 on the Global Graph:

- the blue (B), greeen (G) and red (R) channels, defined on each pixel (node of the graph) as the values (in the range $[0,255]$) assumed by the *RGB* image on that pixel;
- the light intensity variations, which for a pixel (i,j) is defined as

$$TV((i,j)) := \sum_{(i',j')} |In((i,j)) - In((i',j'))|,$$

  where $In = (R+G+B)/3$ is the light intensity and (i',j') varies in the set of pixels of the mask 8-connected to (i,j);
- the excess-colours $ExcB$, $ExcG$, $ExcR$, defined by formulas of the type

$$ExcB := (2B-G-R+510)/4;$$

- the three colour-differences $R-G$, $R-B$ and $G-B$, defined by formulas of the type

$$R-B := (R-B+255)/2.$$

This provides 11 different 1-dimensional Size Graphs (G,ϕ) to compute the corresponding 1-dimensional PBNs.
To the aim of computing 2-dimensional matching distances, 2-dimensional Size Graphs have been formed by coupling some 1-dimensional filtering functions. In particular, the following 2-dimensional filtering functions have been defined on the Global Graph:

- colour channel couplings:
  (B,G), (B,R), (G,R);
- colour channel and opposite colour difference couplings:
  $(B,R-G)$, $(G,R-B)$, $(R,G-B)$;
- excess-colours and opposite colour difference couplings:
  $(ExcB,R-G)$, $(ExcG,R-B)$, $(ExcR,G-B)$.

Since the Global Graph is usually too large, in order to reduce the computational costs, prior to these operations, the images' resolution has been reduced by splitting them into small squared blocks and by taking averages of the three color channels on these blocks. Different sizes of the blocks have been considered, ranging from 3×3 to $8 \times$

8, as shown in Figure 6. For further details, see the numerical experiments.
At last, several different numerical parameters related to the ABCDE analysis of the skin lesions have also been computed, in order to compare the performances of the PBNs with some standard features usually adopted in classification-type algorithms. To this aim, the suitable for comparisons mask of a skin lesion image has been used to calculate the principal axes of inertia of the lesion, its orientation, the smallest rectangle containing the mask (with the same orientation) and its inscribed ellipse. These pieces of information have been used to compute the following parameters.

(A) Asymmetry parameters:

- $Sym1$: the percentage of pixels of the mask whose symmetric with respect to the major axis of inertia lies in the mask;
- $Sym2$: the percentage of pixels of the mask whose symmetric with respect to the line perpendicular to the major axis of inertia passing through the barycenter lies in the mask;
- $Sym3$: the percentage of pixels of the mask whose symmetric with respect to the barycenter lies in the mask;
- $Sym := (Sym1 + Sym2 + Sym3)/3$.

(B) Parameters describing properties of the border:

- The form factor of the mask, defined as $4\pi A/P^2$, where A is the number of pixels of the mask and P is the number of pixels of the boundary;
- Haralick's Circularity (CH), a standard measure of the compactness of a digital space [24];
- Ellipticity (Elt) defined as $4A/(\pi LW)$, where L (respectively W) is length of the longest (respectively shortest) side of the smallest rectangle containing the mask;
- Eccentricity (Ecc), defined as L/W, where L and W are as above.

(C) The Colour Histogram ($Histo$) of the segmented lesion, based on 64 different colours. A Colour Histogram is a vector $\mathbf{H} \in \mathbb{R}^{64}$ whose entries $H_j \in [0,100]$ contain the percentages of pixels of the mask for which the j-th colour is the closest (in the Euclidean metric sense) to their actual RGB values.

(D) The diameter ($Diam$) of the lesion, computed as the maximal distance between two pixels of the boundary. Note that using this feature to compare skin lesions requires the magnification and resolution to be fixed for all the analyzed images.

