OPEN

Novel strategy for applying hierarchical density-based spatial clustering of applications with noise towards spectroscopic analysis and detection of melanocytic lesions

Jason Yuan Ye[a,b], Christopher Yu[c], Tiffany Husman[a], Bryan Chen[a] and Aryaman Trikala[a]

Advancements in dermoscopy techniques have elucidated identifiable characteristics of melanoma which revolve around the asymmetrical constitution of melanocytic lesions consequent of unfettered proliferative growth as a malignant lesion. This study explores the applications of hierarchical density-based spatial clustering of applications with noise (HDBSCAN) in terms of the direct diagnostic implications of applying agglomerative clustering in the spectroscopic analysis of malignant melanocytic lesions and benign dermatologic spots. 100 images of benign (*n*=50) and malignant moles (*n*=50) were sampled from the International Skin Imaging Collaboration Archive and processed through two separate Python algorithms. The first of which deconvolutes the three-digit tupled integer identifiers of pixel color in image composition into three separate matrices corresponding to the red, green and blue color channel. Statistical characterization of integer variance was utilized to determine the optimal channel for comparative analysis between malignant and benign image groups. The second applies HDBSCAN to the matrices, identifying agglomerative clustering in the dataset. The results indicate the potential diagnostic applications of HDBSCAN analysis in fast-processing dermoscopy, as optimization of clustering parameters according to a binary search strategy produced an accuracy of 85% in the classification of malignant and benign melanocytic lesions. *Melanoma Res* 31: 526–532

Melanoma Research 2021, **31**:526–532

Keywords: Clustering, dermoscopy, hierarchical density-based spatial clustering of applications with noise, machine learning

[a]University of California, Los Angeles, [b]UCLA Microbiome Center, David Geffen School of Medicine at UCLA, Los Angeles, California and [c]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Correspondence to Jason Yuan Ye, University of California, Los Angeles, 330 De Neve Drive, Los Angeles, CA 90024, USA
E-mail: jason.ye9@gmail.com

Received 21 May 2021 **Accepted** 16 July 2021

Background

Cutaneous melanoma is a significant problem that persists despite medical and diagnostic innovation. Global incidences of melanoma occur at approximately 160 000 cases a year with an associated mortality of 48 000 deaths annually [1]. Furthermore, metastatic melanoma is highly resistant to conventional therapies, and this necessitates the development of novel methodologies for early diagnoses and detection [2].

Dermoscopy involves the magnification and in-vivo observation of lesions via handheld instruments, combined with diagnostic algorithms based on dermoscopic-histological correlations [1]. Advancements in dermoscopy techniques have elucidated key identifiable physical characteristics of melanoma which revolve around the asymmetrical constitution of melanocytic lesions consequent of unfettered proliferative growth as a malignant lesion [1]. Such factor described in established dermoscopic criteria include atypical pigment networks and branching, irregular pigment blotches, general asymmetry of the pigmentation, atypical vascular patterns and blue-white veils formed on the surfaces of some lesions [3]. Currently, clinical diagnosis of melanocytic lesions is based on the subjective delineation of dermatologic spot or mole borders which gives rise to observational variance [4].

Previous research has shown promise in applying clustering techniques in detecting lesion borders for the purposes of image segmentation in advancing medical diagnosis and treatment. Clustering has been persistently explored as an option regarding border detection in algorithm-based dermoscopy, with techniques such as PCT/median cut algorithm, Markov random field segmentation and nonlinear diffusion.5 Recently, density-based clustering has been popularized as an intelligent method for the selection of query points in image analysis, by treating image data as vectors in a space with a given density and distance from disparate data points.5 In a comparative analysis of density-based and fuzzy c-means clustering, Kockara found that density-based spatial

0960-8931 Copyright © 2021 The Author(s). Published by Wolters Kluwer Health, Inc. DOI: 10.1097/CMR.0000000000000771

clustering of applications with noise (DBSCAN) was more effective in lesion border delineation as opposed to the popularly utilized alternative in image segmentation, Fuzzy C-Means clustering.6

