
UNIVERSIDADE D
COIMBRA

Pedro Miguel Marques Pereira

# Skin Lesion Assessment based on Plenoptic Images for Melanoma Classification

**Tese no âmbito do Programa de doutoramento em Ciências e Tecnologias da Informação orientada pelo Professor Doutor Rui Pedro Pinto de Carvalho e Paiva, e pelo Professor Doutor Sérgio Manuel Maciel de Faria, pelo Professor Doutor Rui Manuel da Fonseca Pinto, e apresentada ao Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologias da Universidade de Coimbra.**

Agosto de 2021

Faculty of Sciences and Technology
Department of Informatics Engineering

# Skin Lesion Assessment based on Plenoptic Images for Melanoma Classification

Pedro Miguel Marques Pereira

Thesis under Doctoral program in Information Sciences and Technologies guided by Professor Doctor Rui Pedro Pinto de Carvalho e Paiva, and by Professor Doctor Sérgio Manuel Maciel de Faria, by Professor Doctor Rui Manuel da Fonseca Pinto, and presented to the Department of Computer Engineering of the Faculty of Sciences and Technologies of the University of Coimbra.

August 2021


UNIVERSIDADE D
COIMBRA

**Author:**
Pedro Miguel Marques Pereira
Department of Informatics Engineering, University of Coimbra (UC)
Multimedia Signal Processing Group, Instituto de Telecomunicações

**Supervisors:**
Prof. Rui Pedro Pinto de Carvalho e Paiva
Centre for Informatics and Systems of the University of Coimbra
Department of Informatics Engineering, University of Coimbra (UC)

Prof. Sérgio Manuel Maciel de Faria
Multimedia Signal Processing Group, Instituto de Telecomunicações
School of Technology and Management (ESTG), Polytechnic of Leiria

Prof. Rui Manuel da Fonseca Pinto
Multimedia Signal Processing Group, Instituto de Telecomunicações
School of Technology and Management (ESTG), Polytechnic of Leiria

This work was supported under FCT the PhD. Grant SFRH/BD/128669/2017 and by the projects DermoPleno (PEst-OE/EEI/LA0008/2016) and PlenoISLA (PTDC/EEI-TEL/28325/2017), in the scope of R&D Unit 50008 (UIDB/EEA/50008/2020), financed by FCT/MEC through national funds, co-funded by a partnership agreement between FEDER, Portugal 2020, and COMPETE 2020.

# Acknowledgement

I reserve this page to thank all whom aided me. The present work would not have been possible without the support of various institutions and the help of many people.

My thanks go firstly to **Professor Sérgio Manuel Maciel de Faria** for pulling me to this project and for his patience and general psychological support. Then to **Professor Rui Manuel da Fonseca Pinto** for his engagement and ideas that ultimately guided the initial PhD. research. Finally to **Professor Rui Pedro Pinto de Carvalho e Paiva**, who became the bridge between this PhD. thesis and the host university, for his remote meetings, which greatly helped me stay positive about the PhD. stance and prospects, and for his work guidance and reviews.

I would also like to express my gratitude to **Professor Pedro Antonio Amado de Assunção** for his suggestions and observations during this research period. And **Professor Luís Miguel Oliveira Pegado de Noronha e Távora**, who always had an hands-on approach and enabled several advances.

To the already mentioned, I would like to thank their over 250 weeks of contributions and all the time spent reviewing my manuscripts, specially to **Professor Lucas Arrabal Thomaz** who is a valuable colleague and brought harmony to the environment workflow. Thank you.

In addition, I would like to thank **all my colleagues** at Instituto de Telecomunicações, Leiria, for their contributions, recommendations, and support during the research activities. It is worth mentioning some of the relevant milestones enabled by some of these peers, without which the contents of this dissertation would not come to exist: **Miguel Oliveira Santos**, for the creation of the final acquisition equipment prototype and, jointly with **José Nunes Dos Santos Filipe**, the initial acquisition and organisation of the dataset, now named *Light Field Image Dataset of Skin Lesions* (SKINL2); **João Oliveira Parracho**, for the continued acquisition of new dataset samples at the local hospital; and **Francisco Ferreira Cunha**, for the adjustments performed to the acquisition equipment to enable more lens magnification and surface quality.

I gratefully acknowledge the funding sources that made my PhD. work possible. This work was initially carried out under the host institution **Instituto de Telecomunicações**, in the scope of project DermoPleno, **PEst-OE/EEI/LA0008/2016**, and later in the scope of project PlenoISLA, **PTDC/EEI-TEL/28325/2017**, both in the scope of **R&D Unit 50008**, through national funds and where applicable co-funded by **FEDER – PT2020 – COMPLETE2020** partnership agreements. In addition, my work was directly supported by the Portuguese **Fundação para a Ciência e Tecnologia** under the PhD. grant **SFRH/BD/128669/2017**.

Last, but definitely not least, I also want to thank my beloved and cherished parents, **Cristina Maria Rainho de Pimentel Teixeira Marques Pereira** and **Pedro Luis Lopes Pereira** for letting me work after hours whenever it was needed without any concerns or preoccupations and for providing to my every need so that I could concentrate my daily life on this work. To my parents and my girlfriend, **Ana Marina Rodrigues Marcelino**, I also thank for their predisposition and encouragement to see the end of this latest academic step.

# Abstract

Over the years, image processing algorithms have achieved many advancements in the medical imaging area, namely in skin lesion detection and classification. Still, skin cancer has maintained its position at the top of the most common cancers all over the world. Early detection of suspicious pigmented skin lesions has a determinant role in clinical prognosis. Among them, melanoma, a malignant type of skin lesions, is the one that causes the most deaths.

Several research works have moved forward the methodology and tools employed by expert dermatology clinicians. Currently, most experts employ a dermatoscope in naked eye examination. However, in recent years, some public datasets of dermoscopy images have emerged, enabling researchers to develop, validate, and assess new computer-aided methods. Such methods include: pre-processing algorithms, aimed at removing artefacts and applying transformations necessary for the following algorithms; segmentation methods, that aim at identifying and separating healthy skin from the lesion region; and classification or recognition methods, which aim at detecting key lesion characteristics and even devise the lesion type. However, none of these methods provide sufficient robustness for widespread usage.

In the pursuit for further advancements in this field, this thesis addresses and improves current segmentation and classification algorithms, provides a new evaluation tool for dermatology experts and researchers (by introducing a light-field dataset of skin lesion images to the field), and proposes several approaches based on algorithms capable of differentiating melanoma from non-melanoma images using 2D and 3D features.

Tackling the challenges in the literature, this thesis first proposes two segmentation approaches, while also performing extensive comparisons with other works, across multiple datasets and performance metrics. From this endeavour, evidence that segmentation-detail can contribute for melanoma discrimination is presented.

Using the Light-field Image Dataset of Skin Lesions (SKINL2), with images collected at the Department of Dermatology of Centro Hospitalar de Leiria (Portugal), several methods are presented as the key contributions of this thesis. First, the acquired skin surface depth is explored, confirming that the use of depth data presents relevant information for melanoma classification (data not present in 2D colour images). Then, further steps are taken to exploit both colour and depth information under a joint process, whilst maintaining the capability of showing the depth contribution to the classification performance. In any of these steps, proposed approaches provide results superior the current state-of-the-art, when applied to the SKINL2 dataset.

**Keywords:** *Medical Image Analysis, Dermoscopy, Skin Lesions, Melanoma, Medical dataset, Image Segmentation, Feature Extraction, Image Classification*

# Resumo

Ao longo dos anos, a área de deteção e classificação de lesões cutâneas sofreu diversos avanços. Ainda assim, o cancro de pele manteve a sua posição como um dos mais comuns em todo o mundo. A deteção atempada de lesões cutâneas pigmentadas tem um papel determinante no prognóstico clínico. Dentre os tipos de lesão, o melanoma, um tipo maligno de lesões cutâneas, é o cancro de pele que causa mais mortes.

Diversos trabalhos impulsionaram as metodologias e ferramentas utilizadas por dermatologistas. Atualmente, os especialistas recorrem a um dermatoscópio para realizar o seu exame visual – um instrumento portátil para rastreio de lesões da pele. Desta forma, nos últimos anos, algumas bases de dados de imagens dermatoscópicas públicas surgiram para permitir a pesquisa, validação e avaliação de novos métodos com recurso ao computador. Esses métodos variam entre: algoritmos de pré-processamento, que visam remover artefactos e aplicar as transformações necessárias aos métodos seguintes; métodos de segmentação, com o objetivo de identificar e separar a pele saudável da lesão de pele; e métodos de classificação ou reconhecimento, que visam detetar características da lesão ou definir o seu tipo. No entanto, nenhum deles confere robustez suficiente para que se possa assumir um uso generalizado.

Na busca de novos avanços, esta tese: aborda algoritmos de segmentação e classificação; fornece uma nova ferramenta para especialistas e pesquisadores de dermatologia (introduzindo uma base de dados de *light-field* da pele); e introduz abordagens experimentais através de algoritmos capazes de diferenciar melanomas de outras lesões, usando informações 2D e 3D.

Esta tese propõe duas abordagens de segmentação, acompanhadas de comparações extensas com outros trabalhos, usando várias bases de dados e métricas de desempenho. Desta forma, mostra-se que contornos da segmentação podem contribuir para a discriminação do melanoma.

Utilizando a base de dados de imagens *light-field* de lesões cutâneas (SKINL2), cujas imagens foram adquiridas no Departamento de Dermatologia do Centro Hospitalar de Leiria, Portugal, vários métodos são apresentados – compondo as principais contribuições desta tese. Em primeiro lugar, o relevo da pele adquirida é explorado, confirmando que o seu uso adiciona capacidades relevantes para a classificação do melanoma. Em seguida são tomadas medidas adicionais para unir as informações de cor e profundidade no mesmo processo de classificação, mantendo-se a capacidade do modelo mostrar a contribuição da profundidade para o processo. Em qualquer uma dessas medidas, as abordagens propostas fornecem resultados superiores aos do estado do conhecimento atual (quando aplicadas à base de dados SKINL2).

**Palavras-chave:** *Análise de Imagens Médicas, Dermatoscopia, Lesões Cutâneas, Melanoma, Dataset Médico, Segmentação de Imagens, Extração de Características, Classificação de Imagens*

# Foreword

THE work performed in this thesis was accomplished within the Multimedia Signal Processing group of the Instituto de Telecomunicações (Leiria, Portugal) in cooperation with the Centre for Informatics and Systems of the University of Coimbra (Coimbra, Portugal), in the context of the following projects:

- **Project DermoPleno:** Dermo-Plenoptic Imaging for Skin Lesion Assessment (**PEst-OE/EEI/LA0008/2016**), in the scope of R&D Unit 50008 (**UIDB/EEA/50008/2020**), financed by Fundação para a Ciência e Tecnologia (FCT) and Ministério da Educação e Ciência (MEC) through national funds, co-funded by a partnership agreement between Fundo Europeu de Desenvolvimento Regional and Portugal 2020. This project focused on research and development of new methods and tools for acquisition, computational analysis, and efficient coding of plenoptic medical imaging content, specifically related to melanoma. The project takes advantage of the most recent advances in imaging technology by relying upon melanoma plenoptic high resolution images, characterised by containing both spatial RGB intensity and directional information of light rays, captured by a light-field (or plenoptic, or holoscopic) camera.

- **Project PlenoISLA:** Plenoptic Imaging for Skin Lesion Assessment (**PTDC/EEI-TEL/28325/2017**), in the scope of R&D Unit 50008 (**UIDB/EEA/50008/2020**), financed by FCT/MEC through national funds, co-funded by a partnership agreement between Fundo Europeu de Desenvolvimento Regional and Portugal 2020, Programa Operacional Competitividade e Internacionalização (COMPETE 2020). This project aimed at the development and implementation of new imaging techniques for non-invasive skin surface characterization, based on recent light-field imaging technology, with the objective to obtain a new set of 3D-based quantitative markers of the skin, allowing the characterization of its morphological and functional structure, that can then be used for prognosis in Dermatology studies. Besides the computational analysis, new techniques for acquisition and compression of the light-field images are also addressed, as well as the perpetuation of the creation of a database of skin lesion plenoptic images for the scientific community by partnering with Centro Hospitalar de Leiria.

In addition, the work addressed in this document was financed by Fundação para a Ciência e Tecnologia, under PhD. Grant **SFRH/BD/128669/2017**.

The research outcome of this work resulted in the following publications:

**Journal papers:**

- **Pereira, P. M. M.**, Fonseca-Pinto, R., Paiva, R. P., Assuncao, P. A. A., Tavora, L. M. N., Thomaz, L. A., and Faria, S. M. M., Dermoscopic skin lesion image segmentation based on Local Binary Pattern Clustering: Comparative study, Biomedical Signal Pro-

cessing and Control, vol.59, pp.1–12, **Q2**, **IF**[1]**:3.137**, **H**[2]**:60**, 2020 (DOI: 10.1016/j.bspc.2020.101924).

- **Pereira, P. M. M.**, Fonseca-Pinto, R., Paiva, R. P., Assuncao, P. A. A., Tavora, L. M. N., Thomaz, L. A., and Faria, S. M. M., Skin lesion classification enhancement using border-line features – The melanoma vs nevus problem, Biomedical Signal Processing and Control, vol.57, pp.1-8, **Q2**, **IF:3.137**, **H:60**, 2020 (DOI: 10.1016/j.bspc.2019.101765).
- **Pereira, P. M. M.**, Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., and Faria, S. M. M, Multiple Instance Learning using 3D Features for Melanoma Detection, IEEE Journal of Biomedical and Health Informatics, –, pp.–, **Q1**, **CORE**[3] **A***, **IF:5.223**, **H:115**, 2021 *(submitted).*
- **Pereira, P. M. M.**, Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., and Faria, S. M. M., Melanoma Classification using Light-Fields with Morlet Scattering Transform and CNN: surface depth as a valuable tool to increase detection rate, Medical Image Analysis, Special Issue on Image Analysis in Dermatology, –, pp.–, **Q1**, **IF:11.148**, **H:122**, 2021 *(submitted).*

**Conference papers:**

- **Pereira, P. M. M.**, Tavora, L. M. N., Fonseca-Pinto, R., Paiva, R. P., Assuncao, P. A. A., and Faria, S. M. M., Image Segmentation using Gradient-based Histogram Thresholding for Skin Lesion Delineation, 12th International Joint Conference on Biomedical Engineering Systems and Technologies, vol.2, pp.84-91, Prague, Czech Republic, February, 2019 (DOI: 10.5220/0007354100840091).
- **Pereira, P. M. M.**, Fonseca-Pinto, R., Paiva, R. P., Tavora, L. M. N., Assuncao, P. A. A., and Faria, S. M. M., Accurate Segmentation of Dermoscopic Images based on Local Binary Pattern Clustering, 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, pp.314-319, Opatija, Croatia, February, 2019 (DOI: 10.23919/MIPRO.2019.8757023).
- Faria, S. M. M., Santos, M., Assuncao, P. A. A., Tavora, L. M. N., Thomaz, L. A., **Pereira, P. M. M.**, Fonseca-Pinto, R., Santiago, F., Dominguez, V., and Henrique, M., Dermatological Imaging using a Focused Plenoptic Camera: the SKINL2 Light Field Dataset., Conference on Telecommunications, pp.1-4, Lisbon, Portugal, June, 2019 (Link: https://www.it.pt/Publications/PaperConference/34765).
- Faria, S. M. M., Filipe, J. N., **Pereira, P. M. M.**, Tavora, L. M. N., Assuncao, P. A. A., San-tos, M. O., Fonseca-Pinto, R., Santiago, F., Dominguez, V., and Henrique, M., Light Field Image Dataset of Skin Lesions, 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp.3905-3908, Berlin, Germany, **CORE C**, July, 2019 (DOI: 10.1109/EMBC.2019.8856578).

[1]Impact Factor (IF) from ™Thomson Reuters
[2]H-Index metric
[3]Computing Research and Education Association of Australasia (CORE) ranking

- **Pereira, P. M. M.**, Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., and Faria, S. M. M., Skin Lesion Classification using Bag-of-3D-Features, Conference on Telecommunications, pp. 1-6, Leiria, Portugal, February, 2021 (DOI: 10.1109/ConfTELE50222.2021.9435509).
- **Pereira, P. M. M.**, Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., and Faria, S. M. M. Skin lesion classification using features of 3D border lines, 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp.1-6, Guadalajara, Mexico, **CORE C**, October, 2021.

**Poster Sessions:**

- **Pereira, P. M. M.**, Dermo-Plenoptic Imaging for Skin Surface Assessment, Presented at Encontro Ciência, Lisbon, Portugal, 2018.
- **Pereira, P. M. M.**, Melanoma Detection based on Light-Field Imaging, Presented at Encontro Ciência, Lisbon, Portugal, 2018.
- **Pereira, P. M. M.**, Skin Lesion Classification using Light-Field Imaging, Presented at Medical Imaging Summer School - Medical Imaging Meets Deep Learning, Favignana, Sicily, 2018.

**Other Publications:**

- Faria, S. M. M., Filipe, J. N., Assuncao, P. A. A., Santos, M. O., Fonseca-Pinto, R., **Pereira, P. M. M.**, Tavora, L. M. N., Santiago, F., Dominguez, V., and Henrique, M., Coding of Still Pictures, Doc. ISO/IEC JTC1/SC29/WG1 M82037, January, 2019.

# Index

# List of Figures

# List of Tables

# List of Abbreviations and Acronyms

| | |
|---|---|
| **ACC** | Accuracy |
| **ANN** | Artificial Neural Networks |
| **AQ** | Wan Quantifier |
| **AS** | Adaptive Snake |
| **Atlas** | EDRA-Interactive Atlas of Dermoscopy |
| **BAC** | Balanced-Accuracy |
| **BE** | Border Error |
| **BoF** | Bag-of-Features |
| **BSc.** | Bachelor of Sciences |
| **BT** | Bradley Threshold |
| **CAD** | Computed Assisted Diagnosis |
| **CDNN** | Convolutional-Deconvolutional Neural Network |
| **CH** | Chan-Vese |
| **CIE** | International Commission on Illumination |
| **cm** | centimetre |
| **CV** | Cross-Validation |
| **DCL-PSI** | Deep Class-specific Learning with Probability based Step-wise Integration |
| **DermoNet** | Densely Linked Convolution Neural Network |
| **DL** | Deep Learning |
| **DNN** | Deep Neural Networks |
| **DFT** | Discrete Fourier Transform |
| **DSc.** | Doctor of Science |
| **ECDNN** | Enhanced CDNN |
| **ECG** | Electrocardiogram |
| **EM-LS** | Expectation-Maximization LevelSet |
| **ESF** | Ensemble of Shape Functions |
| **FBSM** | Fuzzy-Based Split and Merge |
| **FC-LS** | Fuzzy Clustering LevelSet |
| **FCM** | Fuzzy C-Means |
| **FCN** | Fully Convolutional Networks |
| **FCT** | Fundação para a Ciência e a Tecnologia |
| **FDEE** | Fuzzy Differential Evolution Entropy |
| **FDR** | False Detection Rate |

**FNN** Feedforward Neural Network

**FNR** False Negative Rate

**FPFH** Fast Point Feature Histogram

**FPR** False Positive Rate

**GASD** Globally Aligned Spatial Distribution

**GHT** Gradient-based Histogram Thresholding

**GM** Geometric-Mean

**GRSD** Global RSD

**GT** Ground-Truth

**HD** Hammoude Distance

**HSV** Hue-Saturation-Value

**HMM** Hidden Markov Models

**ICD10** International Classification of Diseases

**ID** Identifier

**ILSVRC** ImageNet Large Scale Visual Recognition Challenge

**ISDA** Iterative Single Data Algorithm

**ISS** Intrinsic Shape Signatures

**IT** Iterative Threshold

**JCLMM** Joining Circular-Linear Mixture Model

**JI** Jaccard Index

**JPEG** Joint Photographic Experts Group

**KL** Kullback-Leibler

**KL-LS** Kullback-Leibler based LevelSets

**KL-PLS** Kernel Logistic Partial Least Square Regression

**KMC** K-Means Colour

**KMS** K-Means Colour and Spatial

**KT** Kapur Threshold

**LBP** Local Binary Patterns

**LBPC** Local Binary Patterns Clustering

**LED** Light-Emitting Diode

**LMS** Lankton Mean Separation

**LT** Li Threshold

**MAvsBE** Malignant versus Benign

**MC** Mean-shift Colour

**MCS** Mean-shift Colour and Spatial

| | |
|---|---|
| **MEC** | Ministério da Educação e Ciência |
| **mFCN** | multi-stage FCN |
| **MIL** | Multiple Instance Learning |
| **ML** | Machine Learning |
| **MLA** | Micro-Lenses Array |
| **mm** | millimetre |
| **MSc.** | Masters of Sciences |
| **MSCA** | Multi-scale Superpixel with Cellular Automata |
| **MT** | Moment Threshold |
| **MvsAll** | Melanoma versus All other |
| **MvsN** | Melanoma versus Nevus |
| **NARF** | Normal Aligned Radial Feature |
| **NCA** | Neighborhood Component Analysis |
| **NQ** | Neural Quantization |
| **OT** | Otsu Threshold |
| **PC** | Principal Curvatures |
| **PCT** | Principal Components Transform |
| **PCT-MC** | PCT Median Cut |
| **PFH** | Point Feature Histogram |
| **PH$^2$** | Pedro Hispano Hospital |
| **PhD.** | Doctor of Philosophy |
| **pp** | percentage points |
| **Prof.** | Professor |
| **PS** | Proposed Segmentation |
| **PSL** | Pigmented Skin Lesions |
| **PSO-DEN** | Particle Swarm Optimization on Deep Ensemble Network |
| **RMS** | Root-Mean-Square |
| **RGB** | Red / Green / Blue |
| **ROI** | Region of Interest |
| **RT** | Renyi Threshold |
| **RGB-MC** | RGB Median Cut |
| **RSD** | Radius-based Surface Descriptor |
| **RIFT** | Rotation Invariant Feature Transform |
| **SC3D** | 3D Shape Context |
| **SCDRR** | Sparse Coding with Dynamic Rule-based Refinement |

**SDC** Sørensen-Dice coefficient

**SEN** Sensitivity

**SHOT** Signature of Histograms of OrienTations

**SKINL2** Skin Lesion Light-fields

**SLIC** Simple Linear Iterative Clustering

**SPE** Specificity

**SRM** Statistical Region Merging

**SSLS** Saliency-based Skin Lesion Segmentation

**ST** Shanbhag Threshold

**SVM** Support Vector Machine

**SWSDB** Semi- and Weakly Supervised Directional Bootstrapping

**TBP** Total-Body Photography

**TDR** True Detection Rate

**TNR** True Negative Rate

**TPR** True Positive Rate

**TL** Transfer Learning

**UDA** Unified Dermoscopy Algorithm

**UIF** Uncertainty-Infused Function

**UQ** Wu Quantifier

**USC** Unique Shape Context

**UT** Huang Threshold

**VV** Chan-Vese Vector

**WB** Working Blocks

**WS** Wavelet Scattering

**YCbCr** Luminance - Blue Difference Chroma - Red Difference Chroma

**YT** Yen Threshold

**YUV** Luminance - Blue Chrominance - Red Chrominance

**Z** Depth (Layer/Information)

# Chapter 1

# Introduction

Content



Despite the huge advances in algorithms for detection and classification of objects in images, some applications present challenges that require novel approaches. One of these fields is related to the detection of melanoma skin cancer. Although not being the most common type of skin cancer, is the most likely to grow and spread to other organs. In spite of its lethality, if detected at an early stage, the melanoma can be removed through minor surgery. Therefore, much effort has been made to improve the early detection of such lesions. One of the main issues with melanoma detection is that it is visually similar to another skin abnormality called nevus, which is a very common benign type of skin lesion in the general population. Dermatologists (currently using dermoscopy as the main type of skin cancer screening technology), and even current computer systems, still struggle to discriminated melanoma from atypical nevus skin lesions – while other lesion types are more easily classified.

In this context, the aim of this thesis is to address the issue of melanoma classification, by exploring a new acquisition method that enables the extraction of information related to the depth of the skin surface in order to improve melanoma detection rates. This depth surface allows for the extraction of novel (computer vision) features and to further the understanding regarding skin cancer in the existing literature. For this purpose, a new dataset of skin images was acquired prior to study such depth dimension. In parallel to the acquisition of the skin image dataset, other preliminary research was performed on segmentation methods for dermoscopy related images, including their comparison to relevant literature and their evaluation on the skin lesion classification process (Chapter 3). Afterwards, by using the new dataset (of both colour and skin depth), a new research path has been devised using depth information only. This is made to assess the gains achieved with this new dimension and to exploit the 3D characteristics in the skin lesion surface, thus advancing beyond common 2D features. Specifically, two different classification approaches were proposed and compared with the state-of-the-art (Chapter 4). Finally, having validated the usefulness of the third dimension, further steps were taken to join both colour and depth information under the same classification process, whilst maintaining the capability of showing the depth performance contribution to the process – with the proposal of two classification approaches (Chapter 5).

The remainder of this chapter presents the research motivations of this thesis (Section 1.1) followed by the detailed proposed objectives and contributions (Section 1.2), as well as a thesis outline (Section 1.3).

## 1.1 Motivation and Problem Statement

The complexity of computer models and pattern recognition algorithms has advanced significantly over the years. However, for the last decades, skin cancer has maintained its position at the top of the most common cancers all over the world (Alliance, 2020). A skin lesion is any kind of skin patch that presents different characteristics when compared to its surrounding area. There are many types of skin lesions, which can be described according to their type, configuration, texture, colour, localisation, and distribution, among other clinical signs. Generally, studies tend to focus on pigmented skin lesions (PSL), namely the melanocytic lesions – as does this thesis. This type of lesions is primarily denoted as an abnormal proliferation of melanocytes at the basal epidermis or upper dermis layers that may ultimately be classified as benign or malignant (Cichorek et al., 2013). Its classification is typically based on dermatologists' visual inspection, with support of dermoscopic imaging and the diagnosis by skin biopsy (Vestergaard et al., 2008).

Around the world, dermatologist work force shortage and the uneven global distribution of pathology lab facilities set up the main reasons for the lack of access to prompt detection of skin cancer, contributing to the increased morbidity and melanoma mortality (Feng et al., 2018). Melanoma diagnosis rates have increased dramatically over the past three decades, outpacing almost all other cancers (Alliance, 2020). As of 2020, in the USA, the risk of developing melanoma was of 1 in 38 (2.6%) for Whites, 1 in 1000 (0.1%) for Blacks, and 1 in 167 (0.6%) for Hispanics (Society, 2020). A classical method to identify melanoma is with parameters known as Asymmetry, Border, Colour, and Diameter – coined the "ABCD" rule (Soyer et al., 2004). This method is based on the principle that melanoma lesions are typically asymmetric, are larger than $6mm$ in diameter, have irregular borders, and tend to have more than one colour. Additionally, one-third of all melanomas are thought to arise from pre-existing nevus (a, sometimes, visually similar lesion but of benign origin) – thus detection and removal of such nevus is of utmost importance in the prevention of melanoma (Pampena et al., 2017). The process of lesion identification by specialists is labour intensive, time costly, and error prone, therefore, it could be improved with the use of automated methods.

Following the previous arguments, it is then clear that early detection of suspicious PSL has a determinant role in the clinical treatment. Therefore, non-invasive *in vivo* imaging techniques (e.g. total body photography, automated diagnostic systems, reflectance confocal microscopy, and dermoscopy) have been applied on the development of reliable systems to assist in the clinical diagnosis decision (Smith & MacNeil, 2011). Concerning early detection of melanoma, new solutions of Computer-Aided Diagnosis (CAD) based on Machine Learning (ML) tech-

niques have also been investigated for feature extraction and pattern classification (Korotkov & Garcia, 2012; Lalitha & Geetha, 2014). Beyond pattern analysis, the use of texture as a clue for melanocytic classification has been widely discussed in the literature (Malvehy et al., 2007). Some techniques like texture analysis using statistical methods (Materka & Strzelecki, 1998; Sheha et al., 2012; Riaz et al., 2014; Pereira et al., 2015), model-based methodologies (Materka & Strzelecki, 1998), or bank filters in both frequency and in space-frequency domain (Materka & Strzelecki, 1998; Machado et al., 2016) achieved promising results, showing evidence that texture features are a key for successful image characterisation (within a set of different types of features). However, these 2D image processing approaches (using contact dermoscopy), when analysed by ML algorithms, have revealed a wide range of performance results (Korotkov & Garcia, 2012). These discrepancies are due to the use of different metrics to show the results, to the variable contact pressure of a dermoscope (changing the texture), and due to misleading identification of the most discriminant features in PSL. Beyond the 2D analysis, few works have also exploited the 3D information of PSL. The first attempt to reconstruct 3D images was made with the introduction of the "Nevoscope" (Dhawan et al., 1984). As the result of several consecutive reconstructions, lesion changes were evaluated in regard to thickness, size, colour, and structure. Later characterisation approaches were attempted with photometric stereo imaging (Anwar et al., 2012), allowing to extract features for lesion classification like skin tilt and slant patterns (Smith et al., 2011), and statistical moments of enhanced principal curvatures of skin surfaces (Zhou et al., 2010, 2011).

## 1.2 Objectives and Scientific Contributions

The core work of this thesis is on the research and development of new methods for identification and classification of skin lesions, with the particular scope of discriminating melanomas from other types of lesions. The aim of this thesis is to advance the current state-of-the-art by exploiting light-field (plenoptic) image data in order to extract new 3D melanocytic surface information, in pursuit of improved ML classification results. The major objectives of this thesis are as follows:

**Goal 1 -** Research and develop new feature extraction algorithms and classification approaches to improve the state-of-the-art methods associated with melanoma skin lesions;

**Goal 2 -** Create, structure, and make available a dataset of light-field images featuring melanocytic skin lesions, so to allow international cooperation and further validation with international partners;

**Goal 3 -** Research, evaluate, and benchmark different melanoma detection approaches in order to propose computer-based algorithms capable of differentiating melanoma from non-melanoma images using 2D and 3D features from the novel light-field dataset of skin lesions.

Taking into consideration the main goals, the work that led to this thesis has produced the

following main contributions:

**Contribution 1: Segmentation approaches**

This contribution addresses certain types of segmentation methods absent in the literature, namely: a segmentation approach to extract a detailed lesion border; and an approach that behaves like a dermatologist. The development of this task resulted in the proposal of both a segmentation method, aimed at extracting a more spatially detailed skin-to-lesion separation (published in Pereira et al., 2019a), and a dermatologist-like segmentation method based mainly on colour gradients (published in Pereira et al., 2019b) – presented in Sections 3.2 and 3.1, respectively. The comparison with other literature works is not always easy, especially when different types of image datasets are used and the results are expressed in different metrics. Therefore, a comparison of a broad spectrum of segmentation methods was also performed, based on multiple datasets and various performance metrics. This contribution, published in Pereira et al. (2020a), is presented in Section 3.2.

**Contribution 2: Assessment of segmentation-detail importance for classification**

This contribution evidences that segmentation-details can contribute for melanoma discrimination when the created segmentation masks are used to obtain border-line features for the later classification process. First, in Section 3.4 (Pereira et al., 2020b), the two previously proposed image segmentation methods are exploited to provide border-line features from colour (2D) images. The achieved results confirm that a more spatially detailed border-line definition provides better classification performance, and that these features contribute to improve the performance of existing colour based classification algorithms. Then, in Section 4.3 (Pereira et al., 2021d), a similar experiment is also performed in the 3D domain, where extracted border-line features are obtained from the depth information instead of color. In this case, the achieved results also confirm that higher discrimination is achieved between the targeted classes, when border-line features are added to the classification pipeline.