(E) The colour Entropy of the skin lesion ($Entr$), a measure of the colour variations among the four regions in which the axes of inertia divide the segmented image (*inertial regions*). This is the only non-standard parameter considered in the ABCDE-analysis and is computed as follows: let $\mathbf{H}_0$,...,$\mathbf{H}_4$ be the Colour Histograms of a segmented skin lesion related respectively to the global mask and

to its four inertial regions. For every $i = 0, ..., 4$ let H_{ij} be the entries of the i-th Colour Histogram. Then the Colour Entropy of the segmented skin lesion is defined as the quantity

$$Entr := \sum_{i_2 > i_1 > 0} \sum_{H_{0j} > 0.05} \frac{|H_{i_2 j} - H_{i_1 j}|}{H_{0j}}.$$

4 Classification

PBNs have a standard structure, the one of superimposed triangles. This has an important outcome, in that the relevant information can be condensed in the vertices of those triangles, which form the so-called Persistence Diagrams [22]. Comparison of two images (as far as the criterion intrinsic to the filtering function is concerned) can then be carried out by comparing the sets of these points. Several distances can be defined on the set of PBNs (equivalently, of PDs); a very successful one is the matching distance (see figure 5 and [10] as a general reference). Distance from templates generally produces numbers of some significance with respect to a classification. Unfortunately, there do not exist archetypal naevi or melanomas, so the task is harder than for classical classification problems.
For a fixed database of N images, with n naevi and m melanomas, the software computes 28 symmetric $N \times N$ matrices, each containing the relative distances between the images of the database (of course, the main diagonal of these matrices is null):

- 11 matrices for the 1-dimensional filtering functions defined in section 3 (1-dimensional matching distances between the corresponding PBNs);
- 9 matrices for the 2-dimensional filtering functions defined in section 3 (2-dimensional matching distances between the corresponding 2-dimensional Size Graphs).
- 8 matrices for the ABCDE-features of the parameters defined in section 3 (of the asymmetry parameters, only the parameter *Sym*, which includes the other three, has been used). The relative distance between two images with respect to a generic numerical parameter has been computed as the standard distance in $\mathbb{R}$. A different approach has been used to compute the relative distance between two colour histograms, because a standard Euclidean distance between vectors would have not taken into account that some colours are more similar than others. Therefore, the distance between two colour histograms has been computed by minimizing the cost of matching the two histograms by means of a minimum weight perfect matching algorithm, where the weight associated to the matching of two colours is the distance of the corresponding colours in the *RGB* space equipped with the Euclidean metric.

To perform the image retrieval, all the matrices have been normalized and an average distance has been calculated as a weighted sum of some of these distances. The definition and research of an optimal choice for these weights is a challenging problem.

Here the following approach has been used: among a very large amount of random attempts, select the vector of weights that provides the highest diagnostic accuracy with respect to a fixed classifier over the whole database. In the numerical experiments, three different classifiers have been considered and analyzed.

4.1 Classifiers

Fix a positive integer k and a distance d between the images of the database. For an image I, compute the first k-NN of I with respect to d.
Then the first classifier (Standard Classifier) computes the number of melanomas retrieved among the first k-NN and classifies I as a naevus if and only if this number is lower than km/N. The second and the third classifiers (respectively, the Position and the Distance Classifiers) are modified versions of the Standard Classifier that take into account respectively the positions and the distances of the retrieved images. For example, the Distance Classifier classifies an image as naevus if and only if

$$\sum_{i=1}^{k} \frac{\gamma_{N,m,\mathsf{I}_i}}{d(\mathsf{I},\mathsf{I}_i)} < 0,$$

where $d(\mathsf{I},\mathsf{I}_i)$ is the distance between I and the i-th database image I_i, and the coefficient $\gamma_{N,m,\mathsf{I}_i}$ is equal to $(N-m)/N$ if I_i is a melanoma, and to $-m/N$ otherwise.

5 Numerical experiments

The image retrieval was tested on well-controlled lesion images. In particular, the experiments were conducted on a dataset containing 107 atypical melanocytic skin lesions undergone to excision and histological examination (35 melanomas and 72 melanocytic naevi).
The images were acquired in epiluminescence microscopy with a fixed 16-fold magnification, with the only selection criterion that the lesion had to be entirely visible and with a fixed resolution of 768×576 pixels.
Because of the small dimension of the dataset, it must be emphasized that these experiments cannot lead to significant conclusions, but can nevertheless suggest the road for future developments of the system.
The numerical experiments were performed from four different points of view.