This study explores the applications of hierarchical density-based spatial clustering of applications with noise (HDBSCAN) in terms of the direct diagnostic implications of applying agglomerative clustering in the spectroscopic analysis of malignant melanocytic lesions and benign dermatologic spots. HDBSCAN contrasts DBSCAN in the way that it manages its interpretation of data. It consolidates an ordered representation of density-based analysis across all possible density parameters, whereas DBSCAN is limited by its epsilon parameter for density, which it applies as a global distance parameter.7,8 DBSCAN's limitation in terms of its epsilon parameter means that it fails to discover clusters that might exist with different densities that do not fall within the parameter.8 Different approaches towards automated cluster selection were proposed to handle this problem, and the method proposed by Campello *et al*., [9] excess of mass selection, stands out as a practical and efficacious solution through calculating an optimal global solution to determining appropriate variable densities by finding clusters of high stability and persistence in terms of their devised density value λ. Campello's method consolidates HDBSCAN as a potential use for exploratory data-mining where, functionally, only one parameter of the algorithm, the minimum cluster size of HDBSCAN, is necessitated, making the HDBSCAN alternative a promising avenue for fast-processing dermoscopy. 8 Solving the problem of determining a global epsilon parameter value, clustering-driven data extraction can be focused on a more simplistic form of optimization based upon the linear adjustment of a single parameter, giving HDBSCAN an advantage towards diagnostic utilization, as the single parameter may be optimized to elucidate clinically useful clustering data, as illustrated in our approach.

Methods

Image analysis dataset description

The images used in this analysis were obtained from a balanced data set, randomly sampled from The International Skin Imaging Collaboration (ISIC), which contains a public gallery of up to 69445 primarily dermoscopic images.10 Of the images, 33632 (48%) are confirmed to be melanocytic, and the largest proportion of images are sampled from patients ranging from 40 to 70 years of age, with that demographic comprising 44% of all sampled images.10 The ISIC extracted their images from patient databases of high-risk clinics and tertiary referral centers for melanoma, such as Memorial Sloan Kettering Cancer Center, the Hospital Clinic Barcelona, the University of Queensland and the Medical University Vienna, with institutional review board approval.11 Accordingly, identifying patient information was stripped from images prior to publication and the images published to the ISIC were filtered from an original sample of images based on image quality, associated qualifying diagnoses and histopathological confirmation of diagnoses.11 To demonstrate the generalizability of HDBSCAN in elucidating clinically significant features of dermoscopic images, the ISIC images were sampled from the broad categories of either 'malignant' or 'benign' according to ISIC image gallery filters.