**Contribution 3: Creation of a light-field dataset of skin lesions**

Light-field images were acquired at the Department of Dermatology of Hospital of Leiria and compiled into a publicly available dataset to enable further advancements in the field of skin lesion classification and light-field technology in general. Currently, there are two published (and publicly available) versions of the dataset (Faria et al., 2019c,a). A third version also exists, however it is still in the acquisition phase, despite being already publicly available. This contribution is presented in Section 4.1.

**Contribution 4: Show the discriminative power of skin surface for classification**

In order to investigate if the skin surface depth itself has discriminative information in regard to melanomas, a first algorithm was proposed, as published in Pereira et al. (2021c) and presented in Section 4.2, with the capability of selecting relevant features from a broad set of generic extractors. Then, a second algorithm was also proposed (in Pereira et al., 2021d, and presented in Section 4.3) showing that, even when using depth information alone, lesions' border-line surface detail has relevant information for

the melanoma discrimination process. In addition to these two algorithms, two other approaches were developed in order to provide an algorithm aimed at melanoma discrimination. The first, published in Pereira et al. (2021b) and presented in Section 5.1, is a step in the direction of merging current 2D state-of-the-art results with evaluated 3D characteristics. The second algorithm, published in Pereira et al. (2021a) and presented in Section 5.2, was created to be a single model capable of performing melanoma discrimination independently of the use of either colour, depth, or both information.

## 1.3 Structure

The remainder of this thesis is organised in five chapters, as follows:

**Chapter 2 - State-of-the-Art**

This chapter introduces the main concepts and covers a set of background notions necessary for the understanding of the literature and the remaining chapters. It starts by giving some details on the target of this work: melanoma, skin cancer; followed by the enumeration of several publicly available datasets and relevant metrics for this work. Then, it covers the steps that have become common in the process of skin lesion classification (where melanoma is included), and references previous works that feature 3D image acquisition and classification of skin lesions. Finally, some information on light-field technology is also presented.

**Chapter 3 - Segmentation and Classification of 2D Images**

This chapter is dedicated to the developments accomplished with dermoscopic and macro skin lesion images. In the absence of methods aiming for (round) dermatologist-like and (detailed) computer driven segmentation, a method was devised for each of these types. Then, to evaluate the proposed methods, comparisons to the segmentation methods in the literature are made and their efficiency for classification evaluated. Additionally, this chapter describes an initial assessment of the classification performance of a Deep Learning (DL) algorithm when using Transfer Learning (TL).

**Chapter 4 - Contributions using 3D Depth Maps**

Regarding the dataset of light-field images, this chapter presents the hardware and acquisition setup details of the so-called SKINL2 dataset, as well as its details regarding the collected data. Then, two algorithms are investigated and developed to show that the newly extracted depth dimension comprises information relevant to address the problem of melanoma classification.

**Chapter 5 - Towards Melanoma Classification**

This chapter is dedicated to the results achieved by the use of both colour and surface (depth) information for melanoma classification. Two classification approaches are designed to enable comparison between the use of 3D information, namely: one by using a

colour method combined with the extracted surface information; and a second model capable of performing melanoma discrimination independently of the use of either colour, depth, or both sources of information.

**Chapter 6 - Conclusion**

This chapter summarises the final remarks and conclusions taken from this thesis, as well as the research directions for future works.

# Chapter 2

# State-of-the-Art

Content



This chapter provides some of the necessary background to better understand the remaining chapters. Before embarking on the relevant state-of-the-art for this work it is important to understand its base target, i.e., melanoma, skin cancer, which is detailed in Section 2.1.

In order to enable the broad study of PSL attributes, some image datasets have been made publicly available by others in the past. These datasets can be used as basis for PSL research works (Section 2.2).

Within the context of this thesis, two main tasks were carried out: segmentation and classification of PSL. Depending on the targeted domain, several metrics exist to enable evaluation,

comparison, and selection of methods and algorithms to be used on those datasets – depending on their aim (as detailed Section 2.3).

The current literature for 2D melanoma imaging is well described in Korotkov & Garcia (2012), Oliveira et al. (2016), and Pathan et al. (2018). As it can be inferred from the previous literature works, it is evident that the classification of skin lesion related images includes five steps (detailed in Section 2.4), namely: image acquisition, pre-processing, segmentation, feature extraction, and classification. Particularly for the topic of segmentation, due to its importance for this work, a relevant set of methods is presented and discussed in Section 2.5.

Besides simple 2D imaging, as presented in existing literature, it is also possible to extract 3D information from the skin lesion to improve the 2D classification results (Section 2.6). This can be done by using different acquisition technologies. In the scope of this thesis, the acquisition of such 3D information is performed by resorting to light-field imaging (Section 2.4.1-Light-Fields).

The remainder of this chapter reviews the literature pertaining to each of the mentioned key background topics.

## 2.1 Skin and Skin Cancer

Skin is the largest organ in the human body and consists of two main layers: the epidermis and the dermis (Fig. 2.1). The dermis is composed of sub-layers of collagen and elastic fibers. It provides energy and nutrition to the epidermis. The later is a stratified squamous epithelium, a layered scale-like tissue, which serves as protection against external agents. It consists of 4 types of cells (McGrath & Uitto, 2010):

- Keratinocytes, representing 95% of the epidermis.
- Melanocytes, dendritic cells that distribute packages of melanin pigment (via melanosomes) to surrounding keratinocytes (to give skin and hair their colour).
- Langerhans cells, also dendritic cells, that exist to detect foreign bodies that penetrate the epidermis and deliver them to the local lymph nodes.
- Merkel cells, which function as receptors to the touch sensation.

Although almost any cell in the body can develop cancer, some are more cancer-prone than others. In the case of skin, most cancers develop from non-pigmented cells and not from pigmented melanocytes (Kaufman, 2005). This makes the known 'basal cell carcinoma' and 'squamous cell carcinoma' the most common forms of skin cancer. However, 'melanoma' (also called 'malignant melanoma') is a less common but far more deadly skin cancer. Most melanomas possess at least one type of dermoscopic structures: atypical networks, peripheral streaks, atypical dots or globules, negative pigment network, off-center pigmented blotches,

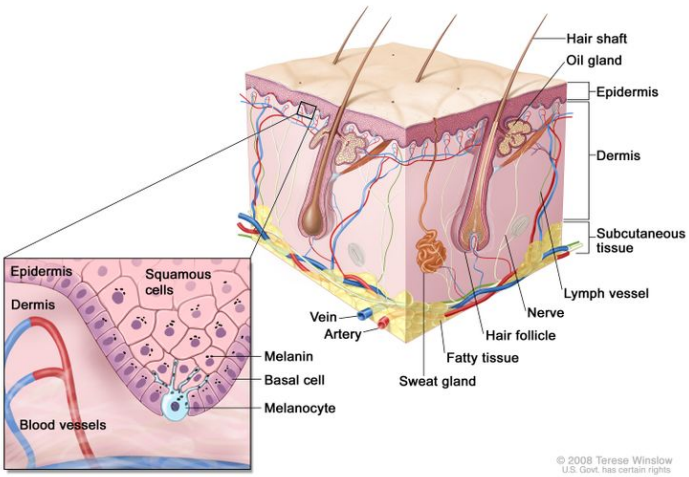



**Figure 2.1:** Anatomy of the skin, showing the epidermis, the dermis, and subcutaneous (hypodermic) tissue. Copyright 2008 by Terese Winslow at http://www.teresewinslow.com/portshow.asp?portfolioid={4B56C61F-9C24-47C6-9F4D-9444E1D75BA2}.

blue-white veil over raised or flat areas, atypical vascular structures, and peripheral brown structureless areas.

The cancerous growths develop when the damaged DNA on skin cells remains unrepaired. This is most often caused by ultraviolet radiation (from sunshine or tanning beds) (Matsumura & Ananthaswamy, 2004; Ravanat et al., 2001; Kiefer, 2007). In a nutshell, this attained genetic defects trigger mutations that may lead the skin cells to multiply rapidly and form what is know as malignant tumours. In turn, these originate in the pigment-producing melanocytes in the basal layer of the epidermis – the ones that produce melanin and give skin its colour and tan. Melanomas often resemble moles, some develop from moles – known as nevus, a similar lesion but of benign origin. The majority of melanomas are black or brown, but they can also be skin-coloured, pink, red, purple, blue, or white. Apart from being mainly caused by intense, occasional UV exposure, one can also be genetically predisposed to the disease.

Melanoma incidence is increasing among the world population. If identified and treated at early stages it is almost always curable (Goldsmith et al., 1992). However, if not, the cancer can progress and spread to other parts of the body. After spreading, it becomes harder to target and treat, likely producing a fatal outcome (AJ et al., 1979; Wick et al., 1980). In resume, early detection of skin cancer constitutes a determinant task in clinical prognosis (Friedman et al., 1985).

Automated melanoma identification is useful to help dermatologists, and it is nowadays becoming possible with the development of new imaging techniques and advances in computing capabilities (both in processing and storage), which are supporting and pushing the development of new algorithms and the creation more image databases. Non-invasive *in vivo* imaging techniques have been applied to develop reliable systems to assist in the clinical diagnosis decision (Smith & MacNeil, 2011). Like in order scenarios, digital imaging proved to be

a great improvement for both health practitioners and patients. Reviewed in Stoecker & Moss (1992), the benefits of computer vision in dermatology soon became another success. Objective non-invasive documentation of skin lesions, digital dermatological image archives, quantitative description of clinical features of cutaneous lesions, among others, are some of the benefits that already became a reality as a consequence of digital imaging.

## 2.2 Datasets

The literature includes some publicly available datasets of collected skin lesion images and skin lesion type label annotations. The five datasets used in this thesis, detailed in the following paragraphs and on Table 2.1, are all of dermatoscopic or macro-imaging origin. Apart from these datasets (comprised by 2D/colour images), the work encompassed by this thesis also produced a novel 3D image dataset. This dataset is detailed in Chapter 4.1.

**Table 2.1:** Relevant 2D Skin Lesion Datasets.

| Dataset | Modality | Samples | Classes | Ref |
|---|---|---|---|---|
| Atlas | dermoscopic | 100 | 2 | Argenziano et al. (2000) |
| Dermofit | macro | 1300 | 10 | (Ballerini et al., 2013) |
| ISIC | dermoscopic/macro | 3438 | 3 | (Collaboration, 2017) |
| MED-NODE | macro | 170 | 2 | (Giotis et al., 2015) |
| $PH^2$ | dermoscopic | 200 | 2 | (Mendonça et al., 2013) |

**Atlas** The EDRA-Interactive Atlas of Dermoscopy (Atlas) dataset (Argenziano et al., 2000) is a representative set of images allowing comparisons among clinical, dermoscopic, and histopathological samples. Images have 700×447 8-bit RGB pixels, and were acquired with a Dermaphot/Optotechnik dermoscope. This dataset does not include segmentation masks. However, it is possible to find in the literature some studies that introduce a segmentation border to the images from this dataset in collaboration with dermatologists. This is the case of the work by Celebi et al. (2010), whose segmentation results can be used as ground-truth. It should be pointed out that with no formal ground-truth masks in the original dataset, there is no guaranty that the selected ones are optimal and valid, nonetheless, whenever this dataset is used in this thesis the Celebi et al. (2010) segmentation masks are considered as ground-truth.

**Dermofit** The Dermofit dataset (Ballerini et al., 2013) is a collection of 1300 high quality focus skin lesion images (with sizes ranging from 177×189 to 3055×1630 8-bit RGB pixels), all collected under similar conditions by the Department of Dermatology of the University of Edinburgh, using a Canon EOS 350D SLR camera (at an average distance of 50 cm) and a flash-ring for controlled lighting. This dataset includes a segmentation ground-truth, generated by a method that uses a statistical model for annotation.

**ISIC** The ISIC dataset (Collaboration, 2017) is a collection of images (of both dermoscopy and macro imagery) created by The International Skin Imaging Collaboration, which is still growing every year. When used for this work, the acquired dataset contained a total of 3438 images divided into 17 lesion types, including segmentation masks and a target classification label obtained from an unspecified number of skin cancer experts, as well as several other information such as: diagnosis type, clinical size, patient approximate age, general anatomic site, patient sex, and whether or not the patient previously had melanoma.

**MED-NODE** The MED-NODE dataset (Giotis et al., 2015) consists of 70 melanoma and 100 nevus images from the digital image archive of the Department of Dermatology of the University Medical Centre Groningen. The images were acquired with a Nikon D3 / Nikon D1x camera and a Nikkor 105 mm f/2.8 micro lens, at an approximate distance of 33 cm between the lens and the targeted lesion. Images of PSL originate only from patients of Caucasian origin (majority of the population in the Netherlands). Prior to the dataset release, rescale and resize (along with other operations like hair removal) were performed manually. These operations are lesion-region dependent since some manual upscaling was performed to set the images range from 349×321 to 1880×1867 pixels. Each image shows a single region-of-interest (ROI) that contains both healthy and lesion skin, and an associated lesion type classification label.

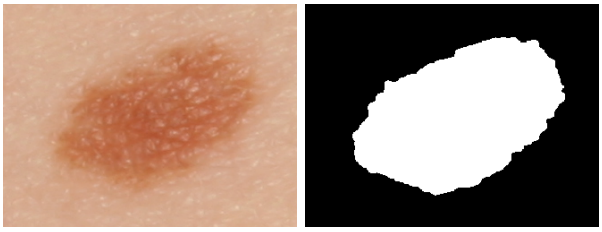
**PH$^2$** The Pedro Hispano Hospital (PH$^2$) dataset (Mendonça et al., 2013) is another set of dermoscopic images. It comprises 200 images (with sizes ranging from 761×570 to 769×577 8-bit RGB pixels) acquired with a Tuebinger Mole Analyzer system using a magnification of 20 times and the corresponding ground-truth segmentation masks (manually annotated by an expert dermatologist). The PH$^2$ has been used in different literature reviews for several dermoscopy purposes (as in Silveira et al., 2009).

## 2.3 Evaluation Metrics

In the following chapters of this thesis, evaluation metrics capable of measuring and enabling the performance assessment of algorithms and methods will be necessary to perform comparisons and extract relevant insights. Some of these metrics are only relevant for skin lesion segmentation, presented in the following Section 2.3.1, while others are only used to measure classification performance, presented in Section 2.3.2.

### 2.3.1 Segmentation Metrics

Typically, a segmentation is used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics. In the case of skin lesion segmentation, the objective is to perform the binary separation of the image pixels into either belonging to the healthy skin region or to the lesion area, as depicted in Fig. 2.2 – where black pixels mark healthy skin region and white pixels mark the (nevus) lesion region.



**(a)** Nevus image **(b)** Ground-Truth mask

**Figure 2.2:** Segmentation Example: (a) nevus image from the Dermofit dataset and (b) corresponding ground-truth image mask.

The process of comparing and determining the better segmentation is generally simpler when performed for binary components, i.e., assigning each pixel to one of two possible types of regions. It is possible to extract metrics such as how much area two segmentations have in common, verify how much area is incorrectly included in the ROI, how far from each other are the two pixel subsets, or establish the best intersection. The following paragraphs detail such metrics, which are used in the experiments and assessments performed in this thesis (in similarity to other literature works: Hance et al., 1996; Celebi et al., 2009b; Garnavi et al., 2011). In addition, for metrics based on set theory, Fig. 2.3 provides an overview of their application.

**Border Error (BE)** The BE is a broadly used metric to assess segmentation algorithms. It was first proposed by Hance et al. (1996) and it can be described by Eq. (2.1). It allows to measure the fraction of non-overlapping area between the proposed segmentation method being evaluated (PS) and the ground-truth segmentation (GT). This area (i.e. number of pixels) of the non-overlapping segmentation parts (exclusive-OR, $\oplus$) is calculated in percentage. In the equation, # represents the number of pixels that a region contains. When PS is an exact match with GT (best scenario), BE is equal to zero (0%) since there are no errors. In contrast, if there are errors, the upper-bound is constrained to the maximum number of pixels the masks have.

$$\mathrm{BE}(\mathrm{PS},\mathrm{GT}) = \frac{\#(\mathrm{PS} \oplus \mathrm{GT})}{\#(\mathrm{GT})} = \frac{\#(\mathrm{PS} \cup \mathrm{GT}) - \#(\mathrm{PS} \cap \mathrm{GT})}{\#(\mathrm{GT})} \tag{2.1}$$

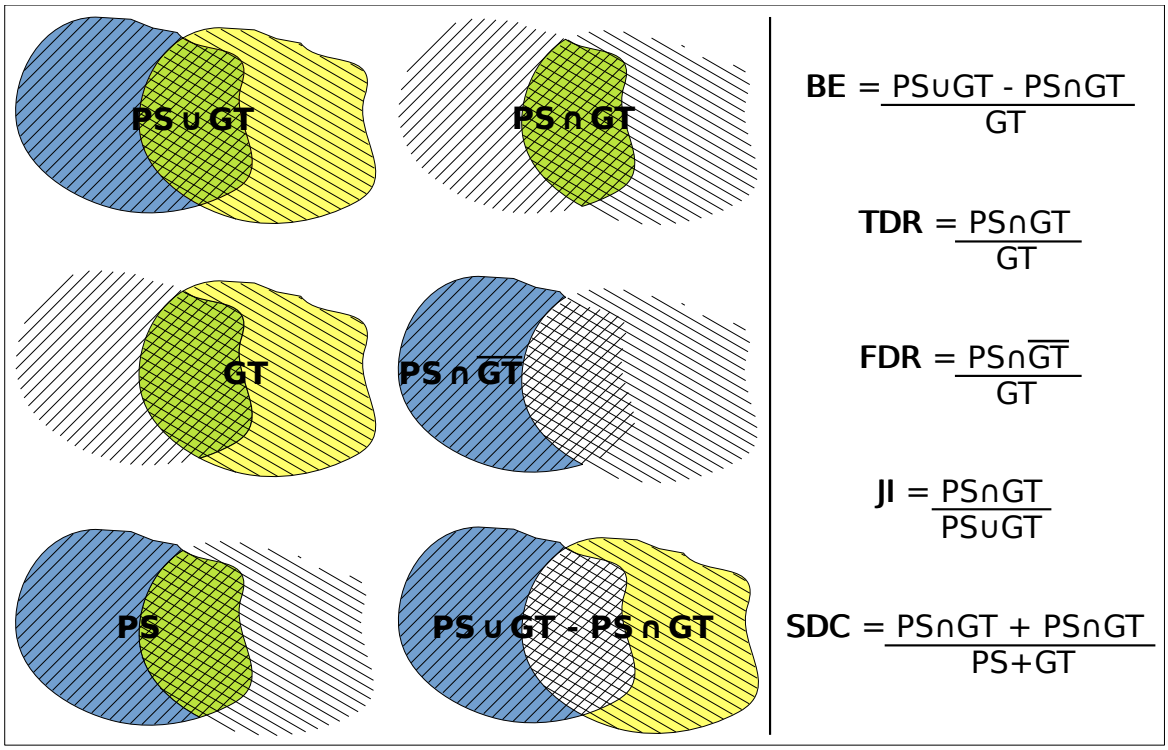



**Figure 2.3:** Segmentation Metrics Visualisation. Given a proposed segmentation (PS) and the ground-truth segmentation (GT), visualise metrics as Venn Diagram calculations.

**True Detection Rate (TDR), False Detection Rate (FDR)** The TDR, in Eq. (2.2), measures the ratio of GT pixels that are correctly classified as lesion in PS, and the FDR, in Eq. (2.3), measures the ratio of GT pixels that are incorrectly classified as lesion in PS. Both TDR and FDR are in the rage of $[0, 1]$. In the best scenario, TDR should be 100% and FDR equal to 0%.

$$\mathrm{TDR}(\mathrm{PS}, \mathrm{GT}) = \frac{\#(\mathrm{PS} \cap \mathrm{GT})}{\#(\mathrm{GT})} \tag{2.2}$$

$$\mathrm{FDR}(\mathrm{PS}, \mathrm{GT}) = \frac{\#(\mathrm{PS} \cap \overline{\mathrm{GT}})}{\#(\mathrm{GT})} \tag{2.3}$$

**Jaccard Index (JI)** The JI, defined in Eq. (2.4), measures the similarity between two sets. It is computed as the size of the intersection divided by the size of the union of the segmentation masks. The metric is also known by "intersection over union index", as it provides the ratio of intersection over union areas between two regions, enhancing the common region over the total ROI, which is seen as a central feature to distinguish segmentation outputs. Nonetheless, the previously mentioned metrics also provide extra relevant details, which is specially convenient when comparing similar JI. In the best scenario, JI should be equal to one, meaning an 100% match between masks. As the masks become more distant, JI will decrease and reach 0% when the two masks no longer overlap.

$$\mathrm{JI}(\mathrm{PS}, \mathrm{GT}) = \frac{\#(\mathrm{PS} \cap \mathrm{GT})}{\#(\mathrm{PS} \cup \mathrm{GT})} \tag{2.4}$$

**Sørensen-Dice coefficient (SDC)** The SDC metric, defined in Eq. (2.5), is another similarity metric, which is seen as equivalent to the JI metric. Given a value for the Jaccard Index JI, the SDC metric value for the Sørensen-Dice coefficient can be calculated (and vice versa), using equations $\mathrm{SDC} = 2\mathrm{JI}/(1 + \mathrm{JI})$ and $\mathrm{JI} = \mathrm{SDC}/(2 - \mathrm{SDC})$, respectively. Additionally, since the SDC does not satisfy the triangle inequality (caused by giving 2 times the importance to the overlapping area), it can be considered a semi-metric version of the JI.

$$\mathrm{SDC}(\mathrm{PS}, \mathrm{GT}) = \frac{\#(\mathrm{PS} \cap \mathrm{GT}) + \#(\mathrm{PS} \cap \mathrm{GT})}{\#(\mathrm{PS}) + \#(\mathrm{GT})} \tag{2.5}$$

**Hausdorff Distance (HD)** The HD metric measures how distant are two non-empty subsets (of a metric space), as shown in Eq. (2.6), where $d$ measures the metric space from a point $p$ to the closest point of a non-empty set (either PS or GT). By definition, two sets are close in terms of the HD if every point of either set is close to some point of the other set. This means that in the best scenario, HD is equal to zero, while the maximum distance upper-bound can be as large as the image diagonal distance (if the two masks comprise one pixel each and are in opposing sides of the diagonal).

$$\mathrm{HD}(\mathrm{PS}, \mathrm{GT}) = \max\left\{\max_{p \in \mathrm{PS}} d(p, \mathrm{GT}), \max_{p \in \mathrm{GT}} d(p, \mathrm{PS})\right\} \tag{2.6}$$

### 2.3.2 Classification Strategies and Metrics

The main objective of the classification process is to assign labels to each element of a given set (images, in the context of this work). More specifically, throughout this thesis, the goal of the classification is to assign a specific skin lesion class label to each image in a given dataset. In order to fulfil this purpose, several classification models based on different feature subsets, samples, and classifiers are evaluated using test sets. Therefore, the predicted label of new classified samples is compared to the known label, in order to evaluate the classification performance. Among several evaluation procedures, cross-validation (CV) (Witten et al., 2005) is the most commonly used in the literature to assess the results of skin lesion classification, since it avoids over-fitting while testing the capacity of the classifier to generalize. The $k$-fold CV (Maglogiannis & Doukas, 2009) and leave-one-out (Barata et al., 2014; Iyatomi et al., 2010) methods are examples of CV proposed for classifying skin lesions in images. The half-and-half test is another evaluation procedure which was applied by Iyatomi et al. (2008).

Statistical measures based on performance metrics are computed to compare the achievements of different classification models according to the outcomes of the classifiers. Possible outcomes of classifiers based on the predicted class and expected class are: true positives, true negatives, false positives, and false negatives; which are typically expressed as a confusion matrix, as shown in Table 2.2. These represent the number of correct (true) and incorrect (false) classifications for each class (positive and negative). For instance, in a classification process

**Table 2.2:** Confusion Matrix.

| | | Predicted class | |
|---|---|---|---|
| | | Positive | Negative |
| Actual class | Positive | True Positive | False Negative |
| | Negative | False Positive | True Negative |

involving two classes, one class may be considered positive and another negative. Positive samples usually represent the most important class (e.g., skin cancer, melanoma), and negative samples the less important (e.g., benign lesions, nevus). Therefore, the true positive rate (TPR), is the number of correctly classified positive samples, as expressed in Eq. (2.7); the true negative rate (TNR), is the number of correctly classified negative samples, as expressed in Eq. (2.8); the false positive rate (FPR), is the number of negative samples incorrectly classified as positive samples, as expressed in Eq. (2.9); and the false negative rate (FNR), is the number of positive samples incorrectly classified as negative samples, as expressed in Eq. (2.10). These rates can be extracted from a confusion matrix (Table 2.2), which is the basis for several metrics used by researchers to measure the classification performance (Alcón et al., 2009; Situ et al., 2008; Smith et al., 2011; Satheesha et al., 2017; Pathan et al., 2018; Hu et al., 2019).

$$\text{TPR} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2.7}$$

$$\text{TNR} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \tag{2.8}$$

$$\text{FPR} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} \tag{2.9}$$

$$\text{FNR} = \frac{\text{False Negative}}{\text{False Negative} + \text{True Positive}} \tag{2.10}$$

For this thesis, as for most of the reference literature, the most relevant metrics are: the accuracy (ACC), that is the percentage of correctly classified positive and negative samples based on all samples, as expressed in Eq. (2.11); the sensitivity (SEN, also known as recall or TPR), that is the percentage of correctly classified positive samples with respect to all positive samples (typically representing the successful melanoma, or malignant lesion, identification rate); and the specificity (SPE, also known as TNR), that is the percentage of correctly classified negative samples with respect to all negative samples. More details about this metrics can be found in Baratloo et al. (2015).

$$\text{ACC} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \tag{2.11}$$

In addition to these evaluation metrics, since skin lesion classification is often an unbalanced classification problem, the balanced-accuracy (BAC, as introduced in Hu et al., 2019) is also used in this work when relevant for similar performance distinction. The BAC metric

corresponds to the average value between sensitivity and specificity, as defined in Eq. (2.12).

$$\text{BAC} = \frac{\text{SEN} + \text{SPE}}{2} = \frac{\text{TPR} + \text{TNR}}{2} \tag{2.12}$$

## 2.4 2D Skin Lesion Analysis Pipeline

According to the literature, established methods commonly comprise five steps for skin lesion detection (as depicted in Fig. 2.4). It starts with Image Acquisition (Section 2.4.1), followed by Pre-Processing (Section 2.4.2) of the data, which tries to enhance several aspects of the image. Then, the aim is to find the image region where the lesion is located, a process known as Segmentation (Section 2.4.3), to the extent of allowing Feature Extraction mechanisms (Section 2.4.4) to be applied over the relevant image area. Finally, some sort of Classification (Section 2.4.5) is applied over the attained features, in order to infer if the lesion area possesses melanoma characteristics.


**Figure 2.4:** Five common steps for skin lesion detection.

Concerning early detection of melanoma, new solutions of CAD based on ML techniques have been investigated for feature extraction and pattern classification Lalitha & Geetha (2014); Hosny et al. (2019). Although it may seem a simple process, no unified procedure exists for any of the mentioned steps (Korotkov & Garcia, 2012). This is so mainly because the existing results are not all coherent with each other. Therefore, there is still a large space for improvement. Morphological differences in skin lesion images directly influence the choice of the method for, e.g., border detection.

### 2.4.1 Image Acquisition

Only recently, dermatologists have started to incorporate the novel imaging techniques in the process of patient diagnoses. The image acquisition technique greatly dictates the type of acquired information (and content) that can be used afterwards for skin lesion assessment. In this thesis, only dermoscopy (or macro skin images) and imaging techniques based on 3D lesion analysis are considered (namely light-fields). Dermoscopy, in particular, is a fast growing method to capture skin imagery and has been growing in both dissemination and utilisation (Psaty & Halpern, 2009); and, as a consequence, dermoscopy is now a common step in the initial examination for many patients.

### Dermoscopy

The field of dermoscopy traces back to 1620 with Pierre Borel (Domínguez-Espinosa, 2014; Kreusch & Rassner, 1990). However, it was only formally developed in the 20th century with Saphier (1921), and, only in 1971, it was first used to explore pigmented lesions (MacKIE, 1971). It quickly turned into a vast research topic in the following years, primarily in Europe and Australia, and only later in the United States (Nehal et al., 2002; Tripp et al., 2002; Hosny et al., 2019; Collaboration, 2017).

As a stepping stone, it has been reported that dermoscopy used in conjunction with clinical examination and patient history results in a 50% improvement in diagnostic accuracy (Kittler et al., 2002) – although the diagnostic classification of small or featureless lesions remains a problem (Blum et al., 2004; Pizzichetta et al., 2001; Menzies et al., 1996; Skvara et al., 2005). It has also been noted that the isolated use of dermoscopy is not a good practice, since some lesions are understood as being suspicious only when viewed in the context of the surrounding skin and with information about the patient's history (Carli et al., 2005). Even so, new criteria are always emerging to help label lesions as melanocytic or non-melanocytic (Argenziano et al., 2003, 2007; Korotkov & Garcia, 2012; Lalitha & Geetha, 2014; Pathan et al., 2018; Oliveira et al., 2018). An example of melanocytic lesions are nevi, which, in general, tend to be symmetric, uniform, display less than three colours, and possess an orderly, perhaps even aesthetically pleasing, architecture. Melanomas however, a visually similar melanocytic lesion type, can be marked as architectural disorder. As stated in Section 2.1, most melanomas posses at least one of nine dermoscopic structures, even on different lighting-environments (polarized and non-polarized) certain structures are better seen under different dermoscopy systems.

Still, given its usage and potential, dermoscopy-related algorithms are constantly emerging, but with variable definitions of specific attributes – complicating diagnosis when using multiple systems. This is why the International Dermoscopy Society created a consensus document for the standardisation of how to effectively convey dermoscopic findings (Malvehy et al., 2007). This was later used to start a study aimed at creating a unified dermoscopy algorithm (UDA) by using the most popular and widely accepted algorithms developed by dermatologists. This project was a multidisciplinary World Wide Web-based study that tracked the ratings and scores of hundreds of histologically confirmed melanocytic lesions, nevi and melanomas, which were evaluated according to some defined metrics (Carli et al., 2005; Henning et al., 2007; Zalaudek et al., 2006).

### 3D Lesion Analysis

Original works that aim for the 3D lesion reconstruction began with the use of the device named "Nevoscope" in 1984. It worked by obtaining images of transilluminated lesions at different angles and then applying a limited-view computed tomography reconstruction algorithm

(Dhawan et al., 1984; Rigel et al., 2010; Kini & Dhawan, 1992). Later, in a similar way as in 2D, features extracted from 3D lesions (generated through stereoscopy) were presented in McDonagh et al. (2008) as input to a Bayesian classifier to distinguish melanoma from non-melanoma. Then, an evolved method was presented in Smith et al. (2011) based on skin tilt, slant patterns, and statistical moments of enhanced principal curvatures of skin surfaces (Zhou et al., 2010, 2011), by using the photometric stereo technique. As in previous works, 3D features were extracted and used as input to both an artificial neural network and a C4.5 tree classifier. The isolated use of the 3D features did not outperform 2D features, however the combined used of both sets of features provided superior performance results.

**Light-Fields** A fairly unexplored image modality in PSL is the recording of skin surface visual information using light-field cameras, which are able to capture not only light intensity but also its direction, providing richer visual information for research on new computational methods (Donatsch et al., 2014; Dansereau et al., 2015). Light-field data allow to obtain multiple views on the lesion and, from there, create a depth map. For instance, light-field depth maps with sub-millimetre accuracy have been used in high precision robotic surgery in Shademan et al. (2016). The use of light-fields in medicine for diagnosis, surgery planning and execution, as well as training of health professionals has been attracting increasing research attention (Makanjuola et al., 2013; Saha et al., 2014; Marshall et al., 2014). Analysis of PSL can also benefit from the additional light-field visual dimensions, as reported in Baghdadchi et al. (2014), where a commercial light-field camera of the first generation has been used to visualise a skin condition, with some constraints in terms of resolution. For a more thorough understanding of the concept that is light-field and its origin, refer to Lippmann (1908b); Adelson & Wang (1992); Ng et al. (2005) and all other works/lectures from Lippmann.