5.1 Classifiers

The aim of the first experiment was to compare the results obtained by the classifiers described in section 4 on the 28 features extracted in section 3.
For the actual computation of the Persistent Betti Numbers on the Global Graphs, the image resolution was reduced by taking averages on squared blocks of dimensions 4×4 pixels. Table 1 summarizes the results, showing the Accuracy percentages obtained by the single features separately. From the table, it appears evident that the PBNs obtain better results than the ABCDE features. Moreover, the best performances are usually

Test 1: classifiers			
Feature	**Std. Class.**	**Pos. Class.**	**Dist. Class.**
	Acc.	Acc.	Acc.
Blue	76,64	80,37	76,64
Green	85,98	85,98	85,98
Red	85,05	86,92	85,05
TV	75,70	74,77	74,77
ExcB	81,31	82,24	82,24
ExcG	81,31	79,44	81,31
ExcR	82,24	80,37	81,31
R-G	80,37	78,50	80,37
R-B	81,31	77,57	80,37
G-B	79,44	81,31	81,31
Border	54,21	57,94	56,07
Histo	75,70	74,77	77,57
FF	54,21	57,94	60,75
CH	54,21	54,21	48,60
Sym	53,27	48,60	49,53
Elt	51,40	49,53	46,73
Ecc	45,79	42,99	51,40
Diam	68,22	66,36	61,68
Entr	46,73	50,47	44,86
(B,G)	84,11	85,05	84,11
(B,R)	85,98	86,92	85,98
(G,R)	85,98	86,92	85,98
(B,R-G)	85,05	85,05	85,98
(G,R-B)	86,92	87,85	86,92
(R,G-B)	85,05	85,05	85,05
(ExcB,R-G)	85,05	86,92	85,05
(ExcG,R-B)	79,44	78,50	80,37
(ExcR,G-B)	85,98	85,05	85,98

Table 1: Numerical results for Test 1. Accuracy percentage results for 28 different features.

achieved by the 2-dimensional PBNs.
The classifiers described in section 3 have usually similar performances. However, in some particular cases, the differences in the accuracy percentages obtained by the three classifiers are not negligible. This highlights an important technical difference between image retrieval and classification: whereas classification calculates an automatic possible diagnosis, image retrieval provides a reliable technical support and remains open to different possible diagnostic results, without replacing the clinician in the classification decision.

5.2 Resolution

The second experiment aimed to compare the results obtained by the Standard Classifier on the features based on Persistent Homology with three different resolutions for computing the PBNs defined on the Global Graph.
Table 2 shows the Accuracy results obtained with three different block dimensions for the different Size Graphs. Significant oscillations in the accuracy results can be noted in many cases. Since they are strongly dependent on the function, a general conclusion cannot be derived. However, it is possible to note that the results tend to improve when passing from 8-dimensional to 4-dimensional blocks: a more detailed analysis seems to provide better classification results. It has to be emphasized that taking averages on

Test 2: resolution					
Feature	**8×8**	**4×4**	**8 vs. 4**	**3×3**	**4 vs. 3**
	Acc.	Acc.	Acc. Diff.	Acc.	Acc.Diff.
Blue	78,50	76,64	**-1,87**	78,50	**+1,87**
Green	84,11	85,98	**+1,87**	84,11	**-1,87**
Red	85,98	85,04	**-0,93**	87,85	**+2,80**
TV	60,75	75,70	**+14,95**	71,03	**-4,67**
ExcB	63,55	81,31	**+17,76**	81,31	**+0,00**
ExcG	72,90	81,31	**+8,41**	78,50	**-2,80**
ExcR	82,24	82,24	**+0,00**	84,11	**+1,87**
R-G	83,18	80,37	**-2,80**	85,98	**+5,61**
R-B	80,37	81,31	**+0,93**	82,24	**+0,93**
G-B	77,57	79,44	**+1,87**	82,24	**+2,80**
(B,G)	84,11	84,11	**+0,00**	n.c.	n.c.
(B,R)	83,18	85,98	**+2,80**	n.c.	n.c.
(G,R)	86,92	85,98	**-0,93**	n.c.	n.c.
(B,R-G)	85,05	85,05	**+0,00**	n.c.	n.c.
(G,R-B)	88,79	86,92	**-1,87**	n.c.	n.c.
(R,G-B)	85,05	85,05	**+0,00**	n.c.	n.c.
(ExcB,R-G)	83,18	85,05	**+1,87**	n.c.	n.c.
(ExcG,R-B)	80,37	79,44	**-0,93**	n.c.	n.c.
(ExcR,G-B)	85,98	85,98	**+0,00**	n.c.	n.c.