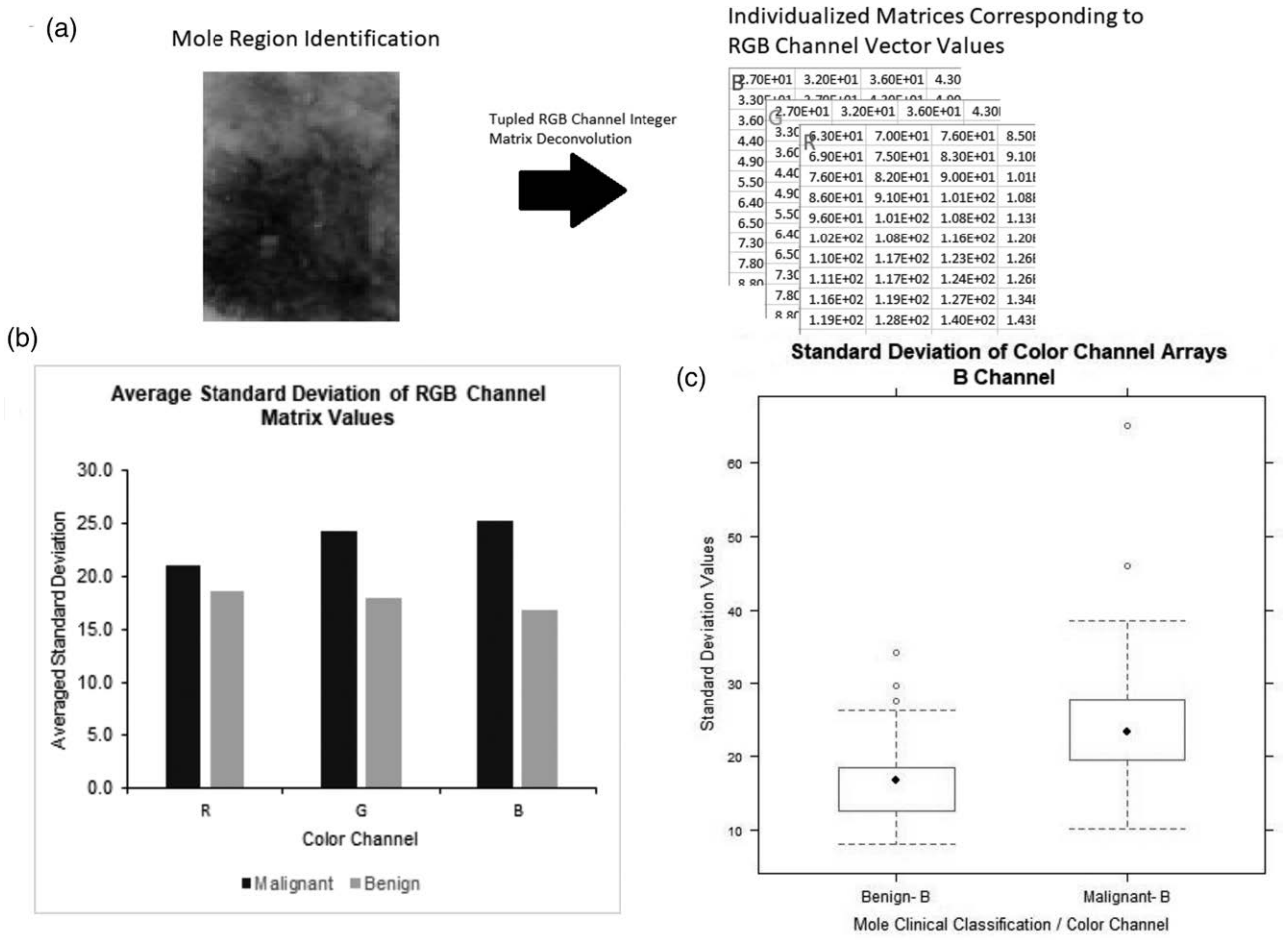
Matrix deconvolution and red, green and blue channel sorting

In total 100 dermoscopic images of either malignant melanocytic lesions ($n = 50$) or benign, nonmelanocytic moles ($n = 50$) were obtained based on random sampling from the ISIC archive and sampled for image data. Images of center region portions of moles were individually deconvoluted using a Python algorithm which automatically separates and quantizes the three-digit tupled integer identifiers of pixel color composition inherent in all images portrayed digitally (Fig 1a). From each individual image, a matrix of integers corresponding to the red, green and blue (RGB) color channels was produced, and the primary endpoint was determining whether the image data retrieved were representative of the physical dermoscopic differences between melanocytic and non-melanocytic lesions or moles (Fig. 1a). Because an individual pixel contains three tupled integer identifiers of color, three matrices were produced per image making 300 total matrices with 100 per color channel with color channel integers located in the matrix with respect to their spatial positioning in the original sample image (Fig. 1a). Accordingly, the average SD of color channel matrix integer values, per each individual matrix of the 300 produced, were obtained and compared between the malignant and benign experimental groups, as shown in Fig. 1b.