### 2.4.2 Pre-Processing

As in other applications, after a clinical or dermoscopic image is acquired, it may not have the adequate quality for subsequent analysis. Typically, PSL images have extraneous artefacts – hairs, air bubbles, ink marks – around or inside the lesion area. Hence, the first step in image processing pipelines is the correction of these artefacts, or the whole image itself (e.g., colour range or distribution correction), so that later border detection (segmentation) is more accurate. Therefore, the good performance of the methods not only contributes to the correct behaviour of the algorithms in the following stages of the analysis, but also loosens the constraints on the image acquisition process. Some common pre-processing methods are: mean filter, wavelet denoising, median filter, Dull Razor method, Wiener filter, and morphological closing (Oliveira et al., 2016) – apart from color adjustments and transformations.

Concerning colour adjustments, many colour spaces have been explored in order to extract more specific information about the lesion colours. One such colour space, relevant for this thesis is the CIE L*a*b* colour space (Barata et al., 2013, 2015; Rastgoo et al., 2015). Most

colour space transformations are performed to adjust colour distributions while others are to enhance certain aspects of the image attributes. For example, conversion from RGB to the CIE L*a*b* colour space is known to directly separate reds from greens and yellows from blues – which can greatly help in some cases, such as in clustering and colour segmentation.

### 2.4.3 Segmentation

In this thesis, the process of discriminating between the lesion area and the healthy skin is called segmentation. This distinction is important so that the forthcoming steps only utilizes the ROI – region of interest, in this case the lesion region – and ignores the surrounding normal skin. Nevertheless, lesion border detection is not a trivial task. Dermatologists, do not usually delineate detailed lesion borders for diagnosis (but rather perform a rough identification of the area for incision, Day & Barbour, 2001), therefore, skin lesion datasets often do not present an accurate segmentation ground-truth. Additionally, humans are generally not very good at discriminating subtle variations in contrast or blur (Claridge & Orun, 2002), thus in perceiving the boundaries of a lesion. This poses some difficulties when trying to find the correct border line, mainly because of high inter- and intra-observer variability in PSL boundary perceptions among dermatologists (Claridge & Orun, 2002; Joel et al., 2002; Iyatomi et al., 2006). Apart from the human factor, the morphological structure of a lesion itself (low lesion-to-skin gradient, multiple lesion regions, among others) can induce more complexity for both manual and automatic segmentation. Up to date, a large number of image segmentation techniques (which span almost all categories of segmentation algorithms, Vestergaard & Menzies, 2008) has been presented by researchers, yet, none of them is suitable for all sorts of applications.

Skin lesion morphological differences in clinical and dermoscopic images directly influence the choice of method for border detection. This is due to the different techniques available for image acquisition alongside with the chaotic mixture of environment-variables (e.g., type of lesion, location, colour conditions, angle of view, and skin types) that increase difficulties in segmentation when using the same imaging modality (Celebi et al., 2009a; Fleming et al., 1998; Xu et al., 1999). Hence the existence of several databases, each with its approach and points of view, where available methods aim to provide robustness for difficult segmentation cases, by adapting it to specific conditions of the image type (e.g., Zhou et al., 2008).

A relevant selection of segmentation methods, pertinent for Chapter 3, is discussed is Section 2.5.

### 2.4.4 Feature Extraction

After proper image adjustments and ROI determination, feature extraction takes place – either with manual definition of such features or with their acquisition via machine learning algorithms. This is a very important task, where the most prominent attributes or aspects

of the image regions are evaluated. This process intends to select an optimal subset of lesion characteristics so that they can be given as input for a classification process.

The first standard set of dermoscopy (manually defined) features was introduced in early 1994 (Nachbar et al.) as the "ABCD" rule mnemonic to distinguish between benign lesions and melanoma. This was later compared with the "7-Point Checklist" analysis (Clemente et al., 1991, which is based on simplified epiluminescence microscopy instead of dermoscopy) in Argenziano et al. (1998). Afterwards, it was expanded to "ABCDE" by including the "E" to denote the lesion evolution over time (Abbasi et al., 2004). Another increment immediately took place the year after with the addend of "FG" for the diagnosis of melanoma (Fox, 2005). In Seidenari et al. (2006), the asymmetry importance for melanoma identification was devised. In Mendes et al. (2016), new methods were introduced to align and asses lesion growth over time. The current "ABCDEFG" rule stands as follows:

**A**symmetry : one half of the tumour does not match the other half
**B**order : edges are ragged, notched, blurred
**C**olour : pigmentation is not uniform
**D**iameter : greater than 6 mm and growing
**E**volving : evolving lesion over time
**F**irm : lesion is firm to touch
**G**rowing : growing rapidly in a few months or weeks

In 1998, the concept of "ugly duckling" appeared (Grob & Bonerandi, 1998) – it stated that different melanocytic nevus/moles in the same person would resemble each-other, but a melanoma lesion would be different. More recently (Sadeghi et al., 2013), an approach has proposed which detects and analyses irregular streaks in dermoscopic images.

Feature extraction may be based on lines, edges, textures, or points. However, in recent years, texture, border and geometric based approaches became more relevant, as expressed in Barata et al. (2018); Oliveira et al. (2018); Baig et al. (2020) – including abstractions acquired with Deep Learning (i.e., machine learned features).

### 2.4.5 Classification

As a final step in the skin lesion pipeline, classification typically assumes the role of the dermatologist or assists on its judgement (CAD systems). Nowadays, there is a diverse collection of approaches, some are built to only distinguish melanoma from non-melanoma images, while others attempt to also recognise other types of skin pathologies.

Assuming the existence of descriptors extracted in a previous step (Section 2.4.4), one or more classification methods can be applied. Their performance depends on both the quality and quantity of the data, the extracted descriptors, and on the chosen classifier. The comparison of classification approaches should be performed on the same dataset and using the same set of

descriptors, to provide adequate results. A research work presented in Maglogiannis & Doukas (2009) performed a unified comparison of 11 of the most common classifier groups used in PSL using over 3639 dermoscopic images. The work comprised three sub-experiments, each aiming different classification outputs (either melanoma vs nevus, or dysplastic vs nevus, or all three). In this study, Support Vector Machines (SVMs) showed the best overall performance, but the author concluded that it was biased due to the set of extracted features.

Machine Learning, in particular for image recognition or classification, has become a major topic in a wide range of research fields because of its ability to learn abstract data models and intrinsic discriminative properties. The datasets used for training, validation and testing are crucial elements required for research on image classification with ML. Amongst the most important, there is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), currently considered one of the standard references used in recent years as a benchmark standard for large-scale object recognition, i.e. image classification, single-object location and object detection. ImageNet has been used by many authors to improve their image classification/recognition algorithms. Its use promoted an exponential growth of research results and significant improvements to the state-of-the-art techniques (Russakovsky et al., 2015).

A recent work in Esteva et al. (2017) uses DL to catalogue a wide range of skin diseases, including melanoma. The authors used the Google Inception v3 (a convolution neural network, CNN) model (Szegedy et al., 2015, 2016) as a starting point for their PSL detection system. This technique, known as TL or Domain Adaptation (Pan & Yang, 2010), assumes that neuron weights only change slightly from task to task and that a static architecture may be maintained. This means, in this case, that if we start with an already good image recognition network, as Google Inception v3 (GoogLeNet) that only had an error rate of 6.7% in the ILSVRC of 2014 (Szegedy et al., 2015), odds are that it will quickly adapt itself to recognise a new set of categories, like the different skin lesions. For example, based on the AlexNet pioneer CNN architecture (Krizhevsky et al., 2012) several works, mainly triggered by the International Skin Imaging Collaboration challenge of 2017, preferred TL instead of starting from scratch (Sousa & de Moraes, 2017; Berseth, 2017; Harangi, 2017; Yu et al., 2017; Murphree & Ngufor, 2017; Menegola et al., 2017; Shoieb et al., 2016; Liao et al., 2016).

In Hosny et al. (2019), a classification approach for segmented colour skin lesion images of three datasets is performed using TL on the AlexNet CNN model (pre-trained in the ILSVRC). In order to increase the number of dataset samples and lower the model overfit probability, augmentation based on image rotation is performed. Data normalisation is also employed, as originally applied for the previously trained ImageNet data (maintaining the same colour feature space). As commonly used in TL methods, the model classification output function is replaced by an appropriate softmax layer for either melanoma and nevus (binary) or melanoma, seborrheic keratosis, and nevus (ternary) discrimination. After fine-tuning the model weights on each dataset, and performing augmentation in both train and test sets, the reported system ACC performance was measured as 96.86%, 97.70%, and 95.91%,

for three different datasets, respectively. Without augmentation, the recorded performance was 88.24%, 91.18%, and 87.31%, for the same datasets. Additional information about Deep Neural Networks (DNN) in skin lesion applications can be found in Senan & Jadhav (2019).

## 2.5 2D Skin Lesion Segmentation Methods

Recent advances in machine learning approaches are rapidly changing the landscape of medical image processing algorithms for detection, recognition and classification. In order to work properly, these methods require datasets with accurate ground-truth image segmentation, both for training and validation of such new computational models (Ker et al., 2018; Oliveira et al., 2018). In the case of skin lesion segmentation, the difficulty of achieving accurate delineation of ROI borders manually has driven research efforts to increase the availability of ground-truth ROI through computational methods (Cheng et al., 2015; Kéchichian et al., 2014).

The creation of accurate segmentation masks is an important step in the classification pipeline, since it enables classification or feature extraction algorithms to filter out features extracted from surrounding healthy skin. As a consequence of not including extraneous information for the studied problem, ML techniques achieve higher success rates (Lee et al., 2018).

This section presents existing segmentation methods that have been applied to skin lesions or that were specially made for that purpose. Skin lesion segmentation methods can be grouped, for example, based on their underlying algorithmic approach. The remainder of this section refers to two groups of methods: those implemented by the author of the this thesis (following other works' descriptions) for application and validation against specific datasets, and those not implemented by the author (since they already have metric information publicly available for the relevant comparisons made in Section 3.2). The first group comprises 27 segmentation methods that are divided into the following seven categories: Threshold (Section 2.5.1), Clustering (Section 2.5.2), Fuzzy Methods (Section 2.5.3), Quantization (Section 2.5.4), Active Contours (Section 2.5.5), and Merging Threshold (Section 2.5.6). This categorisation is shown in Table 2.3, where the first column (ID) is the acronym used to identify each method and the last column corresponds to the associated bibliographic reference. Most of these references indicate the publications where the original methods were proposed. The second group of algorithms, recently proposed in the literature and based on diverse (not-implemented) approaches, are listed in Table 2.4 and discussed in Section 2.5.7.

### 2.5.1 Thresholding Methods

Segmentation algorithms based on thresholding approaches can be generally divided into Global and Local Thresholding, according to whether the entire input image or smaller par-

**Table 2.3:** Conventional segmentation algorithms (Implemented).

| ID | Name | Ref |
|---|---|---|
| | **Thresholding** | |
| UT | Huang Threshold | Huang & Wang (1995) |
| IT | Iterative Threshold | Trussell (1979) |
| KT | Kapur Threshold | Kapur et al. (1985) |
| LT | Li Threshold | Li & Lee (1993) |
| MT | Moment Threshold | Tsai (1985) |
| OT | Otsu Threshold | Otsu (1979) |
| ST | Shanbhag Threshold | Shanbhag (1994) |
| YT | Yen Threshold | Yen et al. (1995) |
| BT | Bradley Threshold | Bradley & Roth (2007) |
| RT | Renyi Threshold | Sahoo et al. (1997) |
| | **Clustering** | |
| KMC | K-Means Colour | MacQueen et al. (1967) |
| KMS | K-Means Colour and Spatial | Ilea & Whelan (2006b) |
| MC | Mean Shift Colour | Fukunaga & Hostetler (1975) |
| MCS | Mean Shift Colour and Spatial | Comaniciu & Meer (2002) |

| ID | Name | Ref |
|---|---|---|
| | **Fuzzy Methods** | |
| FDEE | Fuzzy Differential Evolution Entropy | Sarkar et al. (2014) |
| FC-LS | Fuzzy Clustering LevelSet | Li et al. (2011) |
| FCM | Fuzzy C-Means | Bezdek et al. (1984) |
| | **Quantization** | |
| NQ | Neural Quantization | Dekker (1994) |
| AQ | Wan Quantifier | Wan et al. (1990) |
| UQ | Wu Quantifier | Wu (1991) |
| RGB-MC | RGB Median Cut | Heckbert (1982) |
| PCT-MC | PCT Median Cut | Umbaugh et al. (1993) |
| | **Active Contours** | |
| CH | Chan-Vese | Chan et al. (2001) |
| VV | Chan-Vese Vector | Chan et al. (2000) |
| LMS | Lankton Mean Separation | Lankton & Tannenbaum (2008) |
| | **Merging Threshold** | |
| SRM | Statistical Region Merging | Celebi et al. (2008) |

**Table 2.4:** Machine learning-based algorithms (Not Implemented).

| ID | Name | Ref |
|---|---|---|
| FCN | Fully Convolutional Networks | Long et al. (2015) |
| SSLS | Saliency-based Skin Lesion Segmentation | Ahn et al. (2015) |
| SCDRR | Sparse Coding with Dynamic Rule-based Refinement | Bozorgtabar et al. (2016) |
| MSCA | Multi-scale Superpixel with Cellular Automata | Bi et al. (2016) |
| mFCN | Multi-stage Fully Convolutional Networks | Bi et al. (2017) |
| JCLMM | Joint Circular-Linear Mixture Model | Roy et al. (2017) |
| CDNN | Convolutional-Deconvolutional Neural Network | Yuan et al. (2017) |
| CDNNE | Enhanced CDNN | Yuan & Lo (2019) |
| KL-LS | Kullback-Leibler based Level Sets | Riaz et al. (2019) |
| DermoNet | Densely Linked Convolution Neural Network | Baghersalimi et al. (2019) |
| PSO-DEN | Particle Swarm Optimization on Deep Ensemble Network | Tan et al. (2019) |
| SWSDB | Semi- and Weakly Supervised Directional Bootstrapping | Xie et al. (2019) |
| DCL-PSI | Deep Class-specific Learning with Prob. based Step-wise Int. | Bi et al. (2019) |

titions are used to optimise the process. The presented thresholding methods perform image segmentation by generating a modified version of the image, whose grey-level values are separated using a threshold. The algorithms based on thresholding herein considered are the following:

- *Huang Threshold* (UT) (Huang & Wang, 1995) is a method based on the minimisation of the fuzziness of an image using a membership function to obtain the threshold. This membership function denotes the relationship between a pixel and the region where it belongs (either ROI or background);
- *Iterative Threshold* (IT) (Trussell, 1979) chooses an optimum threshold after successive iterations, providing increasingly cleaner extractions of the ROI. It does so by transforming a smooth grey-level picture into a bi-modal picture (in the greyscale colour space) while maintaining, as close as possible, the average luminance of the picture;
- *Kapur Threshold* (KT) (Kapur et al., 1985) determines the threshold by maximising the entropy on the grey-level histogram;
- *Li Threshold* (LT) (Li & Lee, 1993) addresses the threshold selection problem by minimising the cross entropy between the image and its segmented version, with no assumptions about the distribution;

- *Moment Threshold* (MT) (Tsai, 1985) uses an image moment-preserving principle where the threshold values are deterministically computed in such a way that the grey-level moments of an input picture are preserved in the output;
- *Otsu Threshold* (OT) (Otsu, 1979) is a non-parametric and unsupervised image segmentation method for automatic threshold selection that uses a discriminant criterion. It only utilises the zeroth- and first-order cumulative moments of the grey-level histogram to calculate the optimum threshold, which separates two classes in such a way that their combined intra-class variance is minimal.
- *Shanbhag Threshold* (ST) (Shanbhag, 1994) is a modified version of the Kapur Threshold (Kapur et al., 1985) method, where a more pertinent information measure of the image is obtained, consisting of viewing it as being composed by two fuzzy sets corresponding to two different classes, with membership coefficients associated with their grey-level;
- *Yen Threshold* (YT) (Yen et al., 1995) is a method based on the use of both the discrepancy between the thresholded and the original image, as well as the number of bits required to represent the thresholded image. A maximum correlation criterion for bi-level thresholding is considered by minimising a cost function;
- *Bradley Threshold* (BT) (Bradley & Roth, 2007) is a local threshold method that sets each pixel to black if its brightness is a given percentage lower than the average of the surrounding pixels within a specified window size, otherwise each pixel is set to white;
- *Renyi Threshold* (RT) (Sahoo et al., 1997) is a technique based on Renyi's entropy. Similar to the maximum entropy sum method of Kapur et al. (1985) and the entropic correlation method of Yen et al. (1995), it proposes a thresholding technique using two probability distributions (object and background) derived from the original grey-level distribution of an image, and includes the maximum entropy sum method and the entropic correlation method.

UT, IT, OT, YT, MT, and ST methods were introduced as general-purpose (bi- or multi-level) segmentation algorithms for greyscale images and visually evaluated or compared with several classical images. KT, LT, and RT were built upon two other methods, taking advantage of the entropy concept in different ways. Finally, BT was proposed as a real-time solution for live stream videos and augmented reality solution, by marking pixels as dark or light based on the spatial variation in illumination.

### 2.5.2 Clustering Methods

Another type of segmentation algorithms is based on clustering, where pixels are grouped according to a given metric of similarity. In other words, pixels in the same group have similarities between each other (in a given sense). This type of algorithms is very common in the literature for various applications. A brief description of each one is presented:

- *K-Means Colour* (KMC) (MacQueen et al., 1967) is a method of vector quantization that

aims to partition observations into clusters, in which every single observation belongs to the cluster whose mean value is closer to said observation;

- *K-Means Colour and Spatial* (KMS) (Ilea & Whelan, 2006b) is an adaptive technique for colour-texture segmentation that is a generalisation of the standard KMC. It adds two more dimensions to the problem that sample the local colour smoothness and the local texture complexity. In addition, it also selects the dominant colours from the input image using information from colour histograms, so that proper cluster centres may be selected;

- *Mean-shift Colour* (MC) (Fukunaga & Hostetler, 1975) is a non-parametric density gradient estimation method of generalised kernel, in which a mean-shift estimate kernel is presented for gradient estimation;

- *Mean-shift Colour and Spatial* (MCS) (Comaniciu & Meer, 2002) is a general non-parametric technique for the analysis of a complex multi-modal feature space and to delineate arbitrarily shaped clusters in it, where the basic computational module is the MC. This method provides discontinuous clusters and preserves smoothing and image segmentation by augmenting the feature space with additional (spatial) parameters from the input domain.

As previously stated, both KMS and MCS are extensions of the more traditional KMC and MC algorithms, respectively, which were designed to take advantage of information connecting similar data-points, i.e., spatial information.

The KMC method was proposed with the aim of enabling a process for partitioning N-dimensional populations into $k$ sets on the basis of a sample. The KMS method was designed for image segmentation, aiming to complement KMC with spatial information during the space partitioning process. The starting point for the KMC approach was related to problems in optimal classification where results are theoretically justified, while in KMS results are visually displayed for 6 images and visually compared with Mean-Shift algorithm. In particular, in Ilea & Whelan (2006a), the KMS algorithm is employed by its original authors to perform segmentation on a set of 6 skin cancer images. The KMS results are evaluated using mean, standard deviation and root-mean-square of the Euclidean distance between the pixels of the ground-truth image and the proposed segmentation results.

While both MC and MCS algorithms were proposed to provide further understanding of dense information, they aim for different applications. MC was proposed as a non-parametric density gradient estimation method of generalised kernel that is applied to several pattern recognition problems (Gradient Clustering, Clustering and Data Filtering). Each kernel is derived, guaranteeing it is asymptotically unbiased, consistent, and its estimate is uniformly consistent (evaluation is made visually). On the other hand, MCS was proposed for analysis of a complex multi-modal feature space and to delineate arbitrarily shaped clusters in it, therefore the method has been successfully applied in several tasks and problems, like filtering and segmentation.

### 2.5.3 Fuzzy Methods

This type of segmentation methods uses a mixture of thresholding and clustering concepts, aided by fuzzy logic or representation. The following algorithms are considered:

- *Fuzzy Differential Evolution Entropy* (FDEE) (Sarkar et al., 2014) creates fuzzy partitions from the image histogram based on the entropy theory. Then, the entropy measure is optimised to obtain the thresholds of the image using a differential evolution metaheuristic, which leads to a fast and accurate convergence;
- *Fuzzy Clustering LevelSet* (FC-LS) (Li et al., 2011) is another example of a clustering algorithm based on an hybrid level model, alternating between global and local region competitions. The algorithm directly evolves from the initial segmentation by using spatial information in the fuzzy clustering technique, where the controlling parameters are automated through estimation from the results of fuzzy clustering;
- *Fuzzy C-Means* (FCM) (Bezdek et al., 1984) is a clustering variant, where fuzzy partitions and prototypes are firstly generated for any set of numerical data. Then a generalised least-squares objective function is used as the clustering criterion to aggregate the subset.

Both FDEE and FC-LS were devised targeting the same application, but FDEE was originally evaluated in Sarkar et al. (2014) by resorting to eight classical greyscale images and measuring performance in terms of computational time, mean objective value, and standard deviation of objective values. The FC-LS was proposed for medical image segmentation and it was visually evaluated for different modalities (including carotid artery ultrasound images, liver tumours CT scans, and cerebral tissue MRI slices). Finally, the FCM algorithm was originally implemented in Fortran-IV (Bezdek et al., 1984) that is linked to an original publication of Gustafson & Kessel (1979) through Bezdek (1981), which validates the method capability using two classes that have some degree of overlap.

### 2.5.4 Quantization

Segmentation approaches based on Quantization reduce the number of distinct colour levels. For skin lesion segmentation, quantization is performed until the image is segmented in two levels, i.e., the ROI and the background. In the scope of this work, five quantization methods were considered:

- *Neural Quantization* (NQ) (Dekker, 1994) is an algorithm that uses a self-organising Kohonen neural network to quantize the colour image;
- *Wan Quantifier* (AQ) (Wan et al., 1990) is an variance-based algorithm used for multidimensional data clustering, that uses a sum-of-squared error minimisation criterion between the original image and its quantized version;
- *Wu Quantifier* (UQ) (Wu, 1991) is based on variance minimisation through linear search.

The RGB colour space cube is divided in two, in each of its axes and the division plane that minimises the sum of variances at both sides of the colour space is selected, thus creating two boxes. Then the process is repeated for the box with the largest variance. The process stops when a predefined number of boxes is found, and the boxes' centre of gravity are selected as the representative colours.

- *RGB Median Cut* (RGB-MC) (Heckbert, 1982) maps colours to their nearest neighbours in a colourmap, effectively quantizing and redrawing the original image;
- *PCT Median Cut* (PCT-MC) (Umbaugh et al., 1993) algorithm is based on an optimal transform using the Principal Components Transform (PCT), also known as Karhunen-Loeve or Hotelling transform.

All these methods were developed aiming to advance the capabilities of representing images with less levels of intensity. NQ, AQ, and UQ were introduced as improvements over other state-of-the-art algorithms. NQ is a 24bit-to-8bit image converter for both greyscale and coloured images, which uses half the memory of other algorithms. UQ was proposed in the same scope as AQ, being able to achieve 1/3 of the mean-square error attained by other algorithms. Finally, RGB-MC and PCT-MC are intended to display high-quality reproductions of colour images with small frame buffers. The PCT-MC method is an alternative to RGB Median Cut for two-colour-image segmentation on skin tumour images for extraction of features like tumour border, crust, hair, scale shiny areas, and ulcer, with the underlying objective of developing an automated visual feature identification program.

### 2.5.5 Active Contour

A fifth type of segmentation approach is called Active Contour algorithms, whose main advantage is the ability to delineate objects in potentially noisy 2D images. In such methods the contour is defined by a constrained spline, which is obtained by minimising a cost function. In the context of this work the following methods are used:

- *Chan-Vese* (CH) (Chan et al., 2001) is a model based on active contours proposed to detect objects resorting to techniques of curve evolution, Mumford–Shah segmentation, and level sets. It minimises an energy which can be seen as a particular case of a minimal partition problem. The level set formulation iteratively improves the active contour until the desired boundary, which does not depend on the gradient since it is related to a particular segment of the image;
- *Chan-Vese Vector* (VV) (Chan et al., 2000) is an extension of the previous method that minimises a Mumford-Shah function over the length of the contour, plus the sum of the fitting error over each component of the vector-valued image. Like the original *Chan-Vese* model, the vector-valued model also detects edges both with or without gradient;
- *Lankton Mean Separation* (LMS) (Lankton & Tannenbaum, 2008) considers local image

statistics and develops a contour based on local information. It essentially derives the curve evolution that separates two or more values of a pre-determined set of statistics computed over geometrically determined subsets of the image.

Both CH and VV are very similar as they have the same base method, however, the later is specially tailored for object detection in vector-valued images (RGB). Both methods are visually evaluated with several artificial images and a satellite image. In opposition, LMS was proposed as an active contour algorithm that used local rather than global statistics, and it was visually evaluated using 10 natural and artificial images.

### 2.5.6 Merging Threshold

A different type of algorithm is created when performing both threshold and quantization. In the context of this work, the following method is explored as such:

- *Statistical Region Merging* (SRM) (Celebi et al., 2008) is presented as an inference approach that detects borders in dermoscopy images of PSL using an unsupervised implementation based on the statistical region merging algorithm. The inference problem models the image as an observed instance of an unknown theoretical image, whose statistical regions are to be reconstructed. The optimal statistical regions share a homogeneity property, such that inside any statistical region and given any colour channel, the pixels have the same expectation, whereas the expectations of adjacent statistical regions differ in at least one colour channel. This algorithm also contains a preprocessing stage which first removes the circular black area (by verifying the lightness component of the *CIE L*a*b** colour space), and then applies a median filter (image smoothing) in order to mitigate the detrimental effects of possible artefacts (hairs and bubbles, and skin lines).

SRM was proposed to be a fast and accurate way of producing skin lesion segmentation, which is compared with four state-of-the-art automated methods. This is done resorting to a border error metric and a dataset of images that lacks ground-truth images, where three expert dermatologists produced the target ground-truth segmentation. In Celebi et al. (2008) five algorithms are compared against a dataset with no ground-truths. Relevant to this work are: SRM, FCM, and RGB-MC. Segmentation results are presented for the benign or melanoma image classes. SRM is the only algorithm where the error rates do not increase in the melanoma class. The authors refer that is "possibly due to the presence of higher border irregularity and colour variegation in these lesions".

### 2.5.7 Other approaches

So far, the discussed algorithms are based on traditional approaches, which have been progressively evolving towards a new generation of methods. Taking advantage of recent developments in the field of machine learning, a new type of algorithms has emerged. Among such

recent machine learning methods, a few were specifically designed to segment skin lesions. From the literature, 13 algorithms are considered:

- *Fully Convolutional Networks* (FCN) (Long et al., 2015) is a method based on Convolutional Neural Networks that is presented as a proof-of-concept showing that convolutional networks by themselves, trained end-to-end, pixel-to-pixel, are able to capture feature representations that contain a high-level of semantic information through several convolutional layers. This is done by resorting to a skip architecture that combines semantic information (from deep and coarse layers) with appearance information (from shallow and fine layers) to produce accurate and detailed segmentation;
- *Saliency-based Skin Lesion Segmentation* (SSLS) (Ahn et al., 2015) presents a hair removal technique prior to the actual segmentation, consisting on a pixel level saliency map and a lesion biased Gaussian model;
- *Sparse Coding with Dynamic Rule-based Refinement* (SCDRR) (Bozorgtabar et al., 2016) is an unsupervised skin lesion segmentation method for dermoscopic images that exploits the contextual information of skin image at the superpixel level with a Laplacian multi-task sparse representation;
- *Multi-scale Superpixel with Cellular Automata* (MSCA) (Bi et al., 2016) uses image-wise supervised learning to derive a probabilistic map for automated seed selection. It also enables the inclusion of additional structural information in conjunction with a Multi-scale Superpixel-based Cellular Automata;
- *multi-stage FCN* (mFCN) (Bi et al., 2017) is a recent approach that learns and refines the skin lesion segmentation results across multiple stages. The algorithm then integrates these complementary multi-stage segmentation results in an ensemble-like fashion;
- *Joining Circular-Linear Mixture Model* (JCLMM) (Roy et al., 2017) models hue and chroma information assuming the multi-modal characteristics of skin lesions and deals with heterogeneous margins for different mixture components;
- *Convolutional-Deconvolutional Neural Network* (CDNN) (Yuan et al., 2017) is an automatic method for skin lesion segmentation, which leverages a deep convolutional neural network of 19 layers. The method includes a set of strategies that attempt to ensure effective and efficient learning with limited training data, as well as a loss function based on Jaccard distance to remove the need of sample re-weighting (when using cross entropy due to the strong imbalance between the number of foreground and background pixels);
- *Enhanced CDNN* (ECDNN) (Yuan & Lo, 2019) is an extended version of the previous work in Yuan et al. (2017) that develops a deeper network architecture with smaller kernels to enhance its discriminant capacity. The potential of using a deeper network architecture with smaller convolutional kernels is investigated such that the new model has increased discriminative capacity to handle a larger variety of image acquisition conditions. The use of channels in other colour space, such as Hue-Saturation-Value (HSV) and *CIE L*a*b**, is also investigated as additional inputs to the network that aim for a more efficient training while controlling over-fitting;

- *Kullback-Leibler based LevelSets* (KL-LS) (Riaz et al., 2019) is an active contour based method that uses Kullback-Leibler divergence between the lesion and skin to fit a curve to the lesion boundaries after taking an initial lesion contour. These can be defined using the gradient flow that minimises an appropriate cost function (that is embedded as an external energy term in the existing distance regularised level sets evolution);
- *Densely Linked Convolution Neural Network* (DermoNet) (Baghersalimi et al., 2019) is a deep neural network that encompasses techniques from several other segmentation-aimed networks like autoencoder network funnelling and residual propagation aided with dense convolution blocks;
- *Particle Swarm Optimization on Deep Ensemble Network* (PSO-DEN) (Tan et al., 2019) is an evolving ensemble of deep networks and hybrid clustering models, where the learning hyper-parameters are optimised with a cascading particle swarm optimisation algorithm and a majority voting strategy to combine the prediction results of each base model to produce the final pixelwise classification outcome;
- *Semi- and Weakly Supervised Directional Bootstrapping* (SWSDB) (Xie et al., 2019) is a combination of three deep networks: a coarse segmentation network (adapted from other works and pre-trained on other datasets), a dilated classification network (adapted and pre-trained from another work based on dense object location using dilation convolutions), and an enhanced segmentation network (consisting of an autoenconder, where the encoder and decoder are separated by a enhancement layer); in which the later is trained with a hybrid loss function comprising dice loss and rank loss metrics;
- *Deep Class-specific Learning with Probability based Step-wise Integration* (DCL-PSI) (Bi et al., 2019) is a deep convolutional neural network that is refined using a step-wise integration approach that, using the image label classification probability, iteratively maximises the pixel agreement between different network-models.

In Bi et al. (2017) seven algorithms are compared using a well known dataset. Relevant to this work are: SCDRR, JCLMM, MSCA, SSLS, FCN, and mFCN. The mFCN attained the highest similarity to the dataset ground-truth and is followed by: FCN, SCDRR, MSCA, JCLMM, and SSLS; in this order.