Table 2: Numerical results for Test 2. Accuracy percentage results for the PBNs defined on the Global Graphs with different resolutions.

8×8 blocks produces very low-resolution images, causing an information loss which appears evident just by displaying the resulting images, as shown in Figure 6. On the other hand, when reducing the size of the blocks from 4 to 3, deteriorations and improvements in the accuracy results are more balanced, suggesting that, at least in this case, too much detail is not necessary. Moreover, taking the averages on 3×3 blocks produces very large Global Graphs, requiring a huge computational cost. Thus, the 2-dimensional distances between 2-dimensional Size Graphs were not computed in this case, due to the demanding computational effort.
For the sake of completeness, it is worth stating that in this test the behavior of the sensitivity and specificity results was very similar to that of the diagnostic accuracy presented here.

5.3 Number of Nearest Neighbors

The third experiment was addressed to test the stability of the results with respect to the number of the nearest neighbors.

Figure 7 shows the average behaviors of the diagnostic accuracy with respect to the number of the nearest neighbors. In the test, the Standard Classifier was fixed and blocks of size 4×4 were chosen to simplify the Size Graphs; the figure displays the average results obtained by the 1-dimensional (blue line) and 2-dimensional (green line) Persistent Betti Numbers, by the ABCDE functions (red line) and by all 28 features together (cyan line). Although with some differences from case to case, in general the dependency of the results from the number of NN is not particularly strong. Especially in the case of PBNs, there is a large enough interval of possible choices where the classification results can be considered equivalent. More precisely, the results do not change significantly if the number of nearest neighbors varies from 6 to 15, whereas

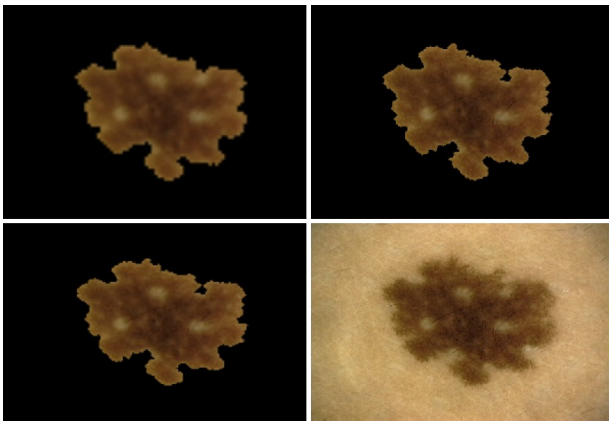


Figure 6: The pictures show the detail deterioration of the image quality caused by the reduction of their resolution. The bottom right picture shows a pigmented skin lesion at its original resolution; the other pictures show its segmented version reduced by splitting it into small squared blocks and by taking averages of the three color channels on these blocks. Results are displayed when the size of the blocks are 3×3 (top right), 4×4 (bottom left), 8×8 (top left).

lower results are achieved when the nearest neighbors number is too low or too high. A different behavior is observed in the case of the ABCDE functions, but in this case the classification results are too low to drive to any conclusion.

5.4 Global distances

In the last numerical experiment the extracted features were matched together to evaluate the performances of the content based image retrieval system.
The Standard Classifier, 10 nearest neighbors and blocks of size 4×4 to simplify the Size Graphs were fixed a priori.
Then three different retrieval systems were implemented as described in section 4, by optimizing respectively the 1-dimensional functions, the 2-dimensional functions plus the Border function, and the ABCDE features separately.
The classification results, summarized in table 3, show that the classification systems

Test 4: Global Retrieval results			
Features	Acc	Sens	Spec
1d PBNs	94,39	94,29	94,44
2d PBNs	92,52	88,57	94,44
ABCDE	86,92	82,86	88,89

Table 3: Numerical results for Test 4. Global Retrieval results obtained by matching some of the features together.