Applying hierarchical density-based spatial clustering of applications with noise

A secondary endpoint was determining the efficacy of the application of HDBSCAN analysis in elucidating the physical differences of malignant and benign moles in terms of objective and quantifiable results. Accordingly, the B channel matrices were identified as a target for clustering due to the high difference in SD in matrix integer values between malignant and benign experimental groups (Fig. 1c). A greater disparity in pixel identifier values will yield more robust clustering in grouping significant cluster-associated matrix vector values from noise. An HDBSCAN algorithm developed in Python was applied to the blue channel matrices corresponding to the two experimental groups and results were assessed comparatively between the groups. The algorithm was applied to 100 samples corresponding to 50 B channel matrices of the malignant experimental group and 50 B channel matrices of the benign experimental group. The algorithm

Fig. 1



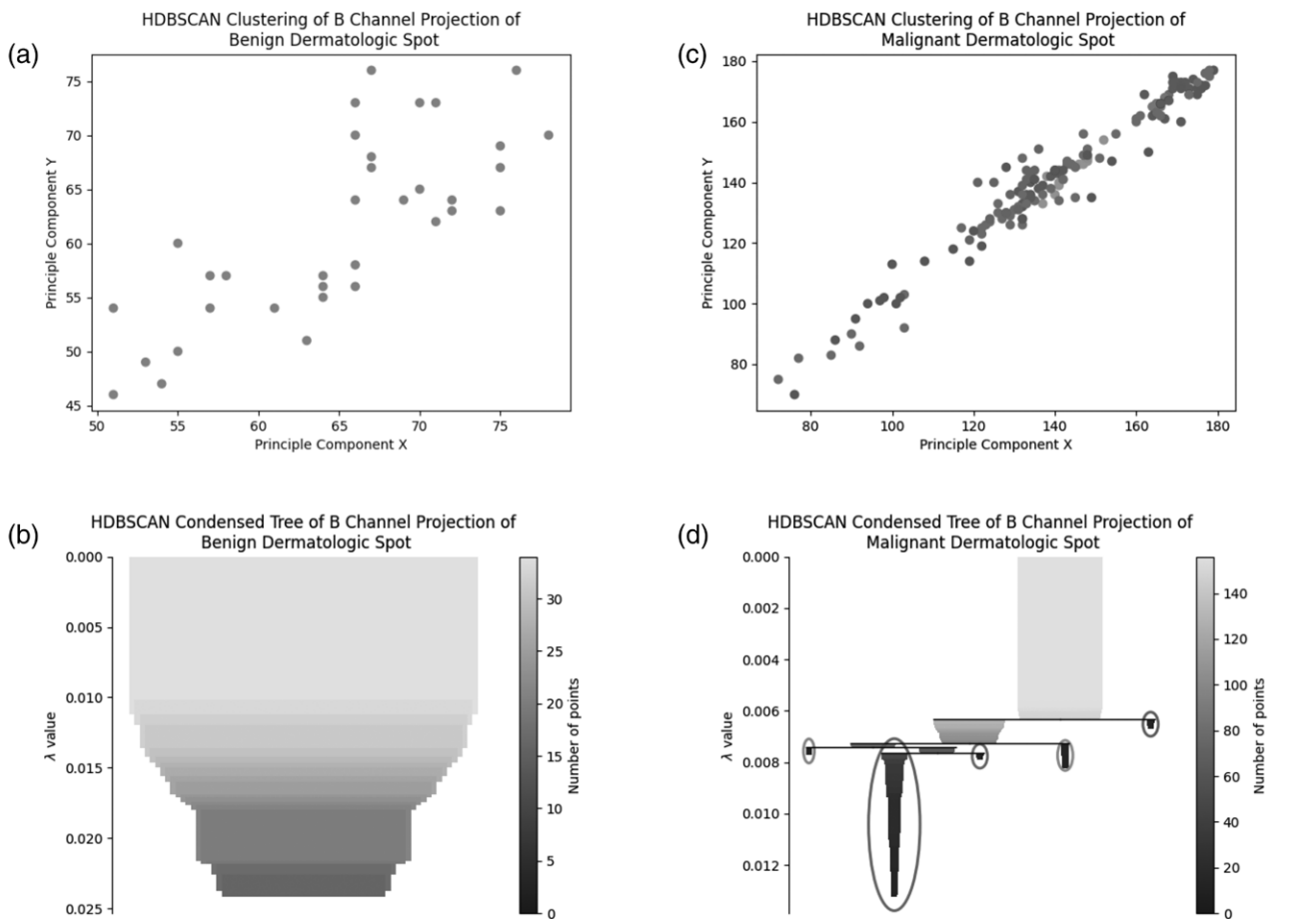
(a) Schematic representation of RGB channel deconvolution into constituent matrices formed from RGB tupled integer pixel color identifiers. (b) Graphical comparison of average standard deviation values between malignant and benign RGB channel matrix integer values corresponding to R, G, and B channels independently. (c) Comparative boxplot of malignant and benign standard deviation value distribution and spread.