## 2.6 3D Skin Lesion Classification

Beyond the 2D analysis, few works have exploited the 3D information of PSL. McDonagh et al. (2008) obtained 3D shape moment invariant features from stereoscopy-generated images for computer-aided diagnosis. In order to automatically distinguish between melanoma and non-melanoma lesions, the features were fed into a Bayesian classifier along with relative colour brightness information, relative variability, and peak and pit density features.

Later, characterization approaches were attempted with photometric stereo (Anwar et al., 2012), allowing to extract features for lesion classification like skin tilt and slant patterns

(Smith et al., 2011), and statistical moments of enhanced principal curvatures of skin surfaces (Zhou et al., 2010). In Zhou et al. (2011), an ensemble classifier comprising three distinct classifiers was tested on enhanced 3D curvature patterns and a selected set of 2D features. The achieved results showed that 3D patterns alone did not outperform traditional 2D features, however, when combined with 2D features, its effectiveness was demonstrated in melanoma diagnosis. Recently, similar technologies using stereo-vision and structured light projections have also mapped melanoma to 3D (Peña Gutiérrez, 2016; Ares Rodríguez et al., 2014; Ding et al., 2015).

The extraction of 3D related features from a light-field surface, in which there are no 3D features specifically studied for melanoma classification, is relevant for this thesis' topic of 3D skin lesion information. Thus, a primary approach towards defining a relevant set of such features is to look at other research fields, where 3D features have been used. Depending on the target recognition task, several 3D features have been developed and generalised across multiple 3D datasets and tasks. This type of generalisation is performed to propose a set of features that capture a broad spectrum of 3D characteristics – typically applied to key regions, which are found by additional methods (hereinafter "keypoint detector"). Keypoint detectors are only necessary when feature extractors operate locally (i.e., in a region surrounding a point of interest), thus said keypoint must be first properly identified. An example of an algorithm capable of performing both keypoint detection and feature extraction (on the identified keypoints) is the Normal Aligned Radial Feature (NARF, Steder et al., 2011). The keypoint detector has two major characteristics. Firstly, keypoints are extracted in areas where the direct underlying surface is stable and the neighbourhood contains major surface changes. The resulting keypoints are located in the local environment of significant geometric structures and not directly on them. Secondly, NARF takes object borders into account, which arise from view dependent non-continuous transitions from the foreground to the background. Thus, the silhouette of an object has a strong influence on the resulting keypoints. The NARF keypoint detector pipeline is as follows: *(i)* transform point cloud to range image; *(ii)* find object borders; *(iii)* compute normals to border points; *(iv)* compute principal curvature for non-border points; *(v)* compute interest value for all points; and *(vi)* isolate keypoints. Having found areas of interest, the NARF feature extractor can now take place. The feature descriptor is computed by defining a normally aligned range value patch around the feature point, computed by constructing a local coordinate system, where the observer looks at the point along the normal. At this point, a star-shaped pattern is projected into the patch (where each beam corresponds to a value in the final descriptor) capturing how much the pixels under the beam change. Then, a unique orientation is extracted from the projection and the values are shifted accordingly, to make this rotation invariant.

Another relevant keypoint detector is the Intrinsic Shape Signatures (ISS) keypoint detector, which employs a saliency measure based on the eigenvalue decomposition of a scatter matrix of the points belonging to a support value (Zhong, 2009). These points are only retained if the ratio between two successive eigenvalues is below a predefined threshold. Their saliency is

determined by the magnitude of the smallest eigenvalue, in order to only include points with large variations along each principal direction. The rationale behind this pruning stage is that points exhibiting a similar spread along the principal directions (where a repeatable canonical reference frame cannot be established) should be avoided because a subsequent description stage would hardly turn out effective. Afterwards, a point will be considered a keypoint if it has the maximum saliency value on a given neighbourhood. Contrary to the NARF detector, the ISS is much more selective and inherently produces less keypoints, reducing the computation time.

In addition to NARF's feature extractor, other relevant methods for this thesis (for 3D characterisation) are now detailed in the following:

- In Marton et al. (2010, 2011), authors define the Radius-based Surface Descriptor (RSD) as a descriptor that depicts the geometric property of a point by estimating the radial relation with its neighbour points. First the radius is modelled as a relation between the distance of two points and the angle between their normals. Then, the maximum radius and minimum radius are recorded as the final features for each point.
- In Kanezaki et al. (2011), the RSD extractor is extended to the Global RSD (GRSD), which computes a global histogram for the whole point cloud. First, the input point cloud is voxelised and the RSD descriptor is generated for every voxel neighbourhood. Then, voxel surfaces are categorised into six possible surfaces based on a set of defined rules using the two RSD features. After categorising all voxels, the GRSD histogram relies on the number of transitions between all of these local categories, which results in 21 features. GRSD allows the use of depth images with or without colour information.
- In Rusu & Cousins (2011), authors define the Principal Curvatures (PC), which returns the eigenvector of the largest eigenvalue along with both the largest and the smallest eigenvalues after performing a Principal Components Analysis on the points normal of a surface patch (in the tangent plane of the given point normal).
- In Lazebnik et al. (2005), the Rotation Invariant Feature Transform (RIFT) is defined such that, given a point, it extracts a sparse set of affine covariant elliptical regions of the surrounding texture using the Harris affine or Laplacian blob detectors, which detect complementary types of structures, and normalise each elliptical region into a unit circle to reduce the affine ambiguity to a rotational one. Then, the method divides the circular normalised patch into four concentric rings with equal width and compute a gradient orientation histogram with eight orientations within each ring. This results in a descriptor of 32 features that is later adjusted for rotational invariance by the radial outward direction at each point.
- In Rusu et al. (2008), a method named Point Feature Histogram (PFH) that encodes the geometric properties of the k-nearest-neighbours of a point is defined, by using the average curvature of the multidimensional histogram around such point. This is done by calculating, for each pair of points, the difference of three angular variables (obtained from a Darboux frame where the third angular variable is normal to the point's plane)

and their euclidean distance. Finally a histogram is created with the 4 variables along each computed pair.

- In Rusu et al. (2009a,b), a variant of PFH, the Fast Point Feature Histogram (FPFH), is proposed as a computational simplification of PFH. In comparison, first, for each point, FPFH uses a method similar to PFH to calculate the three angular variables and obtain a simplified PFH. Then, a weighted neighbouring pairing is used to calculate the final value of the histogram, where the weights depend on the centre point and a neighbour point at a given distance metric space.
- In Tombari et al. (2010b, 2011), a method defined as Signature of Histograms of OrienTations (SHOT) is proposed, based on disambiguated eigenvalue decomposition of the covariance matrix of points within the neighbourhood region, where an isotropic spherical grid defines the signature structure. These locations produce local histograms by counting the number of points within a region of the spherical grid. The juxtaposing of all local histograms with quadrilinear interpolation generates the final collection of features.
- In Wohlkinger & Vincze (2011), authors define the Ensemble of Shape Functions (ESF), which comprises ten 64-sized histograms: three angle related histograms, three area related histograms, three distance related histograms, and one histogram of distance-ratio. The first nine histograms are created by respectively classifying an angle formed by randomly sampled three points, the area created by such three points, and a shape function. While the last is built on the paring-lines generated during the shape function execution.
- In Frome et al. (2004), authors define the 3D Shape Context (SC3D) as a descriptor that captures the local shape of a point cloud at a centre point using the distribution of points in a spherical support. Within this support, a set of bins is formed by equally dividing the azimuth and elevation, and logarithmically spacing the radial dimension. Then, the final descriptor is computed as the weighted sum of the number of points falling into bins.
- In Tombari et al. (2010a), authors define the Unique Shape Context (USC) as an improvement over the SC3D descriptor by adding a unique and unambiguous local reference frame, with the purpose of avoiding computation of multiple features at each keypoint. Given a query point and its spherical support region, a weighted covariance matrix is defined so that three unit vectors of a local reference frame can be computed from the Eigen Vector Decomposition of this matrix. The eigenvectors corresponding to the maximum and minimum eigenvalues are reoriented in order to match the majority of the vectors they depicted, while the sign of the third eigenvector is determined by the cross product. Once the local reference frame is built, the subsequent steps are analogous to those in SC3D.

# Chapter 3

# Segmentation and Classification of 2D Images

Content



Digital image segmentation is a key stage in medical image processing algorithms and machine learning classifiers, where the accuracy of the border-line that defines the ROI is of utmost importance for subsequent algorithms (e.g., classification, where previous identification of discriminative features only belonging to the lesion area may be of critical consideration). This step is common in computer-aided medical systems, which enable early

diagnosis of serious medical conditions, including in the problem of differentiating melanoma and nevus (Linsangan et al., 2018). Most of the existing image segmentation approaches aim at minimising some error metric between computed and ground-truth ROI defined by medical experts. From other works in the literature it is clear that proper segmentation of the lesion area impacts the classification pipeline performance (Burdick et al., 2017).

In the topic of skin lesion classification, employed algorithms range from those using DL, where the algorithm automatically learns which types of features will be employed for classification, to other classic ML algorithms which require hand-crafted features. The use of DL algorithms has achieved significant performances (e.g., Namozov & Cho, 2018; Hosny et al., 2019), however those algorithms require rich (sizeable and precisely annotated) datasets that are not widely available. In DL classification approaches, the segmentation mask can be used to drive the model to lesion area information, instead of having to initially learn to not speculate on the healthy skin colour information. On the other hand, in algorithms that require hand-crafted features, the use of a segmentation mask might be necessary to remove features extracted from healthy skin or even create features regarding the skin shape and size. Therefore, different segmentation masks might originate the inclusion or not of certain pixels (for DL) or features (when hand-crafting).

This chapter presents the contributions to the segmentation and classification of 2D skin lesion images in the scope of this thesis. To satisfy the segmentation needs of the target applications, two algorithms were proposed to address the limitations mentioned above. The first one, named *Gradient-based Histogram Thresholding* (GHT), Pereira et al. (2019b), and discussed in detail in Section 3.1, was created to provide segmentation results consistent with those presented by dermatology experts. The algorithm resorts to the high image gradient separations between lesion and healthy skin, not compromising the coarser delineation of the available ground-truth. The second algorithm, named *Local Binary Patterns Clustering* (LBPC), Pereira et al. (2019a), and described in Section 3.2, was created to be capable of finding more detailed borders of skin lesions, i.e., with much more detail than the usual (round) dermatologist-like segmentation, which is typically available in ground-truths. In addition, a comparative evaluation study was carried out using three datasets (in Pereira et al., 2020a) to demonstrate the performance of the LBPC algorithm against 38 other segmentation algorithms.

On the topic of lesion classification, in Section 3.3, a first pipeline is performed using Transfer Learning (TL) to provide an experimental baseline for the 2D skin lesion imaging classification on the large ISIC dataset (Pereira et al., 2018). In Section 3.4, a different classification approach is proposed to assess the importance of the border-detail of the segmentation algorithms. This approach, in Pereira et al. (2020b), relies on the details of the masks produced by the segmentation algorithms, which are used to provide features for the classification process.

Finally, Section 3.5 summarises this chapter and highlights the discussed materials.

## 3.1 Gradient-based Histogram Thresholding

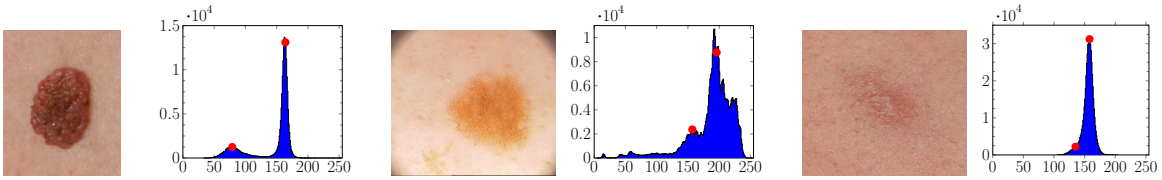
In general, medical image processing systems include a segmentation stage to identify a ROI for further processing, which may include texture and colour analysis, feature extraction, among others. In the case of PSL, due to the rather limited human capability to discriminate slight variations in contrast and blur, precise identification of relevant ROI boundaries poses a problem to dermatologists (Claridge & Orun, 2002). Morphological aspects of skin lesions alongside with the large number of environment-variables further increase the challenge of accurate segmentation of the most useful ROI (Celebi et al., 2009a). This results in significant inter- and intra-observer variability and coarse ROI segmentation. Thus, to reduce the dependence of human factors, different types of computational methods have been used for image segmentation, spanning over a quite considerable range of categories (Pathan et al., 2018).

This section presents an accurate segmentation method for PSL, envisaging delineation of melanoma as the main application. In general, this kind of medical images produce bi-modal histograms and, although this characteristic has been used as the basis of different segmentation methods, it results in either coarse borders or simply fails to provide significant ROI in images with low colour contrast and smooth texture transitions. Therefore, this section addresses the problem of accurate identification of the relevant ROI in such images, which includes the ability to define the external border of the lesion with high level of precision. Taking into account the clinical relevance of this aspect, a gradient-based metric is devised to drive the proposed delineation method across a refinement histogram-based segmentation algorithm. Two different types of medical images are targeted, which increases the challenge of achieving efficient segmentation with accurate border details in both cases, as pointed out in Zhou et al. (2008). For this purpose, the Dermofit dataset of macroscopic images and the $PH^2$ dataset of dermoscopic images are both used in this work.

This section is organised as follows: Section 3.1.1 presents the background that is relevant for the proposed method. Section 3.1.2 describes the Gradient-based metric and Section 3.1.3 introduces the proposed method. In Section 3.1.4 the obtained experimental results are presented, alongside with the associated discussion. Finally, in Section 3.1.5 some conclusions and future work perspectives are presented.

### 3.1.1 Relevant Background

As mentioned above, images of skin lesions exhibit two distinctive regions, one associated to the lesion itself and another to the surrounding skin. This leads to bi-modal histograms, as shown in Fig. 3.1.



**(a)** *A121a* from Dermofit **(b)** *IMD390* from PH$^2$ **(c)** *A67* from Dermofit

**Figure 3.1:** Skin lesion images (left) and the corresponding luminance ($Y$) histograms (right), were red dots represent peaks that correspond to the lesion and skin, respectively.

However, their histograms may present different characteristics. This results not only from the lesion morphology, but also from the use of different image acquisition technologies and lighting conditions. For instance, while accurate segmentation of the skin lesion shown in Fig. 3.1a is quite easy to be obtained directly from its well-defined histogram, in the case of Fig. 3.1c it poses a problem due to the blurry borders and low colour contrast. This can be confirmed by the corresponding histogram shown on the right side of Fig. 3.1c, where the pixels belonging to the relevant ROI are quite difficult to identify.

**Histogram Thresholding** Histogram thresholding techniques have long been used for segmentation of these type of bi-modal images, where the region of the lesion can be distinguished by its different tonality (Korotkov & Garcia, 2012). The underlying idea of these methods is to perform a binary partition of the image based on the luminance level of each pixel, meaning, in this case (e.g. Fig. 3.1), to successfully separate the region of the lesion (darker region ↔ left $Y$-peak: $Y_{Pmin}$) from the surrounding skin (right $Y$-peak: $Y_{Pmax}$). In a simple formulation, the challenge of segmentation, in this application, is to find an optimum criterion to define a threshold value ($Y_{th}$) that leads to an accurate ROI extraction, i.e., the region of the image that contains the lesion. Different threshold techniques have existed for decades (Sahoo et al., 1988) and can be performed either directly on the $Y$-histogram or after a transformation, as proposed in Rajab et al. (2004). Nevertheless, the performance of the method might strongly depend on the distribution of luminance values, as inferred from Fig. 3.1c and even from Fig. 3.1b.

**Clustering** Clustering algorithms have also been used for skin lesion segmentation based on different approaches, as described in Section 2.5.2. These algorithms can be fed with image data in different formats such as RGB, YUV, or YCbCr, but there are also systems using only the luminance Y channel since the inherent fusion process of the RGB channels allows inclusion of the relevant colour information (Maglogiannis & Doukas, 2009). An efficient clustering approach is based on the so called *K-Means*, or Lloyd's algorithm (Lloyd, 1982), which is an iterative data-partitioning algorithm that assigns, for a predefined number of clusters, every input observation to only one of the clusters. In skin lesion images, clustering algorithms may take advantage of the bi-modal characteristics of the histogram to use the corresponding peaks as the initial centroids.

**Other Approaches** Several other segmentation approaches are used in this section as baseline literature: FDEE and FC-LS as fuzzy methods; PCT-MC as quantization approaches; CH and LMS as active contours; and OT and KMC as other traditional methods.

### 3.1.2 Gradient-based Metric

The figure of merit that is herein proposed to assess the accuracy of a given outside border-line is based on the rationale that the segmentation contour separates regions with substantially different tonalities. Accordingly, the magnitude of the image gradient on contour pixels is expected to yield higher values than in other regions (i.e., either inside the lesion or in the remaining skin). Following this argument, segmentation masks whose border-lines exhibit higher gradient magnitudes should separate more accurately the two regions of the image.

For a given image $I$, the gradient magnitude in each pixel can be determined by Eq. (3.1), where $j$ denotes the colour (or luminance) channel under consideration (e.g., $j = R, G, B, Y$). For any point on the border-line defined by segmentation, the projection of the gradient vector $\vec{G}$ onto the orthogonal direction of the line ($\hat{n}_l$), defines an accuracy metric for the border-line. Such projection is given by Eq. (3.2).

$$\left|\vec{G}_j\right| = \left[\left(\frac{\partial I_j}{\partial x}\right)^2 + \left(\frac{\partial I_j}{\partial y}\right)^2\right]^{1/2}, \tag{3.1}$$

$$G_{\perp_{l,j}} = \vec{G}_{l,j} \cdot \hat{n}_l \tag{3.2}$$

This concept is depicted in Fig. 3.2, where one can observe that the orthogonal direction of the segmentation border-line (red) is not aligned with the gradient (blue) at the same point. The higher the projection $G_{\perp_{l,j}}$ computed by Eq. (3.2) the better (i.e. more accurate) is the lesion contour segment.

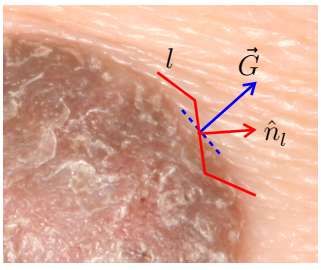



**Figure 3.2:** Segmentation border line (red) and gradient: the higher the projection of the image gradient $\vec{G}$ onto $\hat{n}_l$, the more accurate is the border line (image *P348a* from Dermofit).

Therefore, following the above discussion, the average value of $G_{\perp_{l,j}}$ over all points of a contour line $l$ is used as the gradient-based metric to evaluate how accurately a given segmentation contour represents the outside border on a skin lesion.

### 3.1.3 Proposed Dermatologist-like Segmentation Method

The image segmentation method proposed for skin lesion delineation follows the representation depicted in Fig. 3.3.

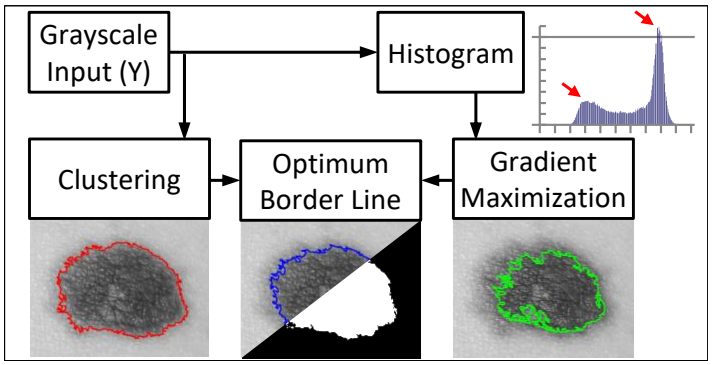



**Figure 3.3:** Gradient-based Histogram Thresholding method workflow: given a grayscale input image, an histogram is produced to find the two dominant colour peaks (arrows) of the image, which provide boundaries for the RGB gradient maximisation step (green segmentation line); the same input image goes through a clustering step (red segmentation line); finally, from the two previous segmentation lines, an optimum border-line optimum is obtained (blue curve) from which image binarisation produces the final mask.

The underlying idea is to find an optimal ROI delineation based on a trade-off between a ROI with the highest gradient magnitude in the orthogonal direction of its border-line, and another ROI with larger area but lower gradient. While the former identifies the sharpest boundary of the skin lesion, the latter contains more boundary information which is also useful for medical analysis and monitoring of temporal evolution. As described in the following, gradient-based histogram thresholding and clustering are used to generate both ROIs for final optimisation and delineation of skin lesions. A more detailed processing chain is shown in Fig. 3.4.

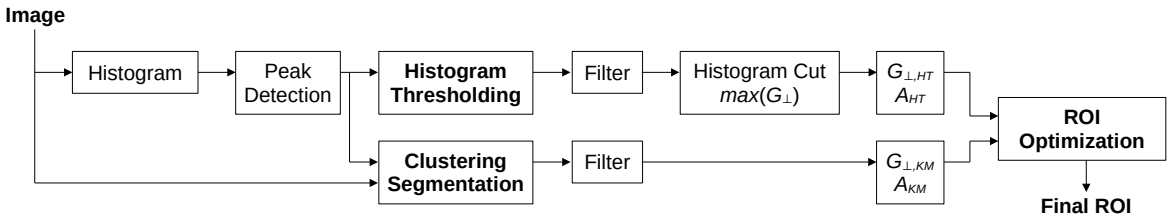



**Figure 3.4:** Proposed Gradient-based Histogram Thresholding scheme.

**Histogram Thresholding** The bi-modal characteristic of skin lesion image histograms is used to determine the two most important peaks, which in turn define the range limits $Y_{Pmin}, Y_{Pmax}$ for all possible thresholds, i.e., the best ROI must be found by cutting the histogram at the optimum threshold $Y_{th^*} \in [Y_{Pmin}, Y_{Pmax}]$ to be found between the two peaks (e.g. red markers on the histograms of Fig. 3.1). In the method shown in Fig. 3.4, histogram thresholding is performed for all values between $Y_{Pmin}$ and $Y_{Pmax}$, generating an equal number of images and the corresponding segmentation masks. After a filtering process to remove small isolated regions and outliers, a clean ROI is determined for each image and

the average gradient $G_{\perp}$ is computed, as defined by Eq. (3.2). Then, the ROI whose border yields the gradient with the maximum average is selected along with its histogram threshold. In summary, this process ensures that the border-line of such ROI is the one with the highest tonality variations across it. In the remaining sections, the ROI obtained by maximising the gradient border through histogram thresholding is identified by $HT$.

**Clustering Segmentation** In the proposed method, clustering is used to identify a coarse ROI for the skin lesion where the boundaries include, in general, the smooth transition regions between the lesion and the surrounding healthy skin. In this work the variant *K-Means++* was selected, due to its faster clustering convergence and also good discriminative performance due to different heuristics used for finding centroids (Arthur & Vassilvitskii, 2007). The iterative clustering process is carried out with a maximum of 200 iterations seeking for two clusters with a global minimum of the euclidean distance to cluster-centre. For the sake of reproducibility, the initial centroids are defined as the histogram peaks. In the remaining sections, the ROI obtained through clustering is identified by $KM$.

**Filtering** Due to noise, inherent illumination variations, and other factors, both the histogram thresholding and clustering methods described above produce ROIs with binary masks that include not only a large blob (the skin lesion region) but also other small isolated regions spread across the whole image. The filtering process devised to remove such unwanted regions assumes that the lesion region limits are fully located within the image, so the first operation is to remove all isolated regions with any boundary coincident with the image borders. This is done by using a flood-fill algorithm based on morphological reconstruction (Soille, 2013). This first cleansing operation is especially relevant when processing images from the $PH^2$ dataset, as they exhibit a black circular frame artificially introduced in the dermatoscope digitisation process. The relevant ROI containing the lesion is then determined by extracting the blob with the largest area in the binary mask, using a labelling procedure (Haralick & Shapiro, 1992, p. 40-48).

**ROI Optimisation** The optimisation step aims at improving delineation of skin lesions by selectively expanding the ROI border that was previously found through histogram thresholding with gradient maximisation, in order to further include relevant areas of transition regions. This is necessary because gradient maximisation often leads to stringent contour lines which only include the inner parts of the lesion and leave out relevant transition regions. Taking into account that clustering-based segmentation usually results in the inclusion of a larger region around the inner part of the lesion, the proposed optimisation procedure achieves the best trade-off between gradient maximisation and increased ROI area to include transition regions. This is done by finding an optimum threshold ($Y_{th^*}$) that maximises, simultaneously, both the gradient of the border-line $G_{\perp}$ and the ROI area.

For each ROI obtained through histogram thresholding ($HT$) and clustering ($KM$), as described above, let us define the following reference values:

- $G_{\perp,KM}$ and $G_{\perp,HT}$: the average gradient of the border, in Eq. (3.2);
- $A_{KM}$ and $A_{HT}$: the area of the ROI, i.e. skin lesion.

The corresponding values associated to an arbitrary histogram threshold $Y_{th}$ are defined as $G_{\perp,Y_{th}}$ and $A_{Y_{th}}$, respectively. The ratios $R_G(Y_{th})$ and $R_A(Y_{th})$, in Eq. (3.3), define the relative gradient and relative area of any ROI obtained with threshold $Y_{th}$, using the $HT$ and $KM$ ROIs as references.

$$R_G(Y_{th}) = \frac{G_{\perp,Y_{th}}}{G_{\perp,KM}} \quad R_A(Y_{th}) = \frac{A_{Y_{th}}}{min(A_{KM}, A_{HT})} \tag{3.3}$$

The optimisation procedure consists in finding the optimum threshold $Y_{th^*}$ that maximises both $R_G(Y_{th})$ and $R_A(Y_{th})$. This is accomplished by maximising their product, provided that the selected maximum does not lead to gradient values below that of the $KM$ ROI ($G_{\perp,KM}$) and the new ROI area falls between those of $HT$ and $KM$ ROIs. This is equivalent to solve the following constrained maximisation problem in Eq. (3.4).

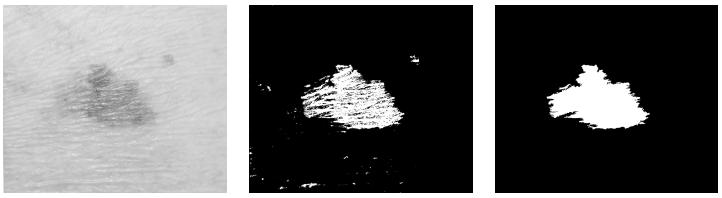
$$Y_{th^*} = \arg\max_{Y_{th} \in Y} R_G(Y_{th})\, R_A(Y_{th}) \quad s.t. \begin{cases} R_G \geq 1, \\ R_A \geq 1, \\ \frac{A_{Y_{th}}}{max(A_{KM}, A_{HT})} \leq 1. \end{cases} \tag{3.4}$$

In summary, the optimal histogram threshold $Y_{th^*}$ to be used for delineation of skin lesions is found through a trade-off between the border gradient and the amount of transition area included in the ROI.

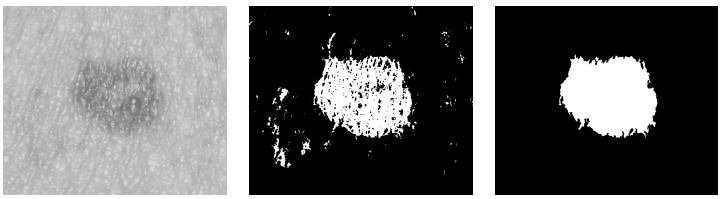
### 3.1.4 Results and Discussion

The performance of the segmentation algorithm described in Section 3.1.3 was evaluated using sets of images from different databases. A total of 195 images from the Dermofit dataset and 32 images from the $PH^2$ dataset were used. This selection followed two main criteria: *i*) images without hair strands crossing the lesion, i.e., as hairless as possible and *ii*) lesion limits within the image, i.e., the whole lesion boundary fully located inside the image.

In the first stage the input grayscale image passes through two segmentation processes, namely Histogram Thresholding and Clustering Segmentation. As pointed out in Section 3.1.3, the output of both algorithms may exhibit some image artefacts, which can be observed in Fig. 3.5b and Fig. 3.5e. Then the efficiency of the filtering stage, that is used after both segmentation algorithms (described in Section 3.1.3), in removing the small isolated regions, is shown in Fig. 3.5c



**(a)** Image *B964* **(b)** Histogram TH **(c)** Filtered

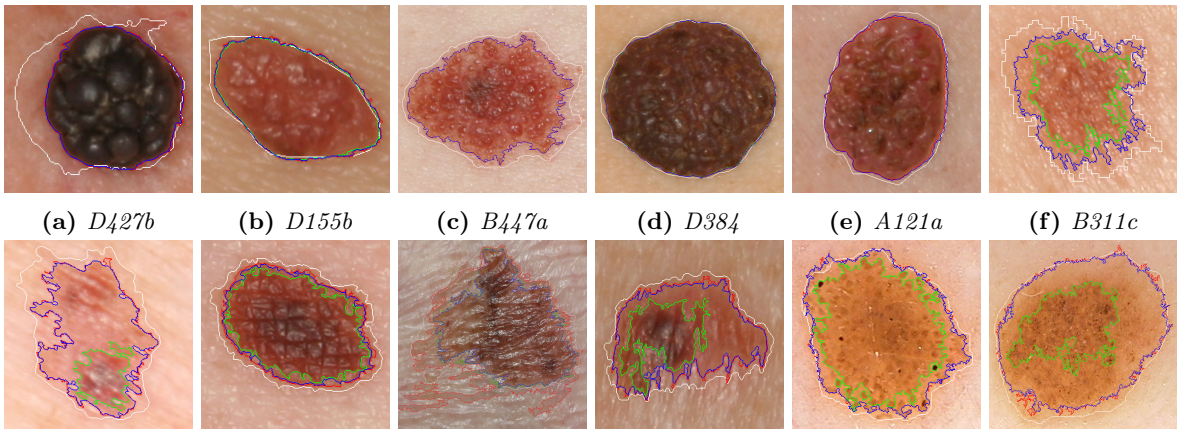


**(d)** image *B241* **(e)** Clustering **(f)** Filtered

**Figure 3.5:** Image segmentation using *Histogram thresholding (HT)* and *K-Means (KM)* clustering: (a,d) original grayscale images from Dermofit, (b,e) segmentation output, (c,f) final binary mask after the filtering operation.

and Fig. 3.5f. In these images it is possible to observe that the filtering process is effective in providing an accurate lesion/skin segmentation mask without harming the border details.

After the segmentation and filtering stages, accurate ROI delineation is performed, following the optimisation procedure described in the previous section. For visual evaluation and discussion, a set of representative types of skin lesions have been selected from the datasets to represent the segmentation results, as can be seen in Fig. 3.6.



**(a)** *D427b* **(b)** *D155b* **(c)** *B447a* **(d)** *D384* **(e)** *A121a* **(f)** *B311c*

**(g)** *A92* **(h)** *B17a* **(i)** *B964* **(j)** *B1075* **(k)** *IMD103* **(l)** *IMD175*

**Figure 3.6:** Skin lesion segmentation using *KM* (red), *HT* (green) and *Proposed* (blue). The white line corresponds to the dataset provided ground-truth (GT). Images (a) to (j) are from Dermofit, and (k) and (l) from $PH^2$.

In Fig. 3.6 the lesion segmentation using KM is represented by a red line, the HT by a green line and the proposed method by a blue line. The white line represents the ground-truth (GT) provided by the dataset. From the representative results presented in Fig. 3.6a to Fig. 3.6e, it can be observed that the algorithms are in general effective in the segmentation of images and delineation of the relevant ROI.