based on the PBNs were able to distinguish naevi and melanomas in a very precise

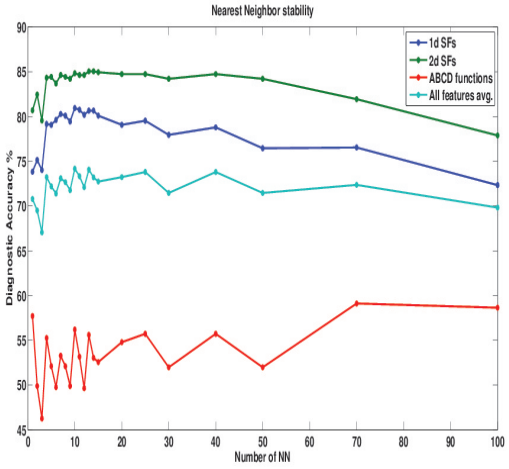


Figure 7: Average Accuracy results obtained by the 28 functions with respect to the NN number.

way. In particular, the results obtained by the 1-dimensional PBNs functions show that only a few of the 107 images were classified in a wrong way by the retrieval based on these functions.
Moreover, although the 2-dimensional PBNs obtained separately the best performances, they failed to achieve better results than the 1-dimensional PBNs in the global case. It is worth mentioning that other numerical experiments, performed with different resolutions and classifiers, led to similar results.
On the other hand, despite the very weak results obtained by the ABCDE functions separately, their optimization improved the performances significantly.
This remark may be the starting point to consider new possibilities, in which the ABCDE features are used at an early stage to discriminate between very different images and then the 1-dimensional PBNs come into play. The authors believe that such hybrid choices should spare computational time without significant losses in the performances.

5.5 Visual results and discussion

Since the final goal of the system is the visualization of the image retrieval, it is worth showing some of the results obtained by the retrieval of the 1-dimensional PBNs functions described in section 5.4. Figure 8 shows the first four neighbors computed for a query image, histologically diagnosed as melanoma. This is one of the rare cases in which a query image has been classified by the software in a wrong way. However, a visual inspection reveals how much this diagnosis could be difficult even for an expert clinician: the query image is symmetric, with a regular border, homogeneously pigmented, and its diameter is smaller than 6 millimeters. The visual retrieval produces a set of neighbors with many clinical features similar to the query, but in this case an

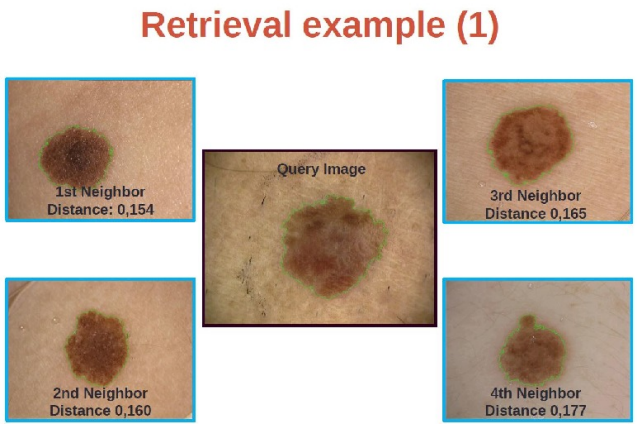


Figure 8: Retrieval of a melanoma with a neighborhood of naevi.

automatic classification would be misleading the diagnostic process. Figure 9 shows

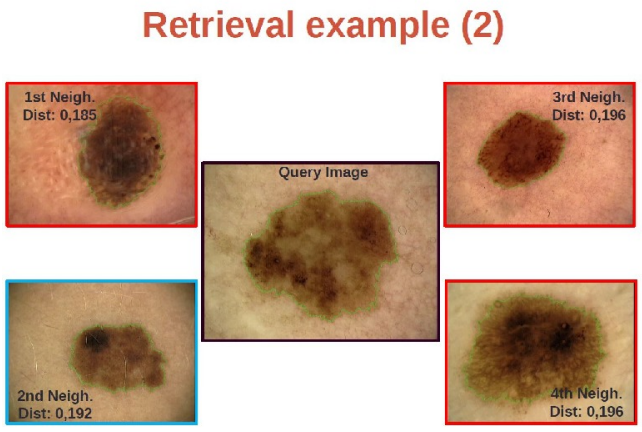


Figure 9: Retrieval of a melanoma with a neighborhood of naevi and melanomas.