converted the B channel matrix projection integer values corresponding to pixel color into vector quantities corresponding to Euclidean distance. The approach allowed for the delineation of grouped color vertices from sparse points, allowing for robust clustering to noise[8]. For each matrix, the algorithm produced a clustering plot of the X and Y principal component of clustering, whereby the X component represents cluster core points and the Y component corresponds to weighted edges or border points of the established clusters (Fig. 2a,c). Data points belonging to clusters were labeled with different colors corresponding to their parent cluster and points representing cluster noise were grayed out. Data points were grouped into clusters or classified as noise based on the clustering selection method of excess of mass (eom), an internal function of HDBSCAN, which serves as an optimized global solution to determining clusters of high stability based on their persistence in continually determined nested parent and child clusters[8]. The automated selection of clusters based on the eom protocol is visually depicted by the condensed tree dendrogram plot, also produced automatically by the HDBSCAN algorithm, shown in Fig. 2b,d, where detected individual clusters are circled and highlighted. The number of clusters per color channel matrix was determined from the circled clusters in the condensed tree dendrogram, as shown in Fig. 3a.

Hierarchical density-based spatial clustering of applications with noise parameter selection and diagnostic criteria

The primary parameter of the HDBSCAN algorithm, the minimum cluster parameter, was adjusted for the determination of malignancy by the presence of clustering. The established diagnostic criteria according to our implementation of the HDBSCAN algorithm was that the presence of clustering in the data set is indicative of a malignant dermatologic spot, and the absence of detected clusters would classify the spot as being benign. The aim of our implementation of the algorithm was to adjust the minimum cluster parameter for exacerbating

Fig. 2



(a,c) HDBSCAN clustering plots of malignant and benign dermatologic spots graphed against the principle component X and Y, which respectively correspond to identified core points of clusters and average weighted distances of data points within the cluster. (b,d) Condensed Tree plots of HDSCAN illustrate detected clusters through circled child cluster branches. No clustering detected in the B channel projection of the benign dermatologic spot, and 5 clusters detected in the malignant dermatologic spot. HDBSCAN, hierarchical density-based spatial clustering of applications with noise.

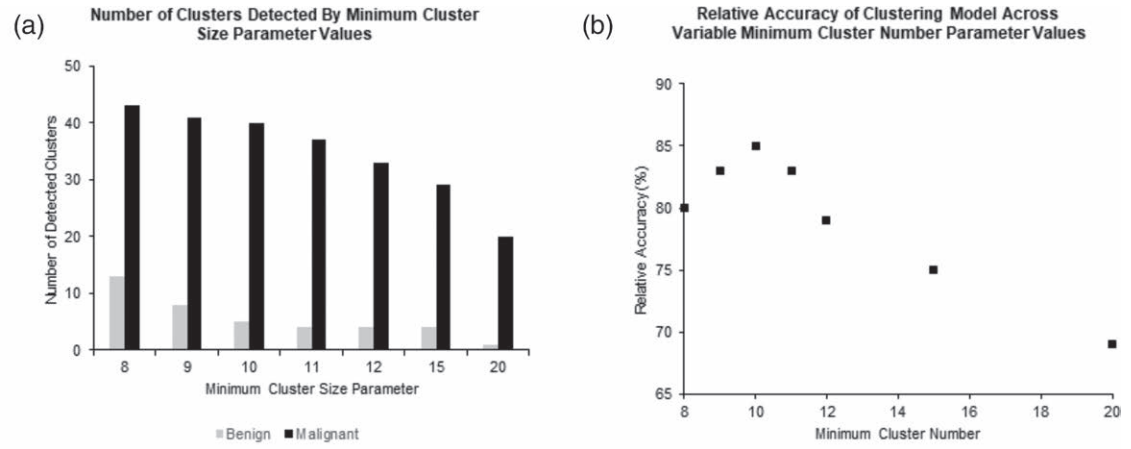
the difference in the presence or absence of clustering between the two experimental groups (Fig. 3c). Accuracy of the HDBSCAN algorithm was determined per value of the minimum cluster parameter by the total number of matrices without clusters in the benign group summed with the total number of matrices with clusters in the malignant group out of the 100 total matrices across both groups (Fig. 3c). Accordingly, the parameter was adjusted continually from its default of a minimum cluster size of 20 towards 0 according to the improvement of accuracy in terms of the aforementioned diagnostic criteria (Fig. 3b,c).