The Histogram Thresholding method ($HT$) is able to achieve accurate delineation when there is a sharp tonality difference between the skin lesion and the surrounding skin. However, as mentioned before, in images with smoother lesion-to-skin transitions, the highest value of $G_\perp$, may result in a segmented region that is smaller than expected. This effect can be seen in images from Fig. 3.6f to Fig. 3.6l. This kind of output is not the most useful from the clinical point of view, as it may exclude a relevant part of the lesion.

In the case of Clustering ROI segmentation ($KM$), in general, the segmented region may include smooth transition regions between the lesion and the surrounding healthy skin. This commonly results in a larger region than that obtained by the Histogram Threshold method, as can be seen in images Fig. 3.6h, Fig. 3.6i, Fig. 3.6j and Fig. 3.6l. In such cases, this might not represent the best option as well.

In order to overcome the $HT$ underestimation of the ROI and the possible $KM$ overestimation, the proposed combined method relies on a trade-off between the ROI and the border gradient. As can be observed in all images of Fig. 3.6, the blue line always represents a more precise delineation of the ROI.

Some authors compare the segmentation results with the ground-truth segmentation masks provided in the databases. Nevertheless, as can be visually observed in Fig. 3.6a, Fig. 3.6b and Fig. 3.6f, the GT borders are not as accurate and spatially detailed as those obtained with the used algorithms. It can also be observed that the GT lines often miss areas with high texture variations.

In a quantitative evaluation, other performance indicators are usually considered as benchmarks, namely Border Error (BE), True Detection Rate (TDR), and False Detection Rate (FDR). The results obtained for these indicators are presented in Table 3.1, alongside with those from the methods presented in Section 3.1.1. It can been seen that the performance of the proposed algorithm (GHT) is generally inline with others published in the literature, though not always consistent for all metrics. However, it should be kept in mind that these indicators use the GT as reference, which does not provide segmentation masks with as much spatial details as those herein obtained. Such difference can be clearly observed in Fig. 3.6

The gradient metric defined in Section 3.1.2 was also used to assess the performance of the proposed method. The quotient of gradient between delineations for both datasets was determined for such purpose and the results are presented in Table 3.2. Observing its first three lines, it can be seen that HT has on average the highest $G_\perp$ values, as the method was optimised for such purpose, though in some cases this also corresponds to inaccurate segmentation. The second group of three lines make it clear that the GHT method outperforms KM while only slightly compromising the $G_\perp$ value in comparison with the maximum of HT. The remaining data on the table also shows that the GHT method produces segmentations with $G_\perp$ values higher than any of the other algorithms previously introduced. This means that skin lesion delineation obtained

**Table 3.1:** Ground Truth (GT) based indicators.

| Method | Dermofit | | | PH2 | | |
|---|---|---|---|---|---|---|
| | BE | TDR | FDR | BE | TDR | FDR |
| OT | 32.569 | 76.386 | 7.483 | 20.907 | 83.999 | 7.061 |
| KMC | 53.475 | 78.439 | 8.581 | 18.174 | 86.041 | 6.313 |
| FDEE | 1.017 | 37.819 | 25.731 | 50.499 | 50.734 | 15.319 |
| FC-LS | 2.439 | 49.500 | 50.084 | 20.352 | 82.846 | 6.872 |
| PCT-MC | 1.504 | 97.456 | 23.096 | 1.107 | 93.442 | 24.524 |
| CH | 2.383 | 69.381 | 35.618 | 48.521 | 68.161 | 14.119 |
| LMS | 2.164 | 0.0003 | 52.467 | 1.604 | 94.543 | 36.550 |
| HT | 36.253 | 64.920 | 9.409 | 34.763 | 65.586 | 11.165 |
| KM | 21.592 | 79.336 | 5.554 | 17.732 | 85.113 | 6.085 |
| **GHT** | 23.034 | 78.283 | 6.008 | 20.336 | 81.676 | 6.950 |

**Table 3.2:** Average Border Gradient Ratio ($G_{\perp,i}/G_{\perp,j}$) Indicator.

| Indicators | Dermofit | PH2 |
|---|---|---|
| $G_{\perp,\mathrm{HT}}$ / $G_{\perp,\mathrm{KM}}$ | 1.173 | 1.271 |
| $G_{\perp,GHT}$ / $G_{\perp,\mathrm{HT}}$ | 0.933 | 0.917 |
| $G_{\perp,GHT}$ / $G_{\perp,\mathrm{KM}}$ | 1.082 | 1.086 |
| $G_{\perp,\mathrm{HT}}$ / $G_{\perp,\mathrm{GT}}$ | 3.908 | 4.742 |
| $G_{\perp,\mathrm{KM}}$ / $G_{\perp,\mathrm{GT}}$ | 3.364 | 3.997 |
| $G_{\perp,GHT}$ / $G_{\perp,\mathrm{GT}}$ | 3.688 | 4.301 |
| $G_{\perp,GHT}$ / $G_{\perp,\mathrm{OT}}$ | 1.831 | 1.916 |
| $G_{\perp,GHT}$ / $G_{\perp,\mathrm{KMC}}$ | 1.831 | 1.921 |
| $G_{\perp,GHT}$ / $G_{\perp,\mathrm{FDEE}}$ | 2.268 | 1.787 |
| $G_{\perp,GHT}$ / $G_{\perp,\mathrm{FC-LS}}$ | 6.429 | 1.829 |
| $G_{\perp,GHT}$ / $G_{\perp,\mathrm{PCT-MC}}$ | 2.894 | 2.701 |
| $G_{\perp,GHT}$ / $G_{\perp,\mathrm{CH}}$ | 1.175 | 1.086 |
| $G_{\perp,GHT}$ / $G_{\perp,\mathrm{LMS}}$ | 3.376 | 3.191 |

by the proposed method is more accurate than the others because the border-line is found where the gradient is higher, i.e., a better discrimination between lesion and normal skin is obtained.

### 3.1.5 Conclusions

This section addressed the segmentation of skin lesion images using both histogram thresholding and clustering algorithms to overcome the limitations presented by each method when used individually. A gradient-based method was devised for optimised thresholding and ROI border quality parameter. The segmentation masks obtained for the final ROIs indicate that this method is accurate in delineation of the relevant lesion regions containing for a wide range of images. The experimental validation, using two publicly available images datasets, shows that the proposed approach is effective in delineating skin lesions with detailed geometry in

regions with diverse tonality variations. The accurate delineation of skin lesions is a relevant achievement to provide discriminative features for machine learning algorithms and also to investigate patterns of temporal evolution of the borders.

## 3.2 Local Binary Pattern Clustering

As mentioned in the previous section, amongst all image processing steps commonly used in dermoscopic images, the identification of the ROI is of central importance in the classification framework (Korotkov & Garcia, 2012). In addition to the ROI delineation, segmentation procedures are also used to extract other skin lesion information, namely the dynamics of its growth process (Mendes et al., 2016).

The manual (round-like) segmentation obtained by dermatologists, and used as ground-truth in the majority of image datasets, is mostly performed to identify surgical borders for excision, lacking an objective rule or metrics. Moreover, variations in lightning conditions can influence contrast and blur, thus precise identification of skin lesion boundaries poses a problem to manual segmentation (Claridge & Orun, 2002). Even when clinicians are guided to perform a cell-like based delineation of the lesion, this procedure has proven to suffer from observer variability (Claridge & Orun, 2002; Joel et al., 2002; Iyatomi et al., 2006). As a consequence, the resulting ground-truth, cell-like based, handmade segmentation lacks definiteness.

In the literature, a broad range of segmentation algorithms have been proposed, mostly covering the above mentioned round-like segmentation. These applications range from smoothing and thresholding, to colour space conversions, to exploit specific aspects of skin dermoscopic images, as reported in Oliveira et al. (2016). In fact, this wide range of methodologies is related with dataset diversity regarding physical acquisition conditions (e.g., light, angle of view), anatomical and local artefacts (e.g., hairs, skin curvatures), and equipment properties (e.g., lens, light, image resolution) (Celebi et al., 2009a; Zhou et al., 2008).

Although cell-like based delineation is useful for assessment of lesion growing dynamics, it is difficult to obtain a ground-truth reference for each image. Note that this type of segmentation is absent in datasets and the manual delineation can be influenced by external factors.

This section presents a contribution to overcome the previously described shortcomings, by proposing an algorithm for cell-like based segmentation, which is independent of human factors. The segmentation algorithm is based on the combination of LBP and K-Means clustering with aim for detailed delineation in dermoscopic images. In comparison with usual dermatologist-like segmentation (i.e., the available ground-truth), the proposed method is capable of finding more realistic borders of skin lesions, i.e., with much more detail. To promote the proposed algorithm, a comparison is performed against 39 other literature works.

This section is organised as follows: Section 3.2.1 presents the background that is relevant for the proposed method. Section 3.2.2 presents the proposed method. Section 3.2.3 presents results achieved using the proposed method and the reviewed methods, as well as comparative analysis. Finally, Section 3.2.4 draws some conclusions and suggestions for future work.

### 3.2.1 Relevant Background

Specifically for this section, it is relevant to know that, in general, regions of normal skin in dermoscopic images present flatter texture when compared to regions within the lesion. This characteristic can be exploited in order to identify those different spots by using LBPs (Ojala et al., 1996). The LBP operator is a 2D texture descriptor that assesses local variations on the image, and codes them in terms of a spatial pattern with an associated grayscale scheme. The underlying idea behind LBP operators is that texture can be represented locally by two complementary components: a spatial pattern and a corresponding strength. In fact, LBPs can be seen as an image operator, whose output is an array of integer labels describing small-scale variations (high frequency content) in the image. These labels, or their statistics, can then be used for further image analysis. Note that there are some variants of the classical LBP algorithm.

### 3.2.2 Proposed Detailed Segmentation Method

The method herein proposed (LBPC) comprises a sequential processing flow, as depicted in Fig. 3.7, which accepts generic dermoscopic images as input. The three main functional blocks are the LBP processing, space transformation, and clustering, all of which implemented and evaluated resorting to Matlab®.

The luminance ($Y$) channel is obtained from a given input image and its pixel-wise LBPs are computed. Then, a set of LBPs are selected and filtered with a Gaussian kernel to expand the region towards the remaining ones, generating an image ($L$) comprising several homogeneous regions. These regions provide rich information for the clustering algorithm, discriminating the various skin regions with different intensity levels. This is due to the fact that the extracted LBPs mark smooth surface areas on the object/image (a characteristic that is not usual in the lesion area, given their chaotic appearance).

The subsequent step consists of a space transformation where the $Y$ and $L$ images are firstly combined to ease visual separation. To this end, first a new three channel image is formed by stacking the $L$ image with the luminance ($Y$) and $L$ again. Then, this image stack is converted using the well-known RGB to *CIE L*a*b** transform, as a means of obtaining a pre-clustering function for the next step. As explained in detail below, the $Y$ and $L$ dimensions have some overlapping information where this transformation leads to a better discrimination of the

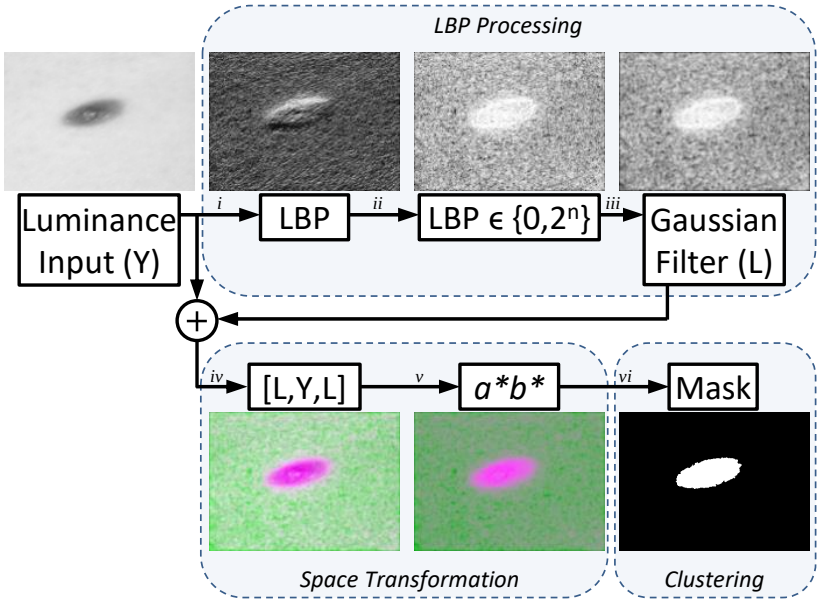



**Figure 3.7:** Local Binary Pattern Clustering method workflow: given a luminance input image $Y$, pixel-wise LBP information is obtained $(i)$ so that a specific set of LBPs is extracted $(ii)$ and smoothed with a Gaussian filter $(iii)$; then, this information, named $L$, is combined with the input $Y$ in a particular fashion $(iv)$ so that, after a conversion to the *CIE* $L^*a^*b^*$ colour space, the $a^*b^*$ channels present the rearranged data into an optimised form $(v)$ which, when fed to a clustering algorithm, groups the information into the two desired regions $(vi)$.

image data into two groups of pixels (lesion *vs* non-lesion), thus easing the final clustering process. With this new space, only value-providing dimensions ($a^*$ and $b^*$) serve as input to a clustering algorithm that separates the information into two cluster regions: of normal skin and lesion skin. The remainder of this section further details the processing pipeline steps shown in Fig. 3.7. Note that the segmentation method works performs equally even if the lesion has not a round shape or if it presents blurry edges.

***i)* Luminance to LBP** Given an RGB image as input, the corresponding luminance image is obtained by means of a weighted sum of R, G, and B channels, as proposed by Recommendation ITU-R BT.601-7 (2011) with four decimal places. The luminance information $Y$ is then converted into a Local Binary Pattern (LBP) (Ojala et al., 1996) code for each image using the definition in Eq. (3.5), where $I_p$ and $I_c$ are, respectively, the intensity of the peripheral and central pixels, and $p$ is the number of neighbouring points. As the neighbourhood of each pixel consists of 8 other pixels, a total of $2^8 = 256$ different labels can be obtained, depending on the relative values of the central pixel and its neighbours.

$$\text{LBP} = \sum_{p=0}^{7} s(I_p - I_c)2^p, \; s(x) = \begin{cases} 1 \; if \; x > 0 \\ 0 \; if \; x \leq 0 \end{cases} \tag{3.5}$$

***ii)* LBP subset** In this step it is important to understand some LBP characteristics. A given LBP can be rotated up to seven times (in the 2D plane) to generate seven other LBPs. This means that many LBPs are merely a rotation of another LBP. For example, LBP of value

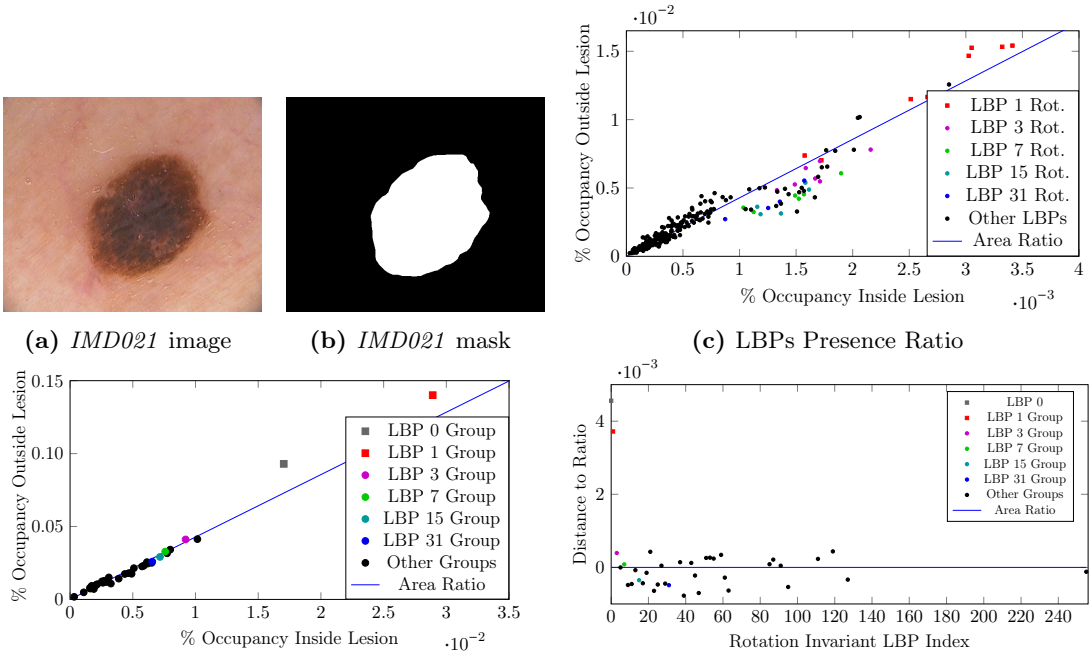
1 (represented by 00000001) can be (binary) rotated to generated LBPs 2, 4, 8, 16, 32, 64, and 128 (represented by 00000010, 00000100, 00001000, 00010000, 00100000, 01000000, and 10000000, respectively). When LBPs are grouped in this fashion they are said to be Rotation Invariant (Ojala et al., 2000). This grouping turns the 256 LBPs into 36 groups of patterns.

By convention, the LBP group number is defined by the lower LBP value present in the group. For example, LBP-group-3 is comprised by LBPs 3 and its rotations (which are: 6, 12, 24, 48, 96, 192, and 129). Following the previous description, it is now possible to interpret that these 36 groups have specific and comprehensible patterns. Namely, LBP 0 can represent a flat surface, since it can be constructed when the central pixel is surrounded by pixels of equal or inferior value. This is relevant because in skin lesion images the skin area is generally smooth and the lesion area is not. Therefore LBPs representing smooth patterns will be largely present in the skin area, while noisy patterns will dominate the lesion area.

Resorting to the image presented in Fig. 3.8a and its ground-truth mask in Fig. 3.8b, plot Fig. 3.8c is constructed by extracting all LBP values generated from the image and checking whether they were originated from either the skin or the lesion region of the ground-truth mask. The plot displays the relationship between the frequency of occurrence (in percentage) of each LBP for each region. The ratio between the two region areas is depicted by the blue line. When an LBP is located below the blue line it means that it occurs more often inside the lesion region. Contrarily, when an LBP is located above the blue line, it means that it occurs more often outside the lesion region (i.e., in the skin area). Finally, if an LBP is on the blue line it means that it occurs by the same proportion in both regions of the image. With this information it is possible to visualise that some LBPs (and special groups) are particularly suited to discriminate either the lesion area or the skin area, which is the case for the LBPs rotated from 1 ("*LBP 1 Rot.*" in red).

Grouping the rotation invariant LBPs from Fig. 3.8c allows for a simpler visualisation, hence Fig. 3.8d is presented. Here, apart from the LBP 0 and 1 group, it is clear that almost all LBP groups tend to be present by the same amounts in either regions. For example, it is clear that LBP group 3 is less dominant than group 0, being also of small importance since it is located much closer to the blue line. Therefore this means that only LBP 0 and 1 groups (which can also be expressed as LBPs of power of 2) are possible candidates to provide information about the type of region the current LBP represents (in this case, the normal skin where the aforementioned property regarding flat textures exists). For a better understanding, data in Fig. 3.8d was rotated (according to the reference blue line slope) to align the reference line with the x-axis. Then, the distance of each LBP group to the line in Fig. 3.8d appears as shown in Fig. 3.8e. The x-axis position of each data points was changed to match the LBP group pattern number.

Hence, in this step, the obtained LBP image is filtered so that only the pixels with an LBP corresponding to one of the selected patterns remain, namely LBP values corresponding to powers of two or zero (LBP $\in \{0, 2^n\}, n \geq 0$), thus limiting the points to only 9 out of 256




**(a)** *IMD021* image **(b)** *IMD021* mask **(c)** LBPs Presence Ratio

**(d)** Invariant LBPs Presence Ratio **(e)** Invariant LBP Significance

**Figure 3.8:** Analysis of image (a) against 256 LBPs (c) and later grouped by binary rotation invariance (d) and (e). Where (e) represents a rotation of plot (d) so that the blue-line matches the x-axis and with points rearranged so that the x-axis represents the LBP value. Blue-line represents the ratio between the number of pixels outside and inside the lesion given its ground-truth mask (b) from the dataset Mendonça et al. (2013).

possible LBP patterns. These selected values are represented by zeros in the image matrix and the remainder of the image is represented by ones, which are concentrated in the lesion area. Only these patterns have been chosen, as they are able to represent the situation where the algorithm is more sensitive to region transitions, i.e., a luminance difference between a pixel value and its neighbours.

***iii)* Gaussian Filter** A Gaussian filter (Haddad & Akansu, 1991), defined by Eq. (3.6), is applied to the binary LBP image, thus creating the new matrix $L$ with values ranging from zero to one. The 2D-Gaussian kernel with a standard deviation $\sigma = 3$ and size of 13 pixels, generates homogeneous regions in the LBP-image, namely in the lesion, thus removing most pixel-level noise. This regions' representation provides rich information to the clustering algorithm during the segmentation process, since they highlight only one of the two image segments. Note that these values were empirically attributed based on some of the early algorithms results. Thus, 13 pixels and $\sigma = 3$, which act as a low passing filter, may not be optimal values for all images.

$$F(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3.6}$$

***iv,v)* Space transformation** As mentioned above, the space transformation is a pre-clustering operation that aims to produce an image with improved discrimination between lesion/non-lesion image regions (steps *iv)* and *v)* in Fig. 3.7), which favours the clustering algorithm ahead. The underlying idea is to use the specific characteristics of the RGB to

*CIE L*a*b** colour space conversion (Joint ISO/CIE Standard, ISO 11664-4:2008(E)/CIE S 014-4/E:2007, 2007) to obtain such discrimination as, in this case, it inherently contributes to the normalisation of the data and dimensionality reduction. In this colour space conversion, the $a^*$ dimension discriminates between $R$ and $G$ by representing $R$ and $G$ in the positive and negative ranges of the $a^*$ axis, respectively. The $b^*$ dimension discriminates between blue and yellow and, finally, the $L^*$ dimension represents lightness.

Taking these characteristics into account, a correspondence between RGB and the image stack $LYL$ previously referred is established, such that $L \rightarrow R$, $Y \rightarrow G$, and $L \rightarrow B$. In one hand, when the $LYL$ image stack is observed in the RGB space, it is viewed as an image containing pink colouring in the lesion area, due to the low concentration of LBP values in $R$ and $B$ channels[1] (i.e., the $L$ channels), and a darker colour in the $Y$ image (due to the lower luminance values of the lesion area). On the other hand, in the non-lesion region, the presence of LBP information (i.e. low $L$ values) flattens the $R$ and $B$ channels, while higher values in $Y$ provide a green colouring, as can be seen in the images of Fig. 3.7. This $LYL \rightarrow RGB$ correspondence is of great importance since it places the lesion and non-lesion information, respectively, in the red ($R$) and green ($G$) channels of the RGB space, which are two opposite sides in the *CIE L*a*b**s' $a^*$ colour space channel. Moreover, the $b^*$ channel provides information to the clustering algorithm, whenever the separation between red and green is not evident (i.e., when the magnitude of lesion-to-skin gradient is low). The lightness channel $L^*$ is discarded since it provides little information to the segmentation problem.

Small blobs of pink colouring may still appear over the non-lesion region since any LBP outside the lesion has the same colouring effect as inside the lesion. However, since the luminance values are lower in this region (where such blobs are located), this pink colouring effect will have a more decolourised appearance, which does not present an issue for the algorithm.

***vi)* Clustering** After the previously described colour space transformation, the resulting channels $a^*$ and $b^*$ are resized into a vector that is fed into a *k-means++* clustering algorithm (Arthur & Vassilvitskii, 2007) – represented by step *vi)* in Fig. 3.7. This algorithm was parameterised with the euclidean distance metric, with a maximum of 100 iterations and set to find 2 clusters. Additionally, the algorithm execution is replicated 3 times to reduce the chances of a bad initialisation (which are already mitigated by the *++* variant of k-means adopted). Afterwards, the best solution is selected by its lowest sums (within each cluster) of point-to-centroid (euclidean) distances.

**Final Remarks** After the clustering step, the information is reshaped back to the number of rows and columns of the input image. This translates into a pair of binary images were each pixel is assigned to either the lesion area or the skin area, thus creating two distinct and mutually exclusive image segments/masks.

[1] Note: red+blue=pink and low concentration of LBPs mean high $L$ values

Due to the nature of the clustering algorithm, the outputted masks are not always in the same order, i.e., it is not safe to assume that the $1^{st}$ segment is always the lesion area. Therefore, an automated process was also conceived to correctly select the mask that depicts the lesion area and the one related to the skin area. Since the segments match the pink and green areas of the stacked image, the variation of the mean pixel values in the image (within each segment) effectively measures the amount of pink in proportion to green that the segment contains, from the $a*$ to the $b*$ channel. That is:

1. Because the pink points present more red than blue information, when merging the values from the $b*$ negative channel (blue) and the values from the $a*$ positive channel (red), the difference between the mean values of those two layers is always positive (or zero);
2. The green is directly acquired by using the negative values in the $a*$ channel, thus its mean value is always negative and distant from 0.

This means that the image mask with the largest average value is always the image lesion area. Note that positive $b*$ values (mapped as yellow) do not exists, but are discarded nonetheless.

At this point, some morphological operations might be necessary to ensure that no small artefacts remain within the lesion perimeter or in the surrounding skin (in this case incorrectly masking the area with holes). These operations are optional since the artefacts provide insight to some structures that might be present in the lesion area, like regression structures that resemble normal skin but are part of the lesion, which are beyond the intent of this work and are therefore excluded from the mask.

### 3.2.3 Results and Discussion

This subsection details about the results obtain with the proposed LBPC method and its robustness and validation with a comprehensive study by direct comparison with other 39 segmentation algorithms using five evaluation metrics. From the 39 algorithms, 26 were implemented and categorised to seven classes, as presented in Table 2.3. The remaining 13 machine-learning segmentation algorithms, presented in Table 2.4, are later compared with the proposed method using the results reported in their original publication.

**Datasets** To enable comparisons with the proposed method, in addition to datasets $PH^2$ and Dermofit, used in the previous Section 3.1, the Atlas dataset is also employed. The use of three datasets provides robustness to the study by enlarging the representativeness of the dermoscopic images and thus improving the generalisation and validation of the results. All available images were used as test data. This is an important methodological aspect to take into account in performance evaluation studies when applied to skin lesion image datasets, because, in general, any single dataset is not well-balanced in terms of the different types of images, different technical characteristics of the acquisition setup conditions, such

as, for example, the type of dermoscope, illumination, and skin surface geometry in the lesion region (Celebi et al., 2008). Another non-uniform characteristic of the chosen image datasets is the identification of ROI (lesion region) and background skin. In some datasets the ROI segmentation is absent, while others include segmentation masks that have been used for the purpose of delineating surgical excision or clinical follow-up.

**Evaluation metrics** To assess the segmentation accuracy among the considered algorithms, this section resorts to five complementary metrics (implemented in Matlab®): BE, HD, TDR, FDR, and JI – detailed in Section 2.3.1. Only TDR and JI increase performance with higher values, the remaining metrics are preferred with lower values.

**Experiment Definition** The achieved results of each algorithm (identified in Table 2.3), using the previously mentioned metrics, are shown in Table 3.3 for the Atlas dataset, Table 3.4 for the $PH^2$ dataset, and Table 3.5 for the Dermofit dataset. Additionally, Table 3.6 shows results, in terms of JI metric, for other approaches using the $PH^2$ dataset. The method's results in Table 3.6 were gathered from different sources and compiled in reference to the most common used dataset and metric.

The performance was measured for each algorithm, in percentage, using the five metrics described in the previous section. Lower BE, HD and FDR indicate better performance, while for TDR and JI better performances are associated to higher percentages. For each performance metric (columns), the segmentation algorithm (rows) that achieves the best performance for each metric is identified by the boldface figure.

Considering the type of images and the nature of the ground-truths (manual or automatic) the JI was used as the first (and most relevant) element under analysis, establishing a similarity metric between the ground-truth and the obtained segmentation for each of the 27 methods.

**Atlas Dataset Results** In the absence of segmentation masks in the Atlas dataset, the SRM segmentation algorithm proposed by Celebi et al. (2008) was applied to all images, acting as segmentation ground-truth. Most segmentations (generated for Table 3.3) tend to be slightly inside the SRM's ground-truth. As it can be seen, LBPC, VV, and BT have the best JI performance.

In the Thresholding group both BT and UT algorithms perform well. Between them, BT appears to be more reliable as it provides better performance for all metrics, except HD and TDR – but only by 0.3 percentage points (*pp*, unit measure of the arithmetic difference between two percentages) that comes with a 1.3*pp* decrease in FDR.

In both Clustering and Active Contours group, the best performing algorithms are KMC and VV. They provide over 80% JI, while the other metrics only go up to 67%.

**Table 3.3:** Segmentation Results for Atlas dataset.

| ID | BE | HD | TDR | FDR | JI |
|---|---|---|---|---|---|
| Threshold | | | | | |
| UT | 27.3±63.5 | 519.0±163.9 | **88.0**±7.7 | 7.6±7.8 | 82.2±13.1 |
| IT | 21.3±13.8 | 490.7±175.8 | 83.0±8.0 | 7.6±4.3 | 80.3±9.1 |
| KT | 31.1±27.7 | 486.8±165.5 | 72.2±28.9 | 12.1±12.4 | 69.7±27.5 |
| LT | 24.7±12.7 | 502.3±154.3 | 78.5±9.5 | 9.0±4.8 | 76.6±9.8 |
| MT | 25.0±15.1 | 483.4±172.1 | 80.1±10.9 | 9.0±5.2 | 77.0±11.0 |
| OT | 21.4±12.0 | 489.4±175.6 | 82.3±8.4 | 7.8±4.4 | 79.9±9.1 |
| ST | 41.6±19.9 | **444.9**±186.0 | 61.4±20.3 | 14.5±6.8 | 59.6±19.0 |
| YT | 28.6±25.5 | 525.6±145.4 | 78.8±27.7 | 11.2±12.2 | 73.1±25.1 |
| BT | **17.3**±10.6 | 543.4±139.8 | 87.7±8.2 | **6.3**±4.3 | **84.0**±8.8 |
| RT | 27.1±24.3 | 500.3±160.4 | 77.1±25.4 | 10.9±12.0 | 73.9±24.0 |
| Clustering | | | | | |
| KMC | **18.9**±10.6 | 485.6±174.5 | **84.9**±6.9 | **6.9**±4.1 | **82.3**±8.1 |
| KMS | 74.8±83.4 | **400.1**±180.3 | 84.5±11.6 | 21.9±15.7 | 63.0±22.4 |
| MC | 51.3±96.3 | 443.1±179.5 | 73.4±28.9 | 15.0±18.4 | 67.3±29.4 |
| MCS | 78.4±29.8 | 430.2±159.3 | 22.4±29.5 | 30.0±15.6 | 22.2±29.3 |
| Fuzzy Methods | | | | | |
| FDEE | 58.0±25.6 | 433.2±166.8 | 42.9±26.0 | 21.9±12.7 | 42.5±25.6 |
| FC-LS | **21.7**±9.3 | **381.6**±291.1 | 80.7±8.7 | **8.1**±4.4 | 78.9±8.4 |
| FCM | 22.4±12.9 | 490.7±176.0 | **81.4**±8.5 | 8.1±4.5 | **79.0**±9.2 |
| Quantization | | | | | |
| NQ | 30.3±56.5 | 493.8±176.8 | 87.6±9.4 | 9.2±10.1 | 80.0±14.1 |
| AQ | 52.1±65.1 | 580.2±109.5 | 95.0±5.8 | 13.8±9.1 | 71.9±17.3 |
| UQ | **18.3**±9.4 | **490.1**±174.4 | 85.2±6.9 | **6.8**±4.1 | **82.7**±7.7 |
| RGB-MC | 61.5±64.1 | 532.1±151.8 | 95.6±6.6 | 16.5±8.5 | 67.9±17.4 |
| PCT-MC | 61.5±64.2 | 549.4±137.5 | **95.6**±6.5 | 16.4±8.5 | 67.9±17.4 |
| Active Contours | | | | | |
| CH | 52.7±34.8 | 404.5±203.4 | 52.4±25.2 | 19.0±11.7 | 51.7±25.3 |
| VV | **16.4**±8.8 | **340.4**±245.8 | 87.1±6.7 | **6.0**±3.8 | **84.5**±7.3 |
| LMS | 104.5±80.1 | 624.7±91.6 | **93.2**±6.2 | 30.8±9.2 | 52.8±15.1 |
| Pattern Clustering (Proposed) | | | | | |
| **LBPC** | **16.2**±8.7 | **71.6**±87.2 | **87.1**±8.4 | **6.2**±4.5 | **84.5**±8.0 |

Apart from a direct JI comparison, other metrics help to characterise what happened during the segmentation process and provide better means to compare similarly-performing algorithms. In the Fuzzy group, both FC-LS and FCM provide similar JI results with the BE, TDR, and FDR metrics differing only 0.68*pp*, 0.74*pp*, and 0.01*pp*, respectively, with FC-LS being better in the BE and FDR metrics. When taking into account the HD metric, it is clear that FC-LS might be preferable instead of the FCM since, overall, its errors are less distant from the ground-truth masks.