the first four neighbors of another melanoma. In this case, the automatic classification is correct, because of the many melanomas present in the neighborhood of the query. Again, the visual performance is satisfying from the clinical point of view, in the sense that the neighbors share many clinical features with the query, especially if one takes into account that melanomas tends to be different from each other. Moreover, despite being a naevus, the second neighbor (bottom left) is similar to the query and could indeed seem a melanoma itself, because of its different colors. An automatic risk computation could produce a medium-high risk score for this lesion, but the retrieval would justify an even greater concern.

These examples show that the retrieval method could prove to be a more useful diag-

nostic support for the clinician than classification and risk assessment-type algorithms, suggesting investigating deeper this method and in particular the algorithm proposed in this article. Summing up, the numerical experiments conducted on this database suggest that the method deserves further investigation. Moreover:

- the system appears to be resilient to the reduction of the images' resolution, at least as far as too much detail is not lost.
- The numerical results of the 2-dimensional Persistent Betti Numbers may not be worth their computational costs for the global distance computation.
- Small differences in the choice of the NN number computed by the retrieval system do not seem to affect the accuracy results in a significant way.
- Retrieval could prove to be a more efficient diagnostic support than classification.

6 Conclusions

As pointed out in [34], at this time, there are no computers that can replace an experienced dermatologist's intuition: a blind trust of the judgement of any computerized system could lead the clinician to a failure in the diagnosis. On the other hand, computerized systems can prove to be an important diagnostic support for the dermatologist in the diagnostic phase.
Starting from this consideration, in this paper the authors propose a new k-NN algorithm for the image retrieval of skin lesions. The main novelty of the algorithm is the extensive use of Persistent Homology theory, with new tools with respect to the previous works [9, 18, 44], which enable a global analysis of the skin lesion. These tools have been tested on a dataset of 107 melanocytic lesions, with promising results, leading to the conclusion that this method deserves further investigation and tests.

Acknowledgements

Support by ARCES, CA-MI S.r.l., IRST-IRCCS and the National Institute of High Mathematics "F. Severi" (INdAM) is gratefully acknowledged.

References

[1] Ackermann, W.: Zum hilbertschen aufbau der reellen zahlen. Math. Ann. **99**, 113–133 (1928)

[2] Biasotti, S., Cerri, A., Frosini, P., Giorgi, D.: A new algorithm for computing the 2-dimensional matching distance between size functions. Patt. Recogn. Lett. **32**, 1735–1746 (2011)

[3] Biasotti, S., Cerri, A., Frosini, P., Giorgi, D., Landi, C.: Multidimensional size functions for shape comparison. J. Math. Imaging Vis. **32**(2), 161–179 (2008)

[4] Burroni, M., Corona, R., et al., G.D.: Melanoma computer-aided diagnosis: reliability and feasibility study. Clin. Canc. Res. **10**(6), 1881–1886 (2004)

[5] Cagliari, F., Di Fabio, B., Ferri, M.: One-dimensional reduction of multidimensional persistent homology. Proc. Amer. Math. Soc. **138**, 3003–3017 (2010)

[6] Cerri, A., Ferri, M., Giorgi, D.: Retrieval of trademark images by means of size functions. Graph. Models **68**(5), 451–471 (2006)

[7] Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of persistence diagrams. In: J.S.B. Mitchell, G. Rote (eds.) Symposium on Computational Geometry, ACM, pp. 263–271 (2005)

[8] d'Amico, M.: A new optimal algorithm for computing size function of shapes. In: Proc. CVPRIP Algorithms III, International Conference on Computer Vision, Pattern Recognition and Image Processing, pp. 107–110 (2000)

[9] d'Amico, M., Ferri, M., Stanganelli, I.: Qualitative asymmetry measure for melanoma detection. In: Proceedings of the 2nd IEEE International Symposium on Biomedical Imaging: Macro to Nano, pp. 1155–1158 (2004)

[10] d'Amico, M., Frosini, P., Landi, C.: Using matching distance in size theory: A survey. Int. J. Imag. Syst. Tech. **16**(5), 154–161 (2006)