Discussion

Of the deconvoluted RGB color channel matrices, the difference of average SD values between the malignant and benign groups was the greatest for the matrices obtained from the blue channel (Fig. 1b). The malignant group ($n=50$) had an average SD of 25.3 whereas the benign group ($n=50$) had an average SD of 16.8 ($P<0.0001$). Results indicated that spectroscopic differences between the benign and malignant dermatologic groups would be best represented quantitatively through the B channel matrices, making the B channel matrix projections the most appropriate target for HDBSCAN analysis. Our results are consistent with previous dermoscopic correlative analyses of the blue color channel plane with malignant dermoscopic features, where blue color skin pigmentation is considered a positive criterion in discrimination of melanocytic lesions from benign lesions [12]. Figure 1c confirms the correlation between the blue channel or plane with melanocytic characteristics and validates the targeted application of HDBSCAN towards the blue channel matrices.

Applying the HDBSCAN algorithm to the blue channel image plane yielded a high difference in the numbers of clusters formed from B channel matrices of the two

Fig. 3



(c)

Amount of Samples with Clusters Detected

Minimum Cluster Parameter Value	Benign (n=50)	Malignant (n=50)	Relative Accuracy (%) (N=100)
8	13	43	80
9	8	41	83
10	5	40	85
11	4	37	83
12	4	33	79
15	1	20	75
20	4	29	69

(a) Graph of the number of clusters detected with the HDBSCAN algorithm by different minimum cluster size parameter values. (b) Graphical representation of determined accuracy according to the presence or absence of clustering between the benign and malignant experimental groups in terms of minimum cluster number. (c) Data table of amount of samples with detected clusters between the benign and malignant experimental groups with corresponding minimum cluster parameter values and relative accuracy calculations. HDBSCAN, hierarchical density-based spatial clustering of applications with noise.

experimental groups, and the adjustment of parameters allowed for the improved differentiation of melanocytic and nonmelanocytic lesions simply based on the presence or absence of clustering, whereby clustering is considered an identifier of malignancy (Fig. 3a). Under the minimum cluster parameter value of 10, only 5 clusters were detected in the benign group and 40 were produced in the malignant group, corresponding to an accuracy of 85% according to the diagnostic criteria (Fig. 3c). Based on the diagnostic accuracy model shown in Fig. 3a,b minimum cluster parameter value of 10 was established under our review to be the most efficacious in determining malignant melanocytic lesions from benign moles across parameter values ranging from 8 to 20.

While skin biopsy is the most established definitive form of diagnosis for cutaneous melanoma following clinical examination, imaging techniques have a potential role in serving as an adjunct to histopathological determinations of melanocytic neoplasms [13]. Surface-level examinations of cutaneous lesions, suspected of being melanocytic, face difficulties in determining morphologically atypical nevi from melanocytic lesions, which are common in cases with patients exhibiting 'atypical mole syndrome' [14]. Furthermore, seborrheic keratoses and pigmented basal cell carcinomas are also difficult to distinguish from melanoma, owing to their superficially irregular appearance, typically indicative of melanoma [15,16]. A commonly adhered to clinical method for examination, the ABCD(E) clinical rule, relates identification of melanoma to the aforementioned acronym, where the letters correspond respectively to asymmetry, border irregularity, color variegation, diameter and evolution in size and shape [14]. According to the ABCD clinical rule, a lesion exhibiting said criteria should undergo excision and biopsy for a histopathological verification of melanoma [14]. Exceptions for the ABCD clinical rule include nodular melanoma, which initially presents as perfectly symmetric tumors, and under such circumstances, the EFC clinical rule, corresponding to elevated tumor, firm consistency and rapid growth, is applied in clinical examination [14]. In terms of conventional melanoma generalized clinical rules can be applied for

preliminary evaluations of lesions; however, subtypes of melanoma, including facial, acral, nail and amelanotic melanoma all require more specific differential diagnoses following initial identification [14].