In the Quantization group, UQ is the best for all metrics except TDR. But, if a small FDR and HD compromise is acceptable to attain higher TDR, then AQ could be the next best option. However, AQ might not be feasible since the BE metric would go from 18.39% to 52.12%, which means that, on average, a lesion segmentation would be always off by half the lesion ground-truth area.

**Table 3.4:** Segmentation Results for $PH^2$ dataset.

| ID | BE | HD | TDR | FDR | JI |
|---|---|---|---|---|---|
| | | | Threshold | | |
| UT | 59.8±147.4 | **468.5**±150.7 | 85.9±8.0 | 10.5±13.4 | 77.1±19.1 |
| IT | 25.1±48.9 | 495.7±141.2 | 87.1±6.5 | 6.5±4.6 | 82.1±10.8 |
| KT | 29.6±30.5 | 485.0±132.3 | 86.8±19.8 | 11.7±16.3 | 76.7±19.2 |
| LT | 20.5±9.1 | 477.8±146.6 | 82.7±7.7 | 6.9±4.3 | 80.6±7.4 |
| MT | 26.2±19.1 | 503.5±122.8 | 89.2±8.4 | 7.6±4.3 | 79.2±9.3 |
| OT | **18.3**±7.2 | 497.8±138.3 | 86.0±6.9 | **6.1**±3.6 | **82.9**±6.1 |
| ST | 33.8±40.0 | 475.7±143.3 | 77.6±17.1 | 9.4±6.9 | 72.8±17.2 |
| YT | 25.0±21.2 | 488.1±147.1 | **89.8**±14.6 | 9.4±12.6 | 79.7±14.4 |
| BT | 32.0±16.6 | 495.7±125.6 | 73.4±17.8 | 11.9±10.2 | 69.9±15.1 |
| RT | 25.5±23.6 | 489.0±142.2 | 88.5±15.6 | 9.8±13.3 | 79.3±15.4 |
| | | | Clustering | | |
| KMC | **17.9**±17.9 | 476.1±155.5 | **87.9**±6.2 | **5.5**±3.3 | **84.4**±7.4 |
| KMS | 131.3±135.1 | **294.8**±147.1 | 87.6±13.7 | 29.0±14.5 | 51.9±23.4 |
| MC | 70.2±79.8 | 410.8±148.6 | 44.3±43.1 | 23.1±22.3 | 40.0±39.9 |
| MCS | 91.5±25.6 | 430.5±114.5 | 8.6±26.3 | 33.4±22.1 | 8.5±25.7 |
| | | | Fuzzy Methods | | |
| FDEE | 45.7±27.0 | 414.6±145.9 | 55.8±27.6 | 14.5±10.6 | 55.1±26.9 |
| FC-LS | 34.8±118.0 | **285.8**±251.7 | 83.4±14.1 | 7.2±8.8 | 81.0±13.9 |
| FCM | **19.8**±11.7 | 496.4±131.5 | **85.5**±7.2 | **6.4**±3.8 | **82.0**±7.3 |
| | | | Quantization | | |
| NQ | 73.5±157.1 | 490.5±132.9 | 88.7±16.9 | 14.5±17.4 | 72.9±23.5 |
| AQ | 109.3±135.7 | 396.4±176.2 | **94.5**±7.4 | 21.3±12.4 | 59.8±22.7 |
| UQ | **16.3**±6.4 | 493.9±139.8 | 87.5±6.6 | **5.4**±3.4 | **84.7**±5.8 |
| RGB-MC | 115.4±125.7 | **395.5**±168.5 | 93.4±10.9 | 23.6±10.7 | 56.6±21.1 |
| PCT-MC | 114.1±125.4 | 397.5±172.3 | 94.1±10.7 | 23.3±10.6 | 57.1±21.1 |
| | | | Active Contours | | |
| CH | 43.8±60.8 | **417.4**±197.4 | 73.9±22.1 | 12.3±13.8 | 68.5±21.2 |
| VV | **26.9**±28.8 | 459.0±183.4 | 82.6±16.2 | **7.5**±4.2 | **77.2**±16.4 |
| LM | 168.0±161.0 | 558.5±67.6 | **94.3**±6.7 | 36.6±12.2 | 46.5±19.8 |
| | | | Merging Threshold | | |
| SRM | **123.7**±173.4 | **244.8**±173.7 | **68.6**±36.2 | **28.3**±22.9 | **48.6**±33.8 |
| | | | Pattern Clustering (Proposed) | | |
| **LBPC** | **14.1**±4.7 | **58.2**±39.3 | **88.4**±6.1 | **5.2**±3.7 | **86.3**±4.6 |

Finally, the proposed LBPC algorithm provides the best results, in terms of JI and BE, and it is second only to VV, in terms of FDR by 0.2*pp*.

**$PH^2$ Dataset Results** Result trends for the $PH^2$ dataset (Table 3.4) are similar to those achieved with the Atlas Dataset, but now they are obtained in respect to a ground-truth provided by the dataset.

In the Thresholding group, looking at JI, both OT and IT perform similarly. However, IT is preferable if a gain of 1.1*pp* in TDR is outperformed by the (error) gain of 0.4*pp* FDR.

In the Clustering, Quantization and Active Contours groups, KMC, UQ, and VV outperform all algorithms in their groups, presenting higher JI.

**Table 3.5:** Segmentation Results for Dermofit dataset.

| ID | BE | HD | TDR | FDR | JI |
|---|---|---|---|---|---|
| | | | Threshold | | |
| UT | 83.9±187.9 | 366.3±306.6 | **81.9**±11.8 | 12.0±13.7 | 70.2±22.4 |
| IT | 65.3±144.7 | 343.3±305.4 | 79.2±11.7 | 11.0±11.1 | 69.8±19.6 |
| KT | 48.3±118.7 | 394.1±290.8 | 75.6±22.2 | 11.8±16.1 | 69.1±22.8 |
| LT | 57.8±126.8 | **327.7**±297.0 | 75.3±12.2 | 10.8±10.1 | 68.5±17.7 |
| MT | 53.1±96.0 | 368.5±307.8 | 80.5±12.0 | 10.6±9.5 | 69.5±17.9 |
| OT | 48.9±103.5 | 331.4±297.3 | 76.9±12.0 | 10.0±9.7 | 70.6±17.1 |
| ST | 65.1±83.1 | 350.6±267.2 | 57.0±21.4 | 15.0±8.8 | 51.3±19.6 |
| YT | 42.4±116.0 | 443.4±291.2 | 80.2±21.3 | 9.8±14.7 | **73.2**±22.1 |
| BT | **35.0**±30.1 | 370.8±307.5 | 72.1±18.6 | **9.4**±7.0 | 68.6±18.4 |
| RT | 48.5±131.6 | 386.5±292.8 | 78.8±19.9 | 11.0±16.0 | 71.7±21.6 |
| | | | Clustering | | |
| KMC | 64.2±136.7 | 349.8±297.0 | 78.6±12.1 | **11.1**±11.8 | **69.5**±19.8 |
| KMS | 212.2±188.6 | 348.5±109.0 | 66.7±15.3 | 43.4±9.2 | 30.7±15.5 |
| MC | 232.4±355.8 | 290.1±212.6 | **89.7**±15.3 | 40.6±34.7 | 51.7±29.3 |
| MCS | **54.5**±35.2 | **286.6**±212.8 | 49.7±27.4 | 16.2±11.5 | 49.0±27.2 |
| | | | Fuzzy Methods | | |
| FDEE | 115.7±133.4 | 372.2±189.4 | 41.6±37.3 | 29.8±21.7 | 27.6±24.0 |
| FC-LS | 238.9±308.8 | **285.3**±204.2 | 50.4±31.3 | 49.9±41.6 | 38.9±35.1 |
| FCM | **60.7**±127.8 | 343.7±299.1 | **77.2**±12.2 | **10.8**±10.1 | **69.0**±18.8 |
| | | | Quantization | | |
| NQ | 177.6±290.3 | 439.8±296.2 | 86.5±17.3 | 25.4±24.1 | 57.5±29.4 |
| AQ | 154.8±243.9 | 627.0±222.9 | 91.3±13.6 | 23.5±18.3 | 56.2±24.7 |
| UQ | **52.6**±108.4 | **321.9**±298.0 | 77.9±12.7 | **10.7**±12.7 | **70.6**±18.5 |
| RGB-MC | 146.5±191.2 | 665.4±215.5 | 95.1±9.0 | 23.5±11.7 | 54.2±22.5 |
| PCT-MC | 145.7±190.4 | 664.1±222.4 | **95.4**±8.2 | 23.4±11.5 | 54.4±22.3 |
| | | | Active Contours | | |
| CH | 244.6±365.5 | 339.3±236.5 | 69.5±30.1 | 39.9±33.1 | 44.7±30.8 |
| VV | 240.6±398.1 | **295.4**±256.5 | **77.7**±24.5 | **33.9**±33.0 | **52.1**±31.0 |
| LMS | **213.0**±86.8 | 369.8±88.0 | 0.1±0.5 | 53.5±12.4 | 0.0±0.3 |
| | | | Merging Threshold | | |
| SRM | **55.0**±65.9 | **167.1**±150.8 | **52.0**±34.2 | **15.0**±12.6 | **50.2**±33.1 |
| | | | Pattern Clustering (Proposed) | | |
| **LBPC** | **29.6**±41.1 | **110.4**±105.1 | **78.7**±21.6 | **7.7**±7.5 | **74.8**±21.4 |

In the Fuzzy group the best algorithm in terms of detection rate is the FCM with 85.54% TDR and 6.44% FDR, however it might not be the best option if its higher erratic behaviour (HD), in comparison to FC-LS, is undesirable. The FC-LS method provides almost 40 less HD than the FCM algorithm, thus yielding smoother segmentation that better mimic the ground-truth at the cost of 2.13*pp* TDR and 0.82*pp* FDR.

Finally, LBPC provides even better results than before, while the SRM provides almost the worst results in Table 3.4.

**Dermofit Dataset Results** The Dermofit dataset results (Table 3.5) are consistent with those obtained for Atlas and $PH^2$.

In the Thresholding group no algorithm stands out as optimal. Most results for the JI metric are within a close range. YT emerges with the highest value and best balance between losses and gains in the remaining metrics in comparison to the remaining peers.

In the Clustering group, KMC leads with the highest JI value. While not having the best TDR value, by 11.1*pp*, it provides the best balance with a 29.5*pp* less FDR in relation to MC.

Likewise, for the Fuzzy group, FCM shows the best values for all metrics except HD.

Similarly, in the Quantization and Active Contours groups, both UQ and VV stand out. They are only outperformed by other algorithms regarding their lower TDR and BE, respectively. But, other algorithms present worse values for all the other metrics, which are disproportional to a the small gain they provide.

Finally, the SRM algorithm provided average results and the LBPC algorithm has the best performance balance in comparison to all others.

**Across Datasets Analysis** Overall, the top 3 algorithms across the datasets (sorted by an average JI performance weighted by the number of images in each dataset) are: LBPC, UQ, and KMC; with 81.7%, 76.3%, and 75.5%, respectively. Generally speaking, Thresholding methods tend to provide results above average, while the remaining groups have a less linear behaviour. For dermoscopic datasets, Clustering, Fuzzy, and Quantization algorithms provide average results, while Active Contours methods tend to perform below average (except VV that is above average). For the macro images dataset (Dermofit), only Thresholding and a short number of methods provide results above average. Most algorithms tend to discard small portions of the ground-truth masks and include surrounding skin. Finally, as previously described, the proposed LBPC provides the best results across the different datasets.

**Recent Works Comparison** As discussed in the previous section, for the considered datasets, the LBPC algorithm was found to be the one with the best JI metric results in comparison to the revised classical algorithm implementations. The study carried out went beyond those approaches and, as shown in Table 3.6, the performance of LBPC was also compared to 13 more recent algorithms based on machine learning techniques (ordered by year of publication) for the $PH^2$ dataset (as in the 13 published algorithm works). The presented JI metric values were gathered from the published papers that describe these methods and from Bi et al. (2017); where now, the proposed method (LBPC) is only second to SWSDB,

**Table 3.6:** Recent Segmentation Results for $PH^2$ dataset.

| ID | FCN | SSLS | SCDRR | MSCA | mFCN | JCLMM | CDNN | CDNNE | KL-LS | DermoNet | PSO-DEN | SWSDB | DCL-PSI | **LC** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| JI | 82.15 | 68.16 | 76.00 | 72.33 | 83.99 | 70.72† | 84.33† | 76.5 | 71.54† | 85.3 | 77.75† | **89.04**† | 85.90 | 86.34 |

† Jaccard index value generated from published Sørensen–Dice metric.

and even in this case for a margin smaller than 3*pp*. This can be explained by the fact that LBPC algorithm is more directed and guided towards the target domain, whereas machine learning algorithms, that depend on ground-truths to learn, need large amounts of data and still might not converge to the desired results.

Although other methods such as Bayraktar et al. (2019) and Kéchichian et al. (2014) exist and could be compared with our proposal, their published results are based on a different set of ground-truths that are incompatible with those of Table 3.6, thus these comparisons were considered out of the scope of this work.

**Summary** Globally, excluding the LBPC algorithm, Thresholding methods perform the best. Nevertheless, having defined JI as the most discriminant performance metric, the LBPC algorithm provides proper results across the datasets. In the presence of more recent and complex algorithms, LBPC slightly behind SWSDB.

### 3.2.4 Conclusions

Automated melanoma identification is crucial to help both dermatologists and computer expert systems. An accurate skin lesion segmentation can improve the initial assessment and help computer vision techniques to provide insights and enhance the users abilities. Therefore, image segmentation of skin lesion plays a major role for forthcoming algorithms. This work provides a unified comparison between several segmentation methods and datasets in order to present a better understanding of some of the currently available techniques. The study compares the proposed method with a specific set of 26 conventional segmentation algorithms, grouped by the type of their methodology, tested for 3 different datasets.

The comparison results showed that different segmentation methods of a same type tend to behave similarly. Even so, their behaviour highly depends on the datasets' image characteristics. In spite of that, some conclusions can still be drawn from this study. The Thresholding type of algorithms seems to be the most constant in terms of expected results, however the results achieved are not that accurate in the general context. In contrast, Active Contours algorithms provide, on average, the worst FDR measurements, since they tend to stick close to the lesion borders, thus typically presenting more inner segmentations. Quantization, as a group, is on average one of the worst in any scenario, however the quantization UQ algorithm actually performs above average. The sole Merging Threshold algorithm, which is reported to have a good performance for the Atlas dataset, does not stand out across the selected metrics on the remaining two datasets. Finally, the proposed Local Binary Pattern based algorithm appears as the leading method when compared to the existing ground-truths. As future research, adaptions of the proposed method will be investigated on other datasets, as well as, potential clinical applications.

Across the observed results, the proposed LBPC method is always superior in terms of BE, HD, and JI. It is also the best performer in the FDR metric for both the $PH^2$ and Dermofit datasets. The best performing algorithms for the TDR metric usually create a segmentation that is too large, covering not only the lesion area but also marking other areas as lesion. Despite their higher TDR, they perform worse in the remaining metrics.

The proposed unsupervised method is tested with three certified datasets and comprises three main phases: LBP image enhancement, space transformation, and binary clustering. It has performed better than the 26 classic methods of segmentation. Additionally, it was also compared against 13 recent skin lesion segmentation approaches, published between 2015 and 2019, for the $PH^2$ dataset. The proposed method outperforms most of the other algorithms, while is computationally less expensive. It behaves independently from the dataset (given its unsupervised setting) but it is sensitive to hairs and possible artefacts present in the image.

Overall, apart from presenting a new segmentation method capable of outperforming the current state-of-the-art, this work provides insightful information about the behaviour and performance of different image segmentation algorithms.

## 3.3 Classification using Transfer Learning

Traditional ANN have been investigated in the past for skin lesion classification. Nowadays, their performance is useful to assist in medical diagnosis and decision processes, namely when performing transfer learning on pre-trained networks.

This section focuses on studying the performance of skin cancer detection using highly-accurate networks, which were developed for ImageNet. To this end, the ISIC dataset (Collaboration, 2017) is selected as the collection of skin lesion images.

This section is organised as follows: Section 3.3.1 presents the background that is relevant for the proposed experiment. Section 3.3.2 presents the transfer learning approach used in this research study and the selected networks with pre-trained weights from ImageNet. Then, in Section 3.3.3, the classification performance is evaluated and discussed. Finally, Section 3.3.4 draws some conclusions.

### 3.3.1 Relevant Background

Recent advances in visual recognition led to the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Deng et al., 2009; Russakovsky et al., 2015), which uses a dataset comprising more than 14 million images (of which 1 million have bounding box annotations) that can be

divided into 1000 different labels – manually validated through crowd-sourcing. The ImageNet Challenge is currently considered to be one of the most important initiatives. Therefore its dataset has became a benchmark standard for large-scale object recognition, i.e., image classification, single-object location, and object detection. Due to its competition-based approach, many authors are constantly improving their image classification/recognition algorithms every year. This has led to an exponential growth of related research and significant advances in state-of-the-art techniques (Russakovsky et al., 2015).

### 3.3.2 Proposed TL Classification Approach

The proposed approach follows a processing pipeline from the input image data to the output classification results. Firstly, before entering the network, a pre-processing stage is responsible for performing data augmentation and then image resizing, to match the network intake. Secondly, these data enters in the pre-trained network whose output is fed to the final classifier. Several classifiers are studied in this work. Different alternatives are separately trained, resorting to both original data and augmented data with 20% random information holdout for later evaluation of the trained network.

**Architectures** In ILSVRC history there are several pre-trained networks, already capable of image classification over 1000 different categories. This work elects 5 of the most frequently used networks, which have shown to be able to adapt to other identification and classification problems. These networks are: Alexnet (Krizhevsky et al., 2012), pioneering networking comprising 25 layers (winner of the 2012 ILSVRC); VGG16 and VGG19 Net (Simonyan & Zisserman, 2014), reinforced the notion that convolutional neural networks must have layers in depth, such that visual data present a hierarchical representation; GoogLeNet (Szegedy et al., 2015), has the Inception module that deviates from the standard sequential layer-stacking approach (winner in 2014); and ResNet50 (He et al., 2016), presents an innovative way of solving the vanishing gradient problem, it comprises 177 layers (winner in 2015).

**Pre-processing** To increase accuracy, data augmentation is performed by using a limited set of random transformations (Lemley et al., 2017). In this work the following transformations were selected: *Intensity Values Adjustment*: increases the contrast of the image; *Contrast-Limited Adaptive Histogram Equalisation*: enhances the contrast of a given grayscale image by transforming the values, such that its distribution matches a uniform/flat histogram (256 bins); *Random Brightness*: induces brightness variation to the image; *Random Edge-Aware Local Contrast*: enhances or flattens the image local contrasts; *Random Sharpness*: sharpens the image using the unsharp-masking method; *PCA Colour Jitter*: modifies the intensities of the RGB channels in the image, according to the PCA transformation; *Random Affine Transformations*: operation between affine spaces that preserves points, straight lines and

planes. As a final note, the augmentation strategies are not all used at the same time. The *PCA Colour Jitter* and *Random Affine Transformations* are always used at the end of the augmentation step, but the remaining operators are only randomly applied with a 10% change (each). After this stage, each image is augmented 200 times, thus effectively making the dataset 200 times larger.

After a possible augmentation step, and before entering the network, all input data (images) is resized to fit the network intake. Apart from Alex-Net, which receives a 277x277 (pixel) RGB image, all other networks accept a 224x244 (pixel) RGB image. Therefore, as a final step before entering the network, the images are resized to their smallest dimension (maintaining aspect ratio) and centre-cropped, to remove the outer border in excess (if any).

**Learning Strategy** As mentioned before, the overall architecture includes ImageNet networks and a transfer learning scheme for feature extraction using alternative classifiers. Since the selected pre-trained architectures already provide highly accurate predictions in the ImageNet challenge, it is assumed that they are also able to extract a great variety of abstract knowledge/features from the given images containing skin lesions. In this transfer learning strategy, the output of the last convolutional layer in the pre-trained ImageNet network is connected to several alternative classifiers. The classifiers used in this work are: the SVM classifier, the K-Nearest Neighbours, the Tree classifier, a Linear classifier, and a NaiveBayes classifier.

### 3.3.3 Results and Discussion

For the study of the performance of skin cancer detection using highly-accurate networks, the ISIC dataset (Collaboration, 2017) is selected as the collection of skin lesion images. This dataset contains a total of 3438 images that can be divided into: 2380 benign and 1058 malignant lesions. These malignant lesions are classified as melanoma, basal cell carcinomas, and squamous cell carcinoma, while the remaining ones are benign. Such classification was obtained from an unspecified number of skin cancer experts.

Using the ImageNet networks as feature extractor on the original 3438 images, while holding out 20% of this data for later testing, the network knowledge provides an average ACC of 61% on the testing data, while the ACC obtained in training data is 87% on average. The overall results are shown in Table 3.7, where it can be observed that the best performing classifier is the KNN with an average ACC of 72% on unseen test data across the different networks and 100% on the training data. Still regarding the training data performance, the SVM and the Tree classifiers achieve accuracies of 99% and 98%, respectively. However, only 62% and 61% ACC is obtained on unseen test data.

**Table 3.7:** Transfer Learning Test Results without using Augmented Data in training.

| Model | AlexNet | | | | VGG16 | | | | VGG19 | | | | GoogleNet | | | | ResNet50 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC |
| SVM | 30.9 | 99.5 | 0.4 | 50.0 | 34.6 | 97.6 | 6.7 | 52.2 | 46.7 | 69.2 | 36.8 | 53.0 | 30.7 | 100 | 0.0 | 50.0 | 30.7 | 100 | 0.0 | 50.0 |
| KNN | 74.5 | 61.1 | 80.5 | 70.8 | 67.7 | 56.9 | 72.5 | 64.7 | 71.3 | 60.2 | 76.3 | 68.3 | 73.8 | 57.3 | 81.1 | 69.2 | 72.8 | 60.2 | 78.4 | 69.3 |
| Tree | 68.7 | 46.9 | 78.4 | 62.7 | 64.2 | 40.8 | 74.6 | 57.7 | 64.0 | 40.3 | 74.6 | 57.5 | 67.7 | 51.2 | 75.0 | 63.1 | 61.3 | 49.3 | 66.6 | 58.0 |
| Linear | 70.3 | 4.3 | 99.6 | 52.0 | 69.4 | 55.9 | 75.4 | 65.7 | 55.5 | 86.3 | 41.8 | 64.1 | 73.9 | 52.1 | 83.6 | 67.9 | 75.8 | 47.9 | 88.2 | 68.1 |
| NaiveBayes | 64.9 | 73.5 | 61.1 | 67.3 | 62.6 | 66.4 | 60.9 | 63.7 | 64.8 | 64.5 | 64.9 | 64.7 | 64.9 | 73.9 | 60.9 | 67.4 | 72.5 | 55.9 | 79.8 | 67.9 |

When data augmentation is used, the performances increase by 9*pp* on the test-set and decrease 12*pp* ACC on the training-set. Table 3.8 is presented for comparison with the previous results. In this case the training-set only comprises augmented images, while the test-set is the same as before. It is observed that image augmentation provides some improvement to the classification results. Despite the small improvement of the KNN classifier, which only gains 0.6*pp* ACC on test data, the SVM classifier more than doubles its performance. Taking into account the training results (not shown here), this increase in performance is justified by the reduction of overfitting resulting from data augmentation.

**Table 3.8:** Transfer Learning Test Results using Augmented Data in the training.

| Model | AlexNet | | | | VGG16 | | | | VGG19 | | | | GoogleNet | | | | ResNet50 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC |
| SVM | 72.6 | 54.5 | 80.7 | 67.6 | 71.9 | 58.8 | 77.7 | 68.3 | 71.9 | 58.8 | 77.7 | 68.3 | 71.6 | 60.7 | 76.5 | 68.6 | 72.8 | 53.1 | 81.5 | 67,3 |
| KNN | 75.1 | 62.6 | 80.7 | 71.7 | 70.3 | 52.1 | 78.4 | 65.3 | 70.3 | 52.1 | 78.4 | 65.3 | 71.8 | 55.5 | 79.0 | 67.3 | 75.4 | 64.9 | 80.0 | 72,5 |
| Tree | 65.5 | 48.8 | 72.9 | 60.9 | 68.0 | 46.9 | 77.3 | 62.1 | 68.0 | 46.9 | 77.3 | 62.1 | 67.5 | 46.0 | 77.1 | 61.6 | 69.3 | 53.1 | 76.5 | 64,8 |
| Linear | 78.0 | 52.1 | 89.5 | 70.8 | 60.8 | 80.1 | 52.3 | 66.2 | 60.8 | 80.1 | 52.3 | 66.2 | 69.4 | 69.2 | 69.5 | 69.4 | 78.5 | 43.1 | 94.1 | 68,6 |
| NaiveBayes | 67.1 | 66.8 | 67.2 | 67.0 | 69.3 | 0.0 | 100 | 50.0 | 69.3 | 0.0 | 100 | 50.0 | 62.6 | 70.6 | 59.0 | 64.8 | 67.7 | 72.5 | 65.5 | 69,0 |

### 3.3.4 Conclusions

ILSVRC winning networks achieve an ACC greater than 95% in the ImageNet dataset. However, when adapted to classify skin lesions, their performance drops to modest results, even when data augmentation is used. This work performed transfer learning to classify skin lesions as malignant or benign using five cornerstone neural network architectures, which have been proven to produce high results on other domains. The results demonstrate that there is significant room for further research, using highly accurate networks and transfer learning for specific classification in the field of medical imaging. In particular, it is necessary to investigate how to improve transfer learning performance using networks trained on completely different domains.

Since the results of this experiment did not uphold the baseline expectations, following experiments focused on hand-crafted features and validation of previous findings.

## 3.4 Classification using 2D Border-Line Features

Machine learning algorithms are progressively assuming an important role as a computational tool to support clinical diagnosis, namely in the classification of PSL. The current classification methods commonly rely on features derived from shape, colour, or texture obtained after image segmentation, but these do not always guarantee the best results. When the new features are combined with the classical ones, the experimental results show higher accuracy, which positively impacts the overall performance of the classification algorithms. To improve the classification accuracy, this work proposes to further exploit the border-line characteristics of the lesion segmentation mask. In the proposed method, these border-line features are used together with the conventional ones to enhance the performance of skin lesion classification algorithms.

The main contribution in this section is to demonstrate the relevance of the proposed features (extracted from the segmentation border-line information) to improve a classification algorithm, in addition to other commonly used features (Jafari et al., 2016). Hence, this work does not aim to validate the segmentation or classification algorithms, but the information gain achieved with the proposed features. The proposed method uses two existing segmentation techniques: Gradient-based Histogram Thresholding (GHT, Pereira et al., 2019b), detailed in Section 3.1, and a variant of the recent Local Binary Patterns Clustering (LBPC, Pereira et al., 2019a, 2020a), detailed in Section 3.2. The border-line features are extracted and used as input for automatic classification of melanocytic lesions using ML algorithms. To this end, only images of melanoma and nevus are used. As previously mentioned, melanoma is the most aggressive form of skin cancer and one-third of all melanomas arise from pre-existing nevus, thus detection and removal of such nevus is of utmost importance in the prevention of melanoma. In summary, this section exploits the additional insight that, based on statistical features, border-line information provides relevant information to enhance discrimination of skin lesions.

The remainder of section is organised as follows: Section 3.4.1 presents the literature involved in this work. Section 3.4.2 introduces the proposed approach, detailing the segmentation techniques, feature extraction and the pigmented skin lesion classification. Section 3.4.3 presents and discusses the results with statistical validation information, and Section 3.4.5 highlights the conclusions and future work.

### 3.4.1 Relevant Background

Only recently, delineation information of skin lesions border-line has emerged as an input feature in ML algorithms. Based on a detailed segmentation border, several feature characteristics of its perimeter may be determined. For such purpose, some research work has been carried out in this topic. In Linsangan et al. (2018), using a clustering algorithm, the authors ex-

tracted features related to the perimeter border from 90 images and classified them as Benign, Malignant, and Unknown. Similarly, in Mane & Shinde (2018) the authors also extracted the perimeter information, as well as colour and texture features. These emerging features provide new dimensions to the solution and have potential to be added directly to most existing approaches. One of such works is presented in Hameed et al. (2018), where several features are benchmarked with Matlab classifiers (present in the classifier app), attaining an average ACC of 87%. Other implementations, like those based on Deep Learning, namely Namozov & Cho (2018); Chen et al. (2019), already obtain state of the art results. However, the inclusion of manually extracted features is more challenging in these approaches. Nevertheless, most neural network applications, like Majumder & Ullah (2018), are easily adaptable since they already behave like a feature classifier. Currently, a feature-based descriptor for skin lesions that mainly includes some types of border-related features was also introduced in Mahdiraji et al. (2018).

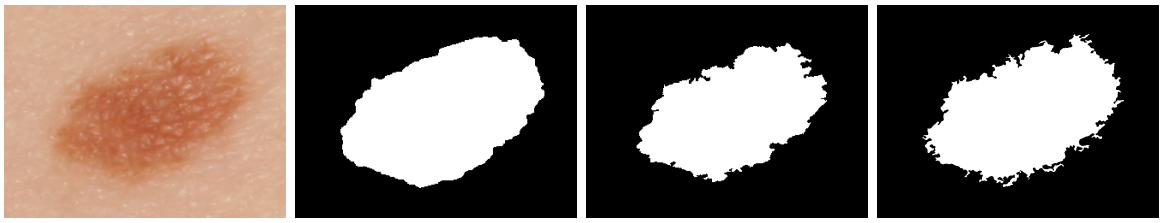
In spite of the fact that such approaches provide state-of-the-art results, they do not utilise the same datasets or classifiers, making the comparison of their results a difficult task. Therefore this comparison is not performed in the scope of this work.

### 3.4.2 Proposed 2D Border-Line Classification Approach

This section presents the method's pipeline, which comprises three steps: Segmentation, Feature Extraction, and Classification. Firstly, the lesion image is segmented with the GHT and the LBPC algorithms from Section 3.1 and Section 3.2, respectively. Then common border-line features of the binary segmentation mask are extracted and used by the lesion classifier in the third and final step.