[11] d'Amico, M., Frosini, P., Landi, C.: Natural pseudo-distance and optimal matching between reduced size functions. Acta Appl. Math. **109**, 527–554 (2010)

[12] Del Mar, C., Green, A., et al., T.C.: Melanocytic lesions excised from the skin: what percentage are malignant? Aust. J. of Public Health **18**, 221–223 (1994)

[13] Dibos, F., Frosini, P., Pasquignon, D.: The use of size functions for comparison of shapes through differential invariants. J. Math. Imaging Vis. **21**(2), 107–118 (2004)

[14] Dreiseitl, S., Binder, M., Hable, K., Kittler, H.: Computer versus human diagnosis of melanoma: evaluation of the feasibility of an automated diagnostic system in a prospective clinical trial. Melanoma Research **19**(3), 180–184 (2009)

[15] Eckhardt, U., Latecki, L.: Digital topology. In: Current topics in Pattern Recongnization Research, Research Trends. Council of Scientific Information, Vilayil Gardens, Trivandrum, India (1994)

[16] Edelsbrunner, H., Harer, J.: Persistent homology - a survey. In: J. Goodman, J. Pach, R. Pollack (eds.) Surveys on Discrete and Computational Geometry. Twenty Years Later, vol. 453, pp. 257–282. Contemporary Mathematics, Amer. Math. Soc., Providence, Rhode Island (2008)

[17] Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., Parkin, D.M.: GLOBOCAN 2008: Cancer incidence and mortality Worldwide: IARC CancerBase No. 10, Lyon, France. Available from: `http://globocan.iarc.fr` (2010)

[18] Ferri, M., Stanganelli, I.: Size functions for the morphological analysis of melanocytic lesions. Int. J. Biomedical Imaging **2010**, 5 pp. (2010). DOI 10.1155/2010/621357. Article ID: 621357

[19] Friedman, R.J., Gutkowicz-Krusin, D., et al., M.J.F.: Diagnostic performance of expert dermoscopists vs a computer-vision system on small-diameter melanomas. Archives of Dermatology **144**(4), 476–482 (2008)

[20] Frosini, P.: Connections between size fonctions and critical points. Mathematical Methods in the Applied Sciences **19**(7), 555–569 (1996)

[21] Frosini, P., Landi, C.: Size theory as a topological tool for computer vision. Pattern Recogn. and Image Anal. **9**, 596–603 (1999)

[22] Frosini, P., Landi, C.: Size functions and formal series. Applicable Algebra in Engineering, Communications and Computing **12**(4), 327–349 (2001)

[23] Gomes, J., Faugeras, O.: Reconciling distance functions and level sets. J. Visual Communic. and Imag. Representation **11**, 209–223 (2000)

[24] Haralick, R.M.: A measure for circularity of digital figures. IEEE Transactions on Systems, Man and Cybernetics, SMC **4**(4), 334–336 (1974)

[25] Khalimsky, E., Kopperman, R., Meyer, P.R.: Computer graphics and connected topologies on finite ordered sets. Topology and its Applications **36**, 1–17 (1990)

[26] Kiani, K., R. Sharafat, A.: E-shaver: An improved dullrazor for digitally removing dark and light-colored hairs in dermoscopic images. Computers in Biology and Medicine **41**(3), 139–145 (2011)

[27] Kong, T.Y., Rosenfeld, A.: Digital topology: Introduction and survey. Computer Vision, Graphics and Image Processing **48**, 357–393 (1989)

[28] Kopf, A.W., Salope, T.G., Slade, J., Marghoob, A.A., Bart, R.S.: Techniques of cutaneous examination for the detection of skin cancer. Cancer **75**(2), 684–690 (1994)

[29] Lee, T., Ng, V., Gallagher, R., Coldman, A., McLean, D.: Dullrazor: A software approach to hair removal from images. Computers in Biology and Medicine **27**, 533–543 (1997)

[30] Li, C., Huang, R., Ding, Z., Gatenby, J.C., Metaxas, D.N., Gore, J.C.: A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI. IEEE Trans. Image Processing **20**(7), 2007–2016 (2011)