The method detailed in this study focuses on the 'C' criteria of the ABCD clinical rule, as the deconvolution of 2D images followed by processing with HDBSCAN focuses the analysis to associating color variegation with the classification of a melanocytic lesion. By our methods, sparse and dense pigmental networks or blobs, generally characteristic of most melanocytic lesions, are represented as clusters elucidated by the HDBSCAN algorithm. This is because pixels of similar and contrastive color to their environment, such as pigment blobs of a melanocytic lesion, are numerically represented as vector points of a similar integer value with significant distance to other points in the matrix.

As such, it comes as no surprise that the analysis is imperfect with an accuracy of 85% in distinguishing melanoma from benign moles, as the analysis mainly applies just one facet of a clinical rule for the preliminary examination. Furthermore, because the analysis was applied to a balanced data set of benign and malignant moles, the mentioned elusive characteristics of various subtypes of melanoma, as well as the presence of morphologically atypical nevi contribute to a decline in accuracy.

A potential limitation to the accuracy of our results may stem from the lack of resolution and clarity of the ISIC database images used in our analysis, as a majority of images published have a resolution of 640×480 pixels [10]. Because each pixel is treated as a data point under our utilization of the HDBSCAN algorithm, it is plausible that a higher resolution, providing more pixels for analysis, would allow for more definitive and accurate clustering of pigmented regions.

A promising alternative for analysis is fluorescent images produced by Wood's light, a cutaneous imaging technique which uses UV-A light to accentuate the contrast in pigmentation between a lesion and surrounding skin [17]. Coupling our methods with techniques such as Wood's light could yield promising results under circumstances where changes in color variegation is a diagnostic endpoint, such as detecting pigment recurrence after dysplastic nevus excisions or monitoring segmental atypical lentiginous naevus for evolution [17]. The contrastive outcomes of clustering between benign and malignant moles demonstrate that our techniques elucidate and profile atypical pigmental variegation and relate digital encoding of images to confirmed clinical diagnoses with some exception to abnormalities (Fig. 3a). Clustering data can potentially be used as an objective and quantitative point of reference for measuring the evolution of a single nevus over time, where observable changes in pigmentation would surely result in spatial changes of clusters as well as clustering detection.

Another common, noninvasive method that facilitates early melanoma diagnoses is total body photography, which involves digitally photographing a patient's entire skin surface with high-resolution images for the purpose of tracking changes in existing lesions and the formation of new lesions [18]. Our methods, which employ a two-part algorithm of deconvolution and then clustering detection, may be advantageous over machine learning techniques in terms of being fast-processing, allowing for the detection of clustering to serve both as a warning of potential malignancy and a reference for the study of evolution in the lesion.

Conclusion

Applying HDBSCAN for the purposes of fast-processing dermoscopic analysis is promising in terms of its clinical applications. This study validates the effectiveness of agglomerative hierarchical clustering in illustrating and quantifying the physical differences of melanocytic lesions and benign, nonmelanocytic moles. HDBSCAN employs efficient implementation into Python with the support of a variety of metrics such as the sckikit-learn library [8]. Images were sampled in the JPG file format and this speaks to the applicability of using Python-based algorithms in potentially generating a tangible application for widespread use by the general public and physicians, assisting in the early detection of melanoma [8]. With an accuracy of 85% based on a sample size of 100, HDBSCAN shows practicality in assimilation into dermoscopic scoring criteria established for clinical diagnosis. Current diagnostic methodologies for melanoma diagnoses fall within methods such as the 7-point checklist for key melanocytic features such as an atypical pigment network or blue-white veil, the Menzies method, and ABCD rule, which all purport a structured analysis of dermatological features [19]. Further analysis into deeper parameters of the algorithm such as a comparison of parent and child clustering sizes and cluster lambda values could provide more insight towards improving the clinical application of HDBSCAN for diagnostic purposes.