**Segmentation** While GHT exploits luminance intensity variations, LBPC is more sensitive to texture patterns, therefore in both methods the luminance (grayscale) image ($Y$) is used, being obtained by means of a weighted sum of R, G, and B channels, as defined in ITU-R (2011) with four decimal places. Particularly for this study, LBPC was changed to not use the *CIE L*a*b** colour space. Instead, the transformed LBP image is subtracted from the grayscale image and the resulting pixel values are input to the *k-means++* clustering algorithm, as it was found to lead to better performances at this study's classification step.

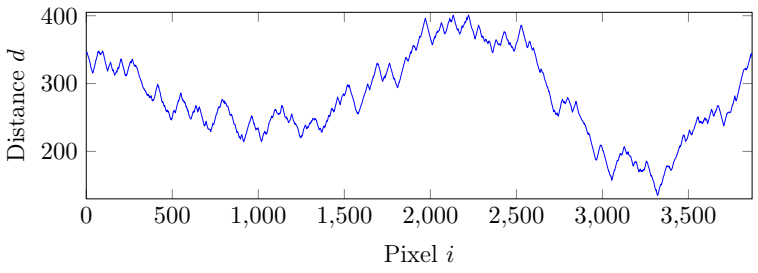
The resulting segmentation output mask obtained with these two techniques can be observed in Fig. 3.9 for image $B355b$ of the Dermofit dataset (Ballerini et al., 2013). As can be seen, the segmentation from both the GHT and LBPC algorithms provides much higher detail on the lesion borders than that of the dataset Ground-Truth (in Fig. 3.9b).



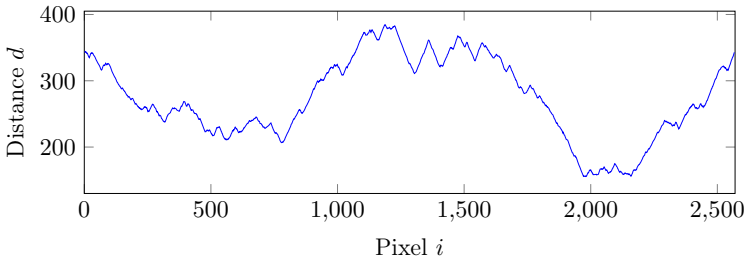
**(a)** Input Image **(b)** Ground-Truth mask **(c)** GHT output mask **(d)** LBPC output mask

**Figure 3.9:** Segmentation results for *B355b* image of the Dermofit dataset.

**Feature Extraction** In order to extract features from the proposed segmentation masks, the detailed border-lines are reshaped from their rounded lesion-shape to an unfolded line, resulting in the lines shown in Fig. 3.10a and Fig. 3.10b. The line unfolding is carried out by firstly calculating the centre of mass of the segmented region. Then, the euclidean distance $d$ (in pixels) from each pixel to the centre of mass is represented by $d(i)$ and, from this representation, the new line-segment is obtained. This unfolded line maintains all the original information, except the lesion shape. As can be observed from Fig. 3.10a and Fig. 3.10b, the lines segments originated from both algorithms have different sizes. This is due to the segmentation boundary generated by each corresponding algorithm. Although the algorithms provide similar shaped-segmentations, GHT displays a smoother curve than LBPC. Hence the LBPC segmentation is intrinsically larger (in terms of perimeter pixels) than GHT. This is further evidenced observing Fig. 3.9.




**(a)** Detail of LBPC border-line




**(b)** Detail of GHT border-line

**Figure 3.10:** Border-lines extracted from *B355b* image of Dermofit dataset.

Based on this representation, a new set of features were extracted from the unfolded border-line (as can be seen in Fig. 3.10a), namely: root-mean-square level *(F1)*; average $d$ value *(F2)*; height of main peak *(F3)* and height and position of second peak of an autocorrelation sequence calculation *(F4-5)*; magnitude of the highest peak of each of the first 6 bins of a Discrete Fourier Transform spectrum (DFT) using 4096 points, where the sampling resolu-

tion is $2\pi/4096$ rad/sample (dividing it in 32 equal-size bins) *(F6-11)*; frequency component corresponding to the 6 points of the previous features *(F12-17)*; sum of values of 5 equal-length segments produced by splitting its periodogram power spectral density (PSD) (Auger & Flandrin, 1995) *(F18-22)*. The number of peaks/segments (*F6-22*) was optimised using correlation analysis.

**Classification** As previously mentioned, the main research question addressed in this section is to verify whether segmentation border details and the type of lesion might be somehow correlated. This is done in a nevus versus melanoma setting. It is known that the melanoma is the most aggressive form of skin cancer and one-third of all melanomas arise from pre-existing nevi. Thus, detection and removal of such nevi is of utmost importance in the prevention of melanoma. If such hypothesis is true, the use of border-line features might prove to be useful in providing additional discriminatory information that will help to improve the classification accuracy of skin lesions. To test and validate the raised hypothesis, three classifiers were used: two of them are based on a linear SVM of similar parameters, while the third implements a Feedforward Neural Network (FNN) for classification. Deep Learning classifiers were not selected for this study due to the selected segmentation algorithms variable length outputs and the datasets' size constraints. The experiments were made in a MSI GT683DXR-423US laptop, which provides an Intel® Core™i7-2670QM CPU @ 2.20GHz with 8GB of RAM.

The first classifier, namely the SMO, employs an SVM classifier using Sequential Minimal Optimization. This classifier was proposed in Jafari et al. (2016) for skin lesions classification and implements a robust supervised learning method with a linear kernel function that is solved iteratively through the sequential minimal optimization. The classifier, imported from Weka 3.8.2, is employed to enable comparison with Jafari et al. (2016). In this algorithm, the SVM problem is broken into a series of smaller sub-problems, which are solved analytically (Platt, 1998). For this method default literature parameters were used: complexity constant of 0.5 and epsilon of $1 \times 10^{-7}$.

The second classifier, namely the ISDA, also employs an SVM classifier, however, instead of solving the problem with the sequential minimal optimization, as in the SMO, this version uses the Iterative Single Data Algorithm proposed in Kecman et al. (2005b). For this classifier, an existing implementation present in Matlab™ R2018b was used. Unlike SMO, ISDA solves a series of one-point minimisation that does not respect the linear constraint and does not explicitly include the bias term in the model. The ISDA implementation uses the same parameters, as in the SMO.

The third classifier, referred to as FFN, is a Feed Forward Network (present in Matlab® R2018b *patternnet* function) that was implemented based on a common *rule of thumb*, which states that the number of neurons $n$ in a network should be determined taking into consideration the number of samples, features (inputs) and possible classifications (outputs), expressed by $n = (\#\text{sample} * (\#\text{inputs} + \#\text{output}))/w$ where, the weight $w$ was set to 2, halving the

result, in order to force lower overfitting probability, as it limits the networks' number of degrees of freedom. In this network, the traditional sigmoid activation function (Cybenko, 1989) was employed. This network was trained using Scaled Conjugate Gradient Backpropagation (Møller, 1993) with cross-entropy as the network performance measurement and no normalisation or regularisation for simplicity. The FFN classifier was included in this experiment because it was shown to be a universal approximator and could, thus, provide better results (Csáji, 2001). The following default literature parameters were used in this method: $5 \times 10^{-5}$ for derivative approximation (sigma), $5 \times 10^{-7}$ for the indefiniteness of the Hessian (lambda), a minimum performance gradient of $1 \times 10^{-6}$, and maximum 6 validation fails.

For the tests, the SVM classifiers were trained using 90% of the available data and tested on the remaining, unseen, 10% of the data. For the FNN, the network was trained on 70% of the data, validated on untrained 20% (to prevent overfitting), and later tested using the remaining 10% of the data. The results obtained in the different classifiers are presented in terms of the average of all tests that have been repeated 10 times using 10-fold Cross-Validation (CV) – 100 executions. Training and test proportions of 70%-30% and 50%-50% were also considered, presenting similar results.

### 3.4.3 Results and Discussion

The obtained results cover two datasets, namely the MED-NODE dataset (Giotis et al., 2015) and the Dermofit dataset (Ballerini et al., 2013), since they provide different acquisition methods and constraints. Particularly for the Dermofit dataset, a pre-processing step occurred with the aim of removing hairs from the images by using the algorithm described in Koehoorn et al. (2015). Although the lesions' diagnostics in this later dataset span across ten different classes, only melanoma and nevus are of interest for this research. Using this criterion, 407 images were selected, obtaining an unbalanced setting of 331 nevi and 76 melanomas.

For each dataset, the image data was evaluated by using the three classifiers described in Section 3.4.2, classification paragraph. In each case, two feature sets were employed. Firstly, 10 features (*F23-F32*) proposed in Jafari et al. (2016) are used as input to the classifiers (five of which assess the lesions' asymmetry aspects, one assesses border condition and four consider lesions' colour attributes). Later, in order to assess the contribution of the detailed border-line information to the classifiers performance, the remaining 22 features (*F1-F22*) described in Section 3.4.2, feature extraction paragraph, are used as input to the classifiers, thus resulting in a total of 32 features. The results obtained in these assessments are expressed in terms of percentage of classification accuracy (ACC), specificity (SPE), and sensitivity (SEN), similarly to Situ et al. (2008); Smith et al. (2011); Satheesha et al. (2017); Pathan et al. (2018); Hu et al. (2019); Pereira et al. (2020b). This experiment, as described, is estimated to take two and half hours to execute – including segmentation and feature extraction of all images in the dataset, and later classification of this data using the four different sets of features for the three classifiers.

**Table 3.9:** Border-Line Results for the MED-NODE dataset.

| Seg. | Ft. (#) | SVM-SMO | | | | SVM-ISDA | | | | FFN | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC |
| GHT | F23-F32 (10) | 73±1.5 | 45±3.7 | 92±3.0 | 69±1.5 | 76±1.2 | 66±2.4 | 83±1.4 | 74±1.9 | **76**±1.9 | 63±4.6 | 84±2.1 | 73±2.3 |
| | F1-F32 (32) | **74**±1.2 | 56±6.4 | 86±6.2 | 71±0.8 | **78**±2.0 | 66±2.4 | 86±1.3 | 76±1.8 | **76**±2.4 | 63±4.7 | 84±2.7 | 73±1.4 |
| LBPC | F23-F32 (10) | 75±1.2 | 49±3.0 | 93±2.0 | 71±1.4 | 77±1.3 | 69±2.2 | 83±0.9 | 76±1.5 | 75±1.7 | 64±4.1 | 83±1.7 | 73±2.8 |
| | F1-F32 (32) | **78**±1.3 | 58±5.6 | 91±3.4 | 74±3,2 | **79**±1.5 | 65±2.7 | 88±1.1 | 76±0.9 | **77**±1.9 | 66±5.8 | 86±2.2 | 76±2.8 |

**MED-NODE dataset** The experimental results for the MED-NODE dataset can be seen in Table 3.9.

When using the GHT segmentation, adding the proposed border-line features led to limited improvements of 1*pp* and 2*pp* on SMO and ISDA, respectively; and no improvements for the FFN classifier. With these results the best classification results correspond to ISDA, which achieves 78% ACC, followed by the FFN with 76% ACC. In this scenario, the main problem faced by classification algorithms is to wrongly classify the melanoma samples as nevus, which leads to poor SPE results. It is, however, worth noting that with SMO the inclusion of the proposed border-line features have substantially increased the SPE by 11*pp*.

The obtained results show that using border-line features extracted from LBPC-based segmentations consistently leads to better results in all classification methods. This is likely to be associated with the dense local texture information provided by LBPs localised detail which, in this case, led to improvements of 3*pp* for the SMO and 2*pp* for the remaining classifiers. Moreover, the SPE also increased 9*pp* with the SMO, showing that for these classifiers the addition of the border-line features to the commonly used features helps solving the main issues in the classification, as discussed in the previous paragraph.

**Dermofit dataset** The Dermofit dataset poses a different challenge to the classifiers due to the unbalanced dataset, despite providing more data for training. As previously mentioned, the dataset was classified using the proposed approach by testing each of the 2 proposed segmentation methods (GHT, LBPC) plus the provided ground truth segmentation, GT. The features were extracted from each segmented image using both methods. Then, they were tested separately with three classifiers (SMO, ISDA, and FFN) to perform the lesion classification, achieving the results depicted in Table 3.10.

When using the provided GT segmentation, the additional border-line-based features led to ACC enhancements of 6*pp*, 6*pp*, and 4*pp* on SMO, ISDA, and FFN, respectively. The best performance was obtained with the ISDA classifier (89% ACC). It is also relevant to notice the significant gains in SPE (46*pp*, 42*pp*, and 13*pp*), which indicates that adding the new features helps the classifiers to better cope with the class imbalance.

**Table 3.10:** Border-Line Results for the Dermofit dataset.

| Seg. | Ft. (#) | SVM-SMO | | | | SVM-ISDA | | | | FFN | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC | ACC | SEN | SPE | BAC |
| GT | F23-F32 (10) | 82±0.1 | 3±1.4 | 100±0.4 | 52±0.5 | 83±0.6 | 17±2.5 | 98±0.3 | 58±1.3 | 84±1.9 | 38±5.6 | 95±1.5 | 67±3.5 |
| | F1-F32 (32) | **88**±0.9 | 49±4.2 | 96±1.1 | 73±1.9 | **89**±0.5 | 59±1.2 | 96±0.5 | 77±0.4 | **88**±0.8 | 51±3.9 | 96±0.8 | 73±2.3 |
| GHT | F23-F32 (10) | 83±0.2 | 9±1.3 | 100±0.3 | 54±0.5 | 88±0.3 | 52±1.6 | 97±0.3 | 75±0.9 | 86±0.8 | 49±2.8 | 96±0.9 | 72±1.8 |
| | F1-F32 (32) | **88**±0.4 | 40±3.9 | 99±1.3 | 69±1.4 | **90**±0.4 | 56±2.1 | 98±0.2 | 77±1.1 | **88**±0.9 | 54±2.9 | 96±0.8 | 75±1.8 |
| LBPC | F23-F32 (10) | 81±0.7 | 5±5.4 | 99±0.8 | 52±2.4 | 84±0.8 | 27±1.9 | 97±0.6 | 62±1.2 | 86±1.1 | 50±5.2 | 95±1.2 | 73±3.1 |
| | F1-F32 (32) | **87**±0.6 | 43±3.8 | 97±1.0 | 70±1.6 | **89**±0.5 | 64±1.7 | 95±0.5 | 79±0.7 | **91**±0.8 | 68±3.5 | 96±0.6 | 82±1.6 |

Concerning GHT-based segmentations, the first observation is that it yields better results than GT even with only the initial features, which can be seen as a indication of a higher quality lesion segmentation. With the additional border-line-based features, gains of 5*pp*, 2*pp*, and 2*pp* were now observed in ACC, with ISDA reaching a top score level of 90%. As with GT, more significant increases were observed in the SPE (31*pp*, 4*pp*, and 5*pp*), showing that, again, the new border-line features make the classifiers better prepared to handle the dataset class imbalance.

Finally, the LBPC segmentation method presents results similar to those previously discussed. In this case, gains of 6*pp*, 5*pp*, and 5*pp* were obtained in the ACC, and of 38*pp*, 36*pp*, and 18*pp* in SPE, with FFN outperforming the other two classifiers.

On a global analysis, GHT features (with the ISDA classifier) and LBPC-based features (with the FFN classifier) led to best results in terms of ACC (90% and 91%), but with the latter exhibiting a better SPE (68% against 56%), indicating it deals better with the previously mentioned dataset class imbalance problem.

### 3.4.4 Statistics

To validate the previous results, the following paragraphs presents statistical information about: 1) the discussed material, providing an insight about each feature's usefulness; 2)the reason why all features are kept, instead of using only the most useful; 3) the classifier's statistical information using the selected features.

To improve the presentation of the results, only the Dermofit dataset and the SMO classifier were considered. The dataset was chosen due to its large class unbalance. The choice of the classifier is related to its low computational constraints.

**Feature Information** Three separate feature selection algorithms were used due to their different capabilities to evaluate the worthiness of an attribute: one to measure the correlation

**Table 3.11:** Border-Line Feature Evaluation using 3 algorithm metrics. Numbers in Attr column refer to the features proposed in Feature Extraction paragraph of Section 3.4.2.

| Correlation | | InfoGain | | Relief | |
|---|---|---|---|---|---|
| Merit | Attr | Merit | Attr | Merit | Attr |
| 26,6±3,0 | 1 | 34,0±1,3 | 1 | 10,2±0,4 | 1 |
| 25,2±3,0 | 2 | 31,6±2,0 | 2 | 9,3±0,5 | 2 |
| 22,1±1,0 | 6 | 17,7±2,0 | 3 | 3,9±0,3 | 12 |
| 12,4±1,7 | 15 | 7,0±1,6 | 19 | 3,6±0,4 | 4 |
| 12,3±2,0 | 16 | 7,5±3,8 | 5 | 2,7±0,7 | 6 |
| 12,2±2,4 | 3 | 6,8±0,7 | 6 | 2,5±0,4 | 3 |
| 12,0±2,0 | 17 | 8,7±4,9 | 12 | 2,1±0,1 | 20 |
| 11,9±2,0 | 14 | 4,6±2,7 | 4 | 2,1±0,1 | 22 |
| 11,0±1,7 | 13 | 5,2±0,7 | 21 | 2,1±0,1 | 21 |
| 8,7±1,7 | 12 | 5,2±0,6 | 18 | 2,1±0,2 | 19 |
| 5,9±2,4 | 18 | 5,2±0,6 | 22 | 1,7±0,2 | 18 |
| 7,5±1,5 | 4 | 4,9±0,6 | 20 | 1,2±0,1 | 10 |
| 6,1±2,6 | 19 | 0,0±0,0 | 17 | 1,1±0,3 | 11 |
| 6,2±1,7 | 11 | 0,0±0,0 | 13 | 0,9±0,3 | 7 |
| 5,2±2,6 | 20 | 0,0±0,0 | 16 | 0,8±0,1 | 8 |
| 4,9±2,5 | 22 | 0,0±0,0 | 14 | 0,9±0,3 | 5 |
| 3,5±1,0 | 7 | 0,0±0,0 | 15 | 0,7±0,2 | 9 |
| 4,4±2,5 | 21 | 0,0±0,0 | 7 | 0,6±0,1 | 15 |
| 3,1±1,3 | 10 | 0,0±0,0 | 8 | 0,6±0,1 | 17 |
| 3,1±1,4 | 8 | 0,0±0,0 | 10 | 0,6±0,1 | 16 |
| 2,7±1,2 | 9 | 0,0±0,0 | 9 | 0,6±0,1 | 14 |
| 1,9±1,3 | 5 | 0,0±0,0 | 11 | 0,5±0,1 | 13 |

(Pearson's) between the attributes and the class, dubbed Correlation; one to measure the information gain with respect to the class, dubbed InfoGain; and another dubbed ReliefF (Kira & Rendell, 1992; Kononenko, 1994; Robnik-Šikonja & Kononenko, 1997) that performs the evaluation by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class. All the above algorithms were executed using their default literature parameters. Table 3.11 shows the results resorted by the algorithms' metric. All experiments were done with 10-fold CV.

In the presented results, it is clear that features *F1* and *F2* are the most relevant. This may be due to their indirect ability to provide a proportionality of the lesions' average size. Then *F6* and *F3* are the next dominant features across the three feature selection algorithms, again providing information about lesion dimension. Apart from these, the following most significant features belong to the magnitude of the 6 peaks of the DFT and the 5 periodogram PSD.

By the magnitudes provided by the algorithm metrics, many features seem to provide very little information. Specifically with InfoGain, there are 10 features that seem to be completely useless. But this is not case, because if they are removed, there is a strong negative combined impact in the classifier.

**Feature Selection** As mentioned in the previous subsection, there are many features that provide very little information or correlation apart from the first two. Fig. 3.11 provides more detail on the feature contribution for the SMO classifier. The initial x-label '$i$' denotes the use of the previously mentioned 10 base-features from Jafari et al. (2016). Afterwards, at each step along the x-axis, each feature ($F$), denoted in the x-label, is incrementally included in the dataset to plot the corresponding ACC, SPE and SPE y-axis values. As can be seen in the this figure, including all the other features moves the SPE up by 18.4*pp* and SPE down by only 0.9*pp*. This, overall, improves the ACC by 2.7*pp*, which is reason to use all features in this work.

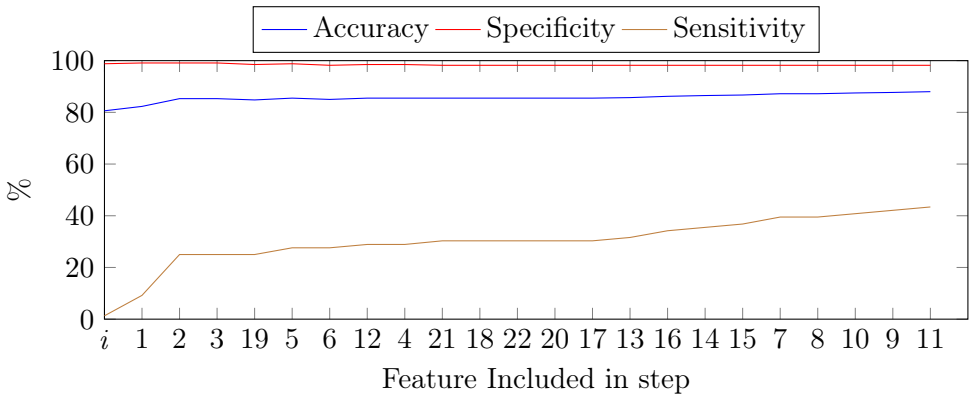



**Figure 3.11:** Feature inclusion plot with Accuracy, Specificity and Sensitivity metrics for the Dermofit dataset and the SMO classifier.

**Classification Significance** The results provided by the classifiers were validated with a corrected paired t-test (Nadeau & Bengio, 2000) to assess if the obtained results with 32 features (32F) are significantly better than the previous 10 features (10F), using a significance $\alpha$ = 0.05. Two hypothesis are tested: *H0)* verifies if the 32F are significantly worse than 10F; and *H1)* verifies if the 32F are significantly better than 10F. If both null-hypothesis are confirmed then it means 32F and 10F are equal. Table 3.12 shows the overall results for the SMO classifier over the three used metrics. The conclusion for the corrected paired t-test is given in column T. The annotation indicates whether a specific result is statistically better (v) or worse (*) than the baseline scheme (10F). Note that they are never statistically equals.

With the presented results it is possible to state that the results obtained with the 32F are significantly better in terms of ACC and SPE, but worse (even if slightly) in terms of SPE, as expected from the previous experiments.

### 3.4.5 Conclusions

This section addressed the importance of the lesion border information on classification of melanocytic skin lesions (namely nevus vs melanoma) in 2D images. Two previously proposed image segmentation methods are exploited to provide the lesion contours, namely the

**Table 3.12:** Border-Line Features Classification Significance Results.

| Seg. | Metric | *H0* | | *H1* | | T |
|---|---|---|---|---|---|---|
| | | ≥ 0 | p-value | ≤ 0 | p-value | |
| GT | Acc. | rejected | 0.0000 | not rejected | 1.0000 | v |
| | SEN | rejected | 0.0000 | not rejected | 1.0000 | v |
| | SPE | not rejected | 1.0000 | rejected | 0.0000 | * |
| GHT | Acc. | rejected | 0.0000 | not rejected | 1.0000 | v |
| | SEN | rejected | 0.0000 | not rejected | 1.0000 | v |
| | SPE | not rejected | 0.9873 | rejected | 0.0127 | * |
| LBP | Acc. | rejected | 0.0000 | not rejected | 1.0000 | v |
| | SEN | rejected | 0.0000 | not rejected | 1.0000 | v |
| | SPE | not rejected | 0.9935 | rejected | 0.0065 | * |

Gradient-based Histogram Thresholding and the Local Binary Pattern Clustering, from which the border-line features are extracted.

The achieved results confirm that segmentation accuracy contributes to enhance the classification performance, namely in methods based on GHT and LBPs, which clearly outperform the GT segmentation provided with the Dermofit dataset. The results obtained with the three considered classifiers confirm that adding border-line lesion features does indeed contribute to improve the performance of automatic classification algorithms.

Moreover, the use of finer segmentation algorithms such as GHT and LBPC was found to be particularly suited for this approach. In fact, the features extracted from their spatially detailed border-lines segmentation improved the classification performance by figures above the gains obtained with the coarser GT segmentation line provided for the Dermofit dataset.

It was shown that using border-line based features together with other commonly used sets can lead to classification results with ACC above 90% in the tested datasets. Additionally, it was shown that these features improve the SPE, which is important when dealing with class imbalanced datasets, as commonly occurs with medical image datasets. Hence, future endeavours might include these types of features to compensate for class imbalance and improve the classification results in general.

## 3.5 Summary

This chapter is focused on the dominant 2D/colour image classification pipeline of skin lesion images. Two segmentation algorithms were proposed, focused on different objectives. One objective was to perform the segmentation like a dermatology expert, by centring the segmentation logic around the fact that skin lesions are bi-model histograms of two dominant peaks.

Another objective was to perform detailed segmentation, capable of finding more realistic borders, which was validated against 39 other literature methods. A comparison was carried out as a contribution to improve dermoscopic image segmentation knowledge against three known datasets. This study allowed to show that, overall, the proposed segmentation method was capable of outperforming the state-of-the-art.

On the topic of classification algorithms, two approaches were proposed. One focused on TL, performing an experimental process that aimed to create a baseline on the current classification expectations to be used with existing DL models. While the other focuses on evaluating the importance of the segmentation mask detail in the classification process, which was performed by assessing the discriminability of the classification of melanoma images. Evidently, the segmentation masks with higher border detail provided higher classification performance in comparison to both the datasets ground-truth masks and the round-like (dermatologist) masks.

With this chapter's insight, further research can be done in the field of skin lesion image segmentation to either improve existing segmentation methods that are lacking in performance or refine the existing top performers. This work allows the detection of which segmentation algorithm is more suitable for a given application by inspecting the strengths and weaknesses of each of the listed algorithms and to decide whether a given algorithm can be further improved.

# Chapter 4

# Contributions using 3D Depth Maps

Content



Light-field imaging technology has been attracting the attention of researchers and engineers due to the ability to capture enriched visual information, which expands the processing capabilities of conventional 2D imaging systems. Light-fields cameras can, for instance, provide dense multiview and multiple focus planes images and extract accurate depth maps. This technology is also emerging in medical imaging research, allowing to find new features and improve classification algorithms, namely those based on ML approaches. Only recently have practical light-field cameras appeared in the market for general use, as a result the availability of light-field content is still a scarce resource for research and development of new image processing algorithms.

This chapter introduces a publicly available light-field image dataset of skin lesions, named SKINL2 (Faria et al., 2019c,a). The dataset currently contains 377 light-fields, captured with a focused plenoptic camera, divided into eight clinical categories, according to the type

of lesion. Each light-field is comprised of 81 different views of the same lesion and a lenslet image. A dermoscopic image of each lesion is also included. This dataset has high potential for advancing medical imaging research and development of new classification algorithms based on light-fields, as well as in clinically-oriented dermatology studies. Therefore, in order to explore and understand the possible gains of this new dimension and exploit the 3D characteristics in the skin lesion surface (thus advancing beyond common features such as, shape, colour, and texture, extracted from dermoscopic RGB images), only 3D information of captured light-fields are used for feature extraction and classification in this chapter.

This chapter describes the developments made with 3D skin lesion surface (depth maps). Skin lesion depth information has not been thoroughly investigated in other works. The hardware and acquisition setup details of the SKINL2 dataset and its contents are discussed in Section 4.1. In the investigations made with this dataset, two algorithms were developed to show that this new dimension has discriminative information relevant for the problem of melanoma classification. The first proposed algorithm, presented in Section 4.2 and published in Pereira et al. (2021c), is a novel approach to this field, where a relevant set of features is investigated to obtain 3D skin lesion characteristics from the depth information. These features were used to train a Bag-of-Features (BoF) model to distinguish between malignant and benign lesions (MAvsBE) or to perform discrimination of melanoma from all other lesion types (MvsAll). The second proposed algorithm, described in Section 4.3 and published in Pereira et al. (2021d), presents a contribution that exploits the lesions' border-line characteristics using this new dimension. A selected group of features is extracted from the depth information to be used for classification of melanoma or nevus images (MvsN) using a quadratic SVM. Finally, Section 4.4 summarises this chapter contributions and highlights the achieved results.

## 4.1 SKINL2 Dataset

The concept of light-field imaging, also known as plenoptic imaging, was first introduced in Lippmann (1908a), as a technique to capture all the information conveyed by the light rays from a visual scene. Following recent technology advances, light-field imaging has been receiving increasing attention, especially since consumer-grade cameras were made available to researchers and the public in general (Raytrix, 2018). The distinctive feature of these devices is their ability for recording not only light intensity, as conventional cameras, but also the direction of light-rays reaching the camera. This is accomplished by using a specially designed array of micro-lenses (MLA) placed in front of the camera sensor (Harris, 2012), each one capturing a different perspective of the scene. More information on other light-field acquisition systems can be found in Zhou & Nayar (2011); Levoy (2006); Levoy & Hanrahan (1996). The ability to capture visual information comprising the light intensity from each point and the direction of the light rays travelling towards the sensor, brings a whole new range of possibilities for 2D and 3D image processing. In fact, light-field image processing allows to render images with different focal planes, depth-of-field or viewing perspectives

(Donatsch et al., 2014; Dansereau et al., 2015), as well as the creation of depth maps (Jeon et al., 2015), which can be used for reconstruction and characterisation of a skin lesion surface, as shown in Fig. 4.1.

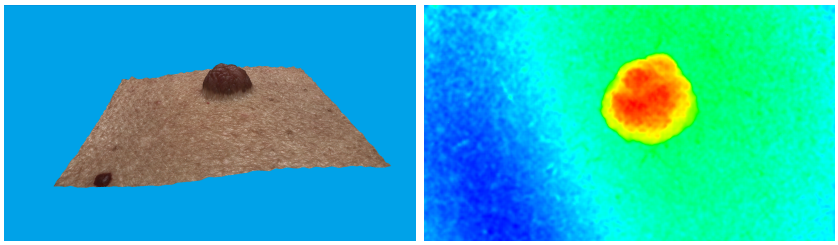


**Figure 4.1:** Example of a 3D skin lesion reconstruction (left) and the corresponding depth map (right).

Dermoscopy and other non-invasive medical imaging technologies have been using digital images of pigmented lesions to detect and assess early signs of malignant lesions. For instance, digital processing of dermoscopic images enables extraction and visualisation of morphological features, which are not discernible by visual inspection. Given its richer content, light-field images add new dimensions of visual information to dermoscopy. Its use is expected to improve medical evaluation by means of a more robust diagnosis, given the additional contribution for higher sensitivity and specificity. In the case of melanoma, the most aggressive form of skin cancer, this technology is particularly relevant because when diagnosed at an early stage, this disease presents high cure rates, and yet using classical imaging techniques it is usually misclassified as a benign nevus.

Sharing publicly available datasets of light-field images with skin lesions is crucial to pursue research at a global scale, and also to establish common references for comparison of results and benchmarking. The new image dataset presented in this section, named Light-field Image Dataset of Skin Lesions (SKINL2), is an enabling resource for pushing forward worldwide research in this specific field. Furthermore, within the scope of the new standard JPEG Pleno (Ebrahimi et al., 2016), new possibilities arise in the development of rendering algorithms, processing techniques and lossy/lossless compression methodologies, specifically tailored for this kind of medical applications. Additionally, the richer content of light-field images of skin lesions can be exploited in terms of the new capabilities of focal plane and depth of field manipulation, as well as 3D-based studies. The main impact of these research results are expected to be in clinical assessment frameworks, which should benefit from the improved feature extraction algorithms and automated classification systems. Some light-field image datasets have been introduced in recent works for different purposes (Rerabek & Ebrahimi, 2016; H. et al., 2016; Guillo et al., 2018), but, to the authors' best knowledge, this is the first dataset of pigmented skin lesion light-field images to be made publicly available.