[31] Li, C., Kao, C., Gore, J.C., Ding, Z.: Minimization of region-scalable fitting energy for image segmentation. IEEE Trans. Image Processing **17**(10), 1940–1949 (2008)

[32] Li, C., Xu, C., Gui, C., Fox, M.D.: Distance regularized level set evolution and its application to image segmentation. IEEE Trans. Image Processing **19**(12), 3243–3254 (2010)

[33] Lucas, R., McMichael, T., Smith, W., Armstrong, B.: Solar Ultraviolet Radiation: Global Burden of Disease from Solar Ultraviolet Radiation. Environmental Burden of Disease Series no. 13, Geneva: World Health Organization (2006)

[34] Masood, A., Al-Jumaily, A.A.: Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. Int. J. Biomedical Imaging **2013**, 22 pp. (2013). DOI 10.1155/2013/323268. Article ID: 323268

[35] Massone, C., Di Stefani, A., Soyer, P.: Dermoscopy for skin cancer detection. Curr. Opin. Oncol. **17**, 147–153 (2005)

[36] Miller, M., Ackermann, A.B.: How accurate are dermatologists in the diagnosis of melanoma? Degree of accuracy and implication. Arch. Dermatol. **128**, 559–560 (1992)

[37] Osher, S., Sethian, J.: Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. J. Comput. Phys. **79**(1), 12–49 (1988)

[38] Pehamberger, H., Binder, M., et al., A.S.: In vivo epiluminescence microscopy: improvement of early diagnosis of melanoma. Invest. Dermatol. **100**, 356S–362S (1993)

[39] Pehamberger, H., Steiner, A., Wolff., K.: In vivo epiluminescence microscopy of pigmented skin lesions. I. Pattern analysis of pigmented skin lesions. Am. Acad. Dermatol. **17**, 571–583 (1987)

[40] Rosenfeld, A.: Digital topology. American Mathematical Monthly **86**, 621–630 (1979)

[41] Seidenari, S., Pellacani, G., Giannetti, A.: Digital videomicroscopy and image analysis with automatic classification for detection of thin melanomas. Melanoma Research **9**(2), 163–171 (1999)

[42] Sethian, J.A.: Level set methods and fast marching methods. Cambridge University Press, Cambridge (1999)

[43] Siegel, R., Ma, J., Zou, Z., Jemal, A.: Cancer statistics 2014. CA: A Cancer Journal for Clinicians **64**(1), 9–29 (2014)

[44] Stanganelli, I., Brucale, A., et al., L.C.: Computer-aided diagnosis of melanocytic lesions. Anticancer Research **25**(6), 4577–4582 (2005)

[45] Stanganelli, I., Bucchi, L.: Epiluminescence microscopy versus clinical evaluation of pigmented skin lesions: effects of Operator's training on reproducibility and accuracy. Dermatology and Venereology Society of the Canton of Ticino. Dermatology **196**(2), 199–203 (1998). PubMed PMID: 9568407

[46] Stanganelli, I., Clemente, C., Mihm Jr., M.C.: Skin Melanoma CD-Multimedial Interactive Atlas for prevention, diagnosis and therapy of Melanoma and pigmented skin lesions. Istituto Oncologico Romagnolo Ed., MAF Turin, Italy (2001). In Italian

[47] Stanganelli, I., Seidenari, S., Serafini, M., Pellacani, G., Bucchi, L.: Diagnosis of pigmented skin lesions by epiluminescence microscopy: determinants of accuracy improvement in a nationwide training programme for practical dermatologists. Public Health **113**(5), 237–242 (1999). PubMed PMID: 10557118

[48] Stanganelli, I., Serafini, M., Cainelli, T., Cristofolini, M., Baldassari, L., Staffa, M., Bucchi, L.: Accuracy of epiluminescence microscopy among practical dermatologists: a study from the Emilia-Romagna region of Italy. Tumori **84**(6), 701–705 (1998). PubMed PMID: 10080681

[49] Vestergaard, M.E., Macaskill, P., Holt, P.E., Menzies, S.W.: 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), 669–676 (2008)

[50] Weinstock, M.A.: Epidemiology, etiology and control of melanoma. Med. Health R.I. **84**(7), 234–236 (2001)