Acknowledgements

Conflicts of interest

There are no conflicts of interest.

References

1 Eggermont AM, Spatz A, Robert C. Cutaneous melanoma. *Lancet (London, England)* 2014; **383**:816–827.

2 Paluncic J, Kovacevic Z, Jansson PJ, Kalinowski D, Merlot AM, Huang ML, *et al*. Roads to melanoma: key pathways and emerging players in melanoma progression and oncogenic signaling. *Biochim Biophys Acta* 2016; **1863**:770–784.

3 Brancaccio G, Russo T, Lallas A, Moscarella E, Agozzino M, Argenziano G. Melanoma: clinical and dermoscopic diagnosis. *G Ital Dermatol Venereol* 2017; **152**:213–223.

4 Lemon J, Kockara S, Halic T, Mete M. Density-based parallel skin lesion border detection with webCL. *BMC bioinformatics* 2015; **16**(Suppl 13):S5.
5 Mete M, Sirakov NM. Lesion detection in demoscopy images with novel density-based and active contour approaches. *BMC Bioinformatics* 2010; **11**(Suppl 6):S23.
6 Kockara S, Mete M, Chen B, Aydin K. Analysis of density based and fuzzy c-means clustering methods on lesion border extraction in dermoscopy images. *BMC Bioinformatics* 2010; **11**(Suppl 6):S26.
7 Sander J, Qin X, Lu Z, Niu N, Kovarsky A. Automatic Extraction of Clusters from Hierarchical Clustering Representations. *Advances in Knowledge Discovery and Data Mining* 2003; 75–87.
8 Malzer C, Baum M. A Hybrid Approach To Hierarchical Density-based Cluster Selection. 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). 2020.
9 Campello RJGB, Moulavi D, Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei J, Tseng VS, Cao L, *et al*. (eds). *Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science*, vol **7819**. Springer; 2013.
10 Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D, *et al*. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv* 2019; 1902.03368.
11 Rotemberg V, Kurtansky N, Betz-Stablein B. *et al*. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci Data* 2021; **8**:34.
12 Argenziano G, Fabbrocini G, Carli P, De Giorgi V, Sammarco E, Delfino M. 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* 1998; **134**:1563–1570.
13 Swetter SM, Tsao H, Bichakjian C, Curiel-Lewandrowski C, Elder D, Gershenwald J, *et al*. Guidelines of care for the management of primary cutaneous melanoma. *J Am Acad Dermatol* 2019; **80**:208–250.
14 Aimilios L, Brancaccio G. Diagnosis of Primary Melanoma: Clinical Presentation. *Cutaneous Melanoma: a pocket guide for diagnosis and management*. by Giuseppe Argenziano. 1st ed. Elsevier/Academic Press; 2017. pp. 28–38.
15 Gill D, Dorevitch A, Marks R. The prevalence of seborrheic keratoses in people aged 15 to 30 years: is the term senile keratosis redundant? *Arch Dermatol* 2000; **136**:759–762.
16 Scrivener Y, Grosshans E, Cribier B. Variations of basal cell carcinomas according to gender, age, location and histopathological subtype. *Br J Dermatol* 2002; **147**:41–47.
17 Paraskevas LR, Halpern AC, Marghoob AA. Utility of the Wood's light: five cases from a pigmented lesion clinic. *Br J Dermatol* 2005; **152**:1039–1044.
18 McGuire L, Disa J, Lee E, Busam K, Nehal K. Melanoma of the Lentigo Maligna Subtype. *Plast Reconstruct Surg.* 2012; **129**:288e–299e.
19 Holmes GA, Vassantachart JM, Limone BA, Zumwalt M, Hirokane J, Jacob SE. Using dermoscopy to identify melanoma and improve diagnostic discrimination. *Fed Pract* 2018; **35**:S39–S45.