Currently, there are two published and publicly available versions of the SKINL2 dataset. A third version, also publicly available, is still being updated with new images. The datasets and all its associated information are available at `http://on.ipleiria.pt/plenoisla`.

The remainder of this section is organised as follows. Section 4.1.1 describes the working principle of plenoptic cameras. Section 4.1.2 presents the light-field acquisition setup and procedure. Section 4.1.3 details the acquired datasets and Section 4.1.4 concludes the section.

### 4.1.1 Plenoptic Cameras

The research efforts on plenoptic cameras has resulted in two different optical designs, commonly referred to as Plenoptic 1.0 and Plenoptic 2.0 (Ahmad et al., 2018; Georgiev, 2009). Both share the principle of placing a MLA between the image sensor and the mains lens, but differ on its relative position. In Plenoptic 1.0 cameras the MLA is placed at the focal plane of the main lens, so the image generated by each micro-lens (micro-image) contains information about only one spatial point in the scene. Each pixel in a micro-image is therefore associated to a particular light ray direction and the spatial resolution of the captured light-field is determined by the number of micro-lenses.

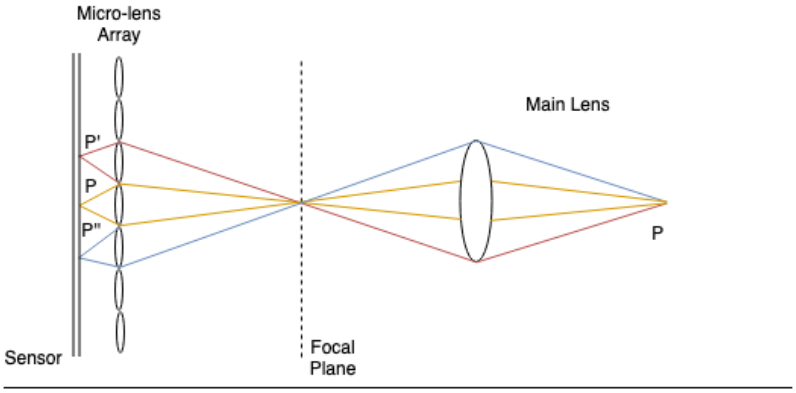



**Figure 4.2:** Plenoptic 2.0 camera diagram.

In Plenoptic 2.0 cameras, the MLA is focused onto the image plane of the main lens, as shown in Fig. 4.2, allowing each micro-lens to record a region of the scene, therefore each point in the scene is visible by different micro-lenses from slightly different perspectives. Although there is a trade-off between spatial and angular resolution that depends on the overlap between micro-images, this enhanced design overcomes the limited spatial resolutions of Plenoptic 1.0 cameras, whose micro-lenses are not focused on the image created by the main lens. A further improvement in some Plenoptic 2.0 cameras is the use of a MLA with different types of lenses, each with a particular focal length, enabling to extend the depth-of-field of the captured light-field.

### 4.1.2 Acquisition

**Setup** The acquisition process uses a Raytrix camera, named R42 Galilean, with a Ricoh $25mm$ f/1.8 lens, which was developed as a Plenoptic 2.0 camera with extended depth of field.

**Figure 4.3:** Acquisition setup: light field camera housing (left) and the camera main lens plus the illumination LED ring (right).

Its MLA comprises three types of micro-lenses, each with a different focal length. As shown in Fig. 4.3, the camera is placed inside a cylindrical acrylic-built black housing. This setup allows the camera to be always placed at the same distance from the imaged object (initially at a distance of $d \approx 197mm$ from skin lesion), thus ensuring optimised focusing conditions and the same magnification factor for all images. The lens is at a distance of $d \approx 140mm$ from the scene, which is illuminated with a ring of 5 neutral white 5050 LEDs, placed inside the tube ($\approx 150mm$ above the imaged object), with the black coating preventing the interference of ambient light. The light intensity of the LEDs is controlled by a computer running dedicated software, thus optimising the illumination conditions for each acquisition.

The camera is provided with the Raytrix's RxLive software (Raytrix, 2019), which is used for both the MLA and metric calibrations, and to record the light-fields. Moreover, RxLive exhibits in real time an all-focus image of the scene, as well as multiview and lenslet images, the corresponding depth map and 3D reconstruction. Information from a given scene can be exported in the form of a total focus image, lenslet or a proprietary format '.ray' (which can afterwards be used to extract total focus/multiview images and depths maps). An API is also available with the same image extraction capabilities of RxLive.

**Methodology** The acquisition methodology comprises first the calibration of the camera and then the light-field acquisition, integrated in the standard procedures of clinical appointments at the Department of Dermatology of Centro Hospitalar de Leiria, Portugal. The camera was calibrated following the manufacturer's documentation, using a calibration target, with a $2.0mm$ point pitch, as described in Johannsen et al. (2013). In this procedure, the position of the camera with respect to the calibration plate was changed using a precision micrometer, so the different distances were established with an accuracy of $0.01mm$.

The light-field images of skin lesions were acquired in dermatology clinical appointments and the procedures were evaluated and approved by a health ethics committee. Additionally, the

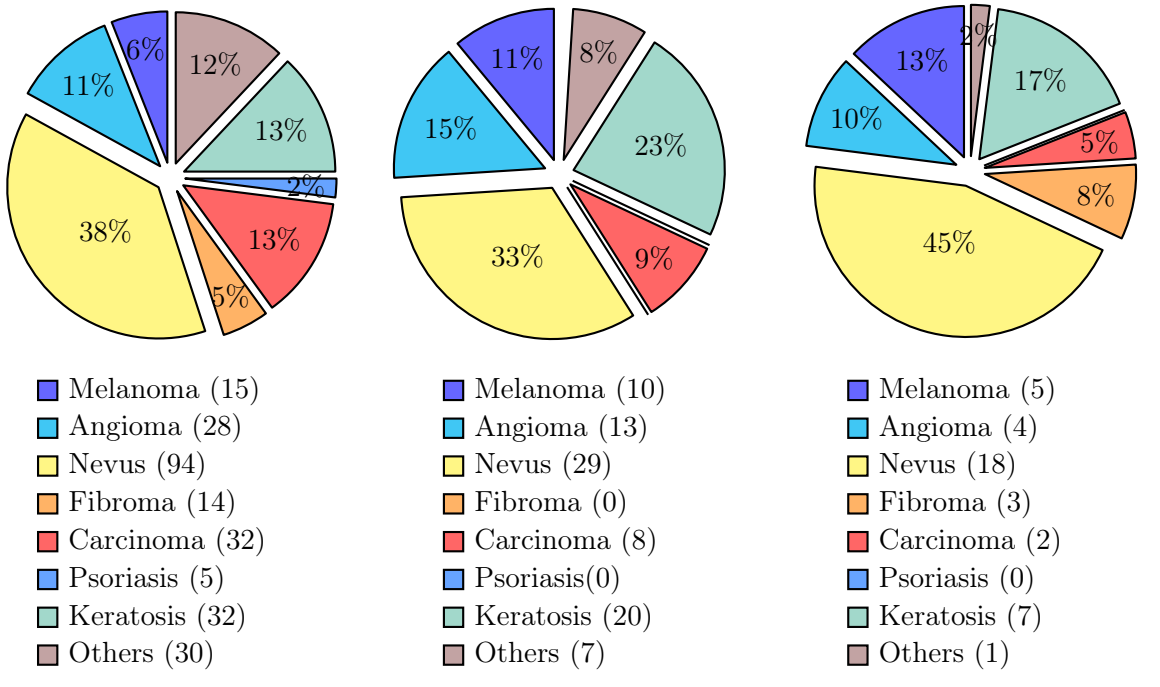
procedure and purpose of the study were explained to all volunteers before signing an informed consent form. Further imaging procedures, such as dermoscopy and standard photography, were also carried out to capture different types of images. Given the variety of skin and lesion tonalities, the light intensity of the LEDs in the acquisition setup was adjusted before capturing every image, in order to prevent either over or underexposure. The skin lesions were manually classified by dermatologists and organised on a clinician diagnosis according to ICD10 (International Classification of Diseases), and a histopathological analysis was done whenever a confirmation was required.

### 4.1.3 Dataset

A quick overview of the datasets composition is shown in Fig. 4.4 and some thumbnail samples are provided in Fig. 4.5. The SKINL2 datasets are classified into eight categories, according to the type of skin lesion/ICD code:

- Melanoma / C43;
- Melanocytic Nevus (Nevus) / D22;
- Basal-cell Carcinoma / D04;
- Seborrheic Keratosis / L82;
- Hemangioma (Angioma) / D18;
- Dermatofibroma (Fibroma) / D23;
- Psoriasis / L40;
- Others.

Any light-field that does not match one of the first seven ICD codes it is archived under *Others*.




**(a)** SKINL2 v1 **(b)** SKINL2 v2 **(c)** SKINL2 v3

**Figure 4.4:** Pathological distribution of captured light-fields.

**(a)** SKINL2 v1 **(b)** SKINL2 v2 **(c)** SKINL2 v3

**Figure 4.5:** Central-view sample images for each version of the SKINL2 dataset. In each, from left to right, top to bottom: Angioma, Carcinoma; Nevus, Melanoma; Keratosis, and Others.

The light-field dataset is made available in two different formats: *i)* a lenselet, with a resolution of $7716 \times 5364$ pixels (8-bit depth RGB components), and *ii)* a matrix of $9 \times 9$ views, each one with a resolution of $3858 \times 2682$ pixels, where each pixel is represented by 16-bit depth RGB components. This matrix of views was obtained from the light-field, using the Raytrix API (Raytrix, 2018), by adjusting the parameter *VirtCamPinholeStd_ ViewOffsetMM_ g* in function *RxSetPar*. The views are horizontally and vertically spaced by $1mm$, and the dimension of $9 \times 9$ was considered the ideal number of views for estimation of the depth maps, following the analysis in Wanner (2014). Each light-field also includes a version of the central view in the all-focus format (as shown in Fig. 4.5) and the depth map obtained with the Raytrix API. Additionally, a regular dermoscopic image of each lesion is also provided, with a resolution of $1920 \times 1080$ pixels, being each pixel represented by 8-bit depth RGB components. In summary, this means that, for each recorded light-field, 82 files exist in the database.

**SKINL2 v1** The SKINL2 v1 dataset comprises 250 light-fields. The light-fields in the dataset are representative of the diversity of observed pathologies (see Fig. 4.4a) with nevus as the largest pathology group represented.

**SKINL2 v2** A second version of the dataset is also available with 87 images. This version was adquired during a shorter time frame and therefore possesses fewer images. Some changes were introduced in the setup, with the distance between the camera being reduced to $d \approx 95mm$ (and so the LED ring to $d \approx 105mm$). These changes were introduced in order to increase the magnification by $\approx 30\%$ when compared to the SKINL2 v1 dataset setup, thus enhancing the spatial detail and, consequently, the depth map resolution. This magnification is visible when comparing Fig. 4.5a with Fig. 4.5b.

**SKINL2 v3** The third version of the dataset is still in the acquisition phase at the moment of publication of this thesis, however its progress was affected by the interruption on medical

screening during the Covid-19 pandemic. As it is visible in Fig. 4.4c, the third version is still very small in comparison to the previous iterations.

### 4.1.4 Conclusions

This section described a new publicly available light-field dataset currently containing 377 skin lesions acquired with a focused plenoptic Raytrix camera. These light-field images are classified into eight different categories, according to the type of skin lesion. Moreover, the dermoscopic images corresponding to each lesion are also provided in the dataset together with the total focus image, a matrix of views, and depth maps. This publicly available set of light-field images intends to contribute for further research in the fields of medical imaging and clinical diagnosis support, as performed in the JPEG proposal Faria et al. (2019b).

## 4.2 Classification using Bag-of-3D-Features

Various feature extraction methods resorting to 3D information are already described in the literature, as presented in Section 2.6. From those, a total of 11 were selected to be used in this work to either characterise the underlying 3D image globally or characterise it in key local regions. The adequate selection of these regions is also a research topic, therefore two different methods are studied.

The main contribution of this section is to demonstrate that 3D information from skin lesions contains relevant discriminative features capable of providing high classification precision of melanoma versus nevus. Such third dimension, that is beyond conventional colour, texture, and shape proves to be beneficial for the classification process. This research's main focus is to to evidence that skin surface topology has potential discriminative information for the classification. However, it does not aim to use RGB information nor improve existing algorithms.

The remainder of this section is organised as follows: Section 4.2.1 presents the background that is relevant for the proposed method. Section 4.2.2 presents the proposed approach, describing feature extraction and relevant classification details. Section 4.2.3 presents and discusses the attained results and Section 4.2.4 highlights the conclusions.

### 4.2.1 Relevant Background

Since BoF models (Sivic & Zisserman, 2003; Csurka et al., 2004) were proposed for skin lesion classification in 2008 (Situ et al.), several research works have been published resorting to

it (Hu et al., 2019). In image classification, a bag of features is a vector of occurrence counts of a dictionary of local image features, which can also be understood as a histogram over extracted image features. This type of structure can be employed for classification resorting to a SVM classifier. The image features that compose the BoF models are designed for object detection, classification, or retrieval generally fall under two base types: signature or histogram (Serratosa & Sanfeliu, 2006). Signature-based feature extractors aim to register specific object characteristics or attributes capable of providing discrimination against other objects or scenes, effectively, a signature of a set is a unambiguous representation of its histogram. Examples of signature based features extractors are: Normal Aligned Radial Features (NARF) (Steder et al., 2011), Radius-based Surface Descriptor (RSD) (Marton et al., 2010, 2011), Global RSD (GRSD) (Kanezaki et al., 2011), and Principal Curvatures (PC) (Rusu & Cousins, 2011). In contrast, histogram-based extractors aim to produce a summarised representation of the underlying data, typically, the presence of a set of features and their occurrence count. Examples of histogram based features extractors are: Rotation Invariant Feature Transform (RIFT) (Lazebnik et al., 2005), Point Feature Histogram (PFH) (Rusu et al., 2008), Fast PFH (FPFH) (Rusu et al., 2009a,b), Signature of Histograms of OrienTations (SHOT) (Tombari et al., 2010b, 2011), Ensemble of Shape Functions (ESF) (Wohlkinger & Vincze, 2011), 3D Shape Context (SC3D) (Frome et al., 2004), and Unique Shape Context (USC) (Tombari et al., 2010a). More details on these features are described in Section 2.6.

In some cases, having a large number of features can be a problem, e.g., when several features are extracted but their relevance for the intended solution is unknown or when there are insufficient data samples. A useful method to reduce the number of features, by selecting the most meaningful ones, is the Neighborhood Component Analysis (NCA) (Yang et al., 2012). NCA is a non-parametric algorithm that enables feature selection with the goal of maximising prediction accuracy of regression and classification algorithms.

### 4.2.2 Proposed Bag-of-3D-Features Classification Approach

The main goal of the work described in this section is to perform the classification of malignant skin lesions based on 3D surface information. To this end, the utilised methodology comprises a BoF approach, as in Sivic & Zisserman (2003); Csurka et al. (2004); Situ et al. (2008), with a dataset holdout of 30% on the SKINL2 dataset where, as a pre-processing stage, pixel values in the RGB channels of all images were replaced by zeros (since some of the selected feature extractors also consider colour information). This means the colour information is not used, only the depth. The following paragraphs provide added details to the pipeline, namely about the selected keypoint detectors and feature extractors, as well as information about the BoF model.

**Features** A total of 5215 features were extracted from each image using 11 features extractors. These extractors were selected based on the relevance of their characteristics for

the input signal (3D information). RIFT (32 features) was selected because it provides invariant to illumination, viewpoint, scale, and rotation. Like RIFT, NARF (42 features) and PFH/FPFH (125/33 features) also possess some of these characteristics, PFH/FPFH, in particular, provides robustness against outliers and noise. Other features extractors as SHOT (361 features), SC3D (1989 features), and USC (1969 features) also provide robustness against noise. Additionally, both SHOT and USC are reported to provide uniqueness amongst detection, as well as unambiguous representations. Finally, ESF (640 features), PC (5 features), RSD (2 features), and GRSD (21 features) were selected for being descriptive, simple, and intuitive shape descriptions. ESF has proven to be efficient and expressive, while GRSD adds expressiveness to the simple RSD, by partitioning the image point cloud into several voxel-surfaces of understandable shapes.

Apart from ESF and GRSD, the other feature extractors operate on specified image keypoints, which must be predetermined. In order to provide such set of locations, two keypoint detectors were selected: NARF and ISS. The NARF detector seems specially suited for skin lesion imaging since it selects locations of high surface changes and takes object borders into account, as in skin to lesion borders, which have already been noticed to have relevant information (Pereira et al., 2020b). Then, ISS is also selected because, like NARF, it produces keypoints which tend to be at saliency regions, like the lesion border or texture-full regions inside the lesion, but in a more selective manner (outputting less keypoints).

**BoF model** The dataset images were divided into a training and a testing set, the former with 70% and the latter with 30%. A set of keypoints are extracted from the training process, producing 5215 features each, and a SVM model is trained on the histograms produced after applying k-means clustering to those features. A SVM model is selected for this work because the mentioned dataset provides fewer images than typically necessary for DL approaches. When building the BoF model, for classification of malignant versus benign (MAvsBE) lesions or for melanoma versus all other lesions (MvsAll), the used SVM classifier is a polynomial kernel of second order and has box constraint of 1. Since all experiments are defined for binary classification, the SVM solver is the Iterative Single Data Algorithm, which minimises by a series of one-point minimisations and does not respect the linear constraint nor explicitly includes the bias term (Kecman et al., 2005a).

Because some of the features might not contribute for the adequate label separation, or might effectively injure the model's capability, during the classification training process the feature selection is also performed with NCA (on the targeted 70%) before training the BoF model. Fitting of the NCA model is done with all training samples and using a stochastic gradient descent solver. The NCA algorithm is susceptible to overfitting but possesses a parameter to prevent it through regularisation. This value is fine tuned via grid-search in the range of $[0; 0.003]$, where 20 equidistant grid-points are selected. At the end, the NCA model provides a relevance-weight for each of the 5215 features. In this experiment, only those with a relevance superior to $0.02 * \max(1, \max(f_w))$ are selected, where $f_w$ is a vector with all provided NCA feature weights.

### 4.2.3 Results and Discussion

The proposed pipeline was applied to the publicly available SKINL2 dataset. Particularly in this work, the second version of this dataset was used, due to its increase in lens magnification of about 30% (which means more detail) in comparison to the first version of the dataset. At the time of this study the dataset comprised 19 malignant lesion images (9 melanomas, 9 basal cell carcinomas, and 1 squamous cell carcinoma) and 66 benign lesion images (32 nevi, 13 angiomas, and 21 seborrheic keratoses), which undergo the pre-processing, feature extraction and classification processes, described in Section 4.2.2.

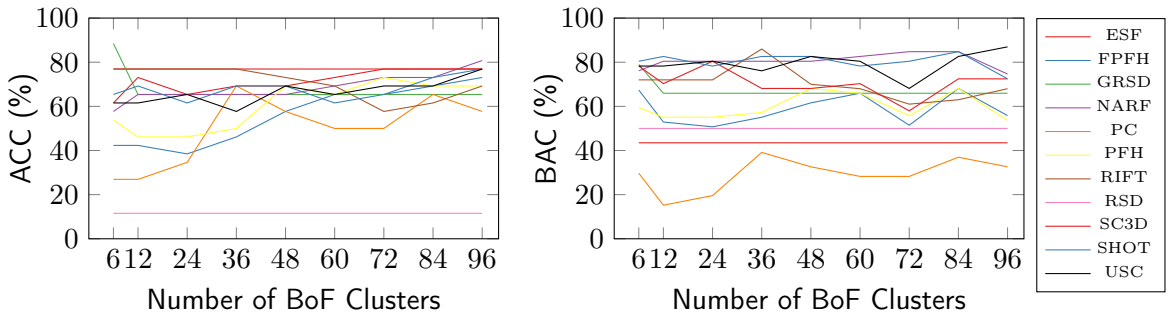
The results obtained from these assessments are recorded in terms of percentage of classification accuracy (ACC), specificity (SPE), and sensitivity (SEN). In addition, because this is an unbalanced problem, the named balanced-accuracy (BAC) is also used (as in Hu et al. (2019)) as it corresponds to the average value between SEN and SPE.

**Table 4.1:** Bag-of-3D-Features Overall Top Results.

| Experiment | Detector | #Clusters | ACC | SEN | SPE | BAC |
|---|---|---|---|---|---|---|
| MAvsBE | NARF | 84 | 73.08 | 75.00 | 66.67 | 70.83 |
| | | 96 | 80.77 | 100.00 | 15.00 | 57.50 |
| MvsAll | NARF | 48 | 84.62 | 66.67 | 86.96 | 76.81 |
| | | 96 | 84.62 | 66.67 | 86.96 | 76.81 |
| | ISS | 96 | 84.62 | 66.67 | 86.96 | 76.81 |

The main experimental results are shown in Table 4.1 and the individual behaviour of the different extracted features is plotted in Fig. 4.6. Table 4.2 adds to Table 4.1 with results resorting to the features selected after running the NCA algorithm. Only best BAC results are shown in the tables. In these tables, results are shown without background shading while values above 75% are highlighted in grayscale towards 100%. Given the available dataset samples previous described in this section, in these tables the column "Experiment" indicates the classification objectives, being either "MAvsBE" for malignant versus benign lesions or "MvsAll" for melanoma versus all other skin lesion types. Additionally, as mentioned in Section 4.2.2, a keypoint detector is necessary for most of the feature extractors, therefore column "Detector" is present to indicate which of the two selected keypoint extractors was used. As the BoF model pipeline uses a k-means clustering algorithm, column "#Clusters" expresses the number of specified clusters. In this research the number of clusters was defined as: either 6 or multiples of 12 up to 96, in a total of nine variations, as represented by x-axis of Fig. 4.6.

From Table 4.1, the highest ACC result obtained for detecting malignant skin lesions is 80.77%, when the BoF pipeline uses 96 histogram bins for classification. Although this is not the best overall result, some clinicians find it appealing as it presents 100% SEN, meaning that no life-threatening condition goes unchecked. It is important to notice that, in this case, the SPE metric indicates that correct classification of benign lesions occur only 15.00% of the time, meaning




**(a)** Accuracy **(b)** Balanced-Accuracy

**Figure 4.6:** Performance of each feature extractor for *Me vs All* classification problem: (a) ACC metric and (b) BAC metric.

that 83.33% of the benign lesions are incorrectly labelled as malignant instead of benign. A more balanced solution is achieved with 73.08% ACC when merging some of the data points, by using less clusters (84). This solution comprises 75.00% SEN and 66.67% SPE, meaning that 25% of the malignant lesions pass as benign but only 33.33% of the benign lesions get classified as malignant, in comparison to the previous 83.33%. Focusing on the melanoma lesion type, a higher ACC result of 84.62% is achieved when performing direct comparison between melanoma and the remaining skin lesion images. In this case, the best results achieved using the NARF detector are also attainable using the ISS detector. However, balanced SEN and SPE results show that: the algorithm is only capable of correctly classify melanomas 66.67% of the time (SEN), while the benign lesions are correctly classified 86.96% of the time (SPE).

The achieved balance of the metrics shows that some information exists in the 3D surface that enables a level of discrimination between skin lesion types. By observing the individual behaviour of the different feature extractors in Fig. 4.6, it is possible to infer their contributions towards the current melanoma classification results shown in Table 4.1. Such behaviour is not uniform across all features extractors, but from the accuracy metric in Fig. 4.6a it can be seen that there is a trend to provide superior accuracy results as the number of clusters increases, although not all feature extractors follow this rule. An exception to this trend occurs, for instance, for RSD features, which provide a constant 11.54% accuracy, resulting from the classification of every sample as melanoma. This also means that the BAC metric for the RSD, in Fig. 4.6b, is 50% $((100 + 0)/2)$. Another example of a non-discriminative set of features is the PC, which always presents a BAC performance lower than RSD, e.g. 32.61% for 96 Clusters, despite presenting a higher accuracy. A possible reason for this type of contradictory results is the use of a unbalanced dataset. RSD always labels samples as melanoma, the smaller class (lower accuracy), while PC mostly labels samples as nevus, the larger class (higher accuracy). Apart from the mentioned outliers, what stands out the most in Fig. 4.6b is that several individual feature extractor results (26, at different cluster settings) achieve BAC performances that are superior to the recorded 76.81% in Table 4.1, which are only achievable with balanced SEN and SPE settings, thus also generating high accuracy values. In particular, results above 80% are attained when using either NARF, RIFT, SC3D, SHOT, or USC. Specifically, USC and RIFT are able to reach a top performance of 86.96% BAC.

**Table 4.2:** Bag-of-3D-Features Overall Top Results after NCA.

| Experiment | Detector | #Clusters | ACC | SEN | SPE | BAC |
|---|---|---|---|---|---|---|
| MAvsBE | NARF | 6 | 61.54 | 100.00 | 50.00 | 75.00 |
| | | 84 | 69.23 | 83.33 | 65.00 | 74.17 |
| MvsAll | NARF | 24 | 84.62 | 66.67 | 86.96 | 76.81 |
| | | 48 | 88.46 | 100.00 | 86.96 | 93.48 |

With the previously mentioned insights, it becomes clear that the BoF model is not able to withstand the presence of non-discriminative features and performs poorly in comparison to use only a feature extractor's individual-set of features. Nevertheless, the combination of various subsets of different feature extractors could still yield even higher performance results. For this reason, the NCA feature selector is introduced in the experimental setup. Results obtained after applying the NCA selection are depicted in Table 4.2. Only the experiments with the highest performing BAC are shown in the table (changing the selected "#Clusters" column from Table 4.1 to Table 4.2).

When applying NCA to the Malignant versus Benign problem, 21 features are selected and the performance of the BAC metric increases from the previous 70.83% to 75.00%. Also, when applying NCA to the melanoma versus all other lesions problem, fewer (14) features are selected, but the BAC metric achieves the best performance of 93.48% (from the previous 76.81% in Table 4.1). Independently of the problem, the best training results were achieved with a lambda of 0.0028. In comparison to the results presented in Table 4.1, the NCA selected feature subset provides significant improvements (in Table 4.2). With 48 clusters, the BoF model achieves 88.46% accuracy by only using depth-based features, in the melanoma versus all other lesions problem. In addition, the generated model presents the capability of correctly identifying all melanoma samples (100.00% SEN), while only incorrectly labelling 13.04% of benign lesions as malignant ones (86.96% SPE). As expected, these results are far superior than using an individual set of features with only one feature extractor, e.g. 93.48% BAC for NCA, in Table 4.2, against 86.96% BAC for USC at 96 Cluster, in Fig. 4.6b.

### 4.2.4 Conclusions

Despite recent advances in the classification technology, classification (or discrimination) of melanoma versus nevus still remains difficult to achieve, due to its similarity at an early stage of the lesion development. A reliable solution might depend on the use of new acquisition modalities instead of the widely available, and utilised, 2D dermoscopic images, which may introduce new fairly unexploited dimensions.

The main contribution of this work is exploitation of depth information from light-field images for classification of skin lesions. This newly introduced type of 3D data was specifically

acquired for this purpose and has shown the ability to provide rich information for image classification. Several literature methods already exist to extract 3D surface information for general classification purposes. However, these features were not originally developed for skin lesion classification. Some of these works jointly extract and classify 3D surface information resorting to a Bag-of-Features model.

As previously exposed, classification between benign and malignant lesions achieved 75.00% BAC, comprising 61.54% accuracy, 100.00% SEN, and 50.00% SPE. In a more explicit setting, discrimination of melanomas against all other available skin lesions was achieved with 88.46% accuracy, 100.00% SEN, and 86.96% SPE, with a BAC of 93.48%. These results evidence the usefulness of unexploited 3D lesion surface information in the classification process of skin lesions.

## 4.3 Classification using 3D Border-Lines Features

Depth information can be obtained from images through different approaches and acquisition setups. In this section, depth maps are extracted from dense light-fields of the SKINL2 dataset. Each light-field presents over $10,000,000$ pixels, yet if only a small line of less than $5,000$ pixels is extracted along the lesions' perimeter region, it might contain relevant information to classify the type of lesion. From such border-line, features can be calculated using solely the depth information present along the set of connected pixels. Assuming that surface-level information (texture) differs from melanoma to nevus, the respective border-lines are expected to have structural differences between them or a different overall geometric behaviour. This is similar to what was performed in Section 3.4 but now using depth information of the skin surface. Thus, discriminative features capable of extracting relevant information about such type of details must be used. Features extracted from electrocardiogram (ECG) signals seem specially suitable for this task, since in ECG classification problems it is necessary to discriminate patterns from fine variations along a one-dimensional signal. Therefore, a similar type of features may be employed for classification of skin lesions, based on the depth values of the border-lines.

Overall, the main contribution of this section is the exploitation of 3D information from skin lesions, aiming to achieve high discrimination in the classification of melanoma versus nevus and, consequently, showing that this third dimension provides significant information for classification. Different from previous studies, this work investigates new 3D information from the segmentation mask border-line to provide evidence that skin surface topology has potential discriminative information.

The remainder of the section is organised as follows: Section 4.3.1 presents relevant background and other similar experiments. Section 4.3.2 describes the proposed approach and details about data pre-processing, feature extraction and classification. Finally, Section 4.3.3

presents and discusses the experimental results and Section 4.3.4 exhibits the conclusions and future work.

### 4.3.1 Relevant Background

Classification results obtained using only dermoscopic information are rather limited, as only planar information can be retrieved from such data. To overcome this limitation, in McDonagh et al. (2008); Smith et al. (2011), a method using a stereoscopy technology is presented. Although the literature addressing 3D surface studies of melanoma and other skin lesions is almost nonexistent, some previous research indicate that improved results arise when using depth information (McDonagh et al., 2008; Smith et al., 2011). In Satheesha et al. (2017), artificial 3D information was generated for datasets to improve the classification results.

Similarly to the method proposed in Chapter 3.4 (which uses dermoscopic images), published in Pereira et al. (2020b), in this work only depth information located at the lesions' border region is utilised for classification. The remaining image data of the lesion is discarded. This depth information in the border-line is represented by a one-dimensional signal, from which a set of discriminative features is extracted.

In regard to feature extraction, one can consider two main approaches: either DL or hand-crafted features. In this work, the latter option was used since the SKINL2 dataset is very small in comparison to what is normally necessary for DL approaches. Due to the reduced size of the dataset and the large amount of pixel data in each light-field, the global depth map of the skin lesion is reduced in size, although keeping its 3D discriminative characteristics. This is done by only considering few border-lines of the segmented lesion. Such data reduction is also necessary to avoid over-training, as pointed out in McDonagh et al. (2008); Smith et al. (2011).

The depth information of a border-line can be analysed as a time-series, like other types of known signals such as ECG for instance. Thus, relevant characteristics can be discriminated by extracting the same type of features. Examples include regression/prediction coefficients as in Zhao & Zhang (2005), localised entropy values as in L. & Z. (2016), or some form of wavelet observation as in Leonarduzzi et al. (2010).

In Zhao & Zhang (2005), the authors present a feature extraction approach for reliable heart rhythm recognition. After data pre-processing and feature extraction steps, the classifier recognition of 6 types of heart rhythm reaches 99.68% by receiving two sets of features: the transform coefficients of a wavelet transform; and the values of auto regressive modelling applied to the temporal structures of ECG wave forms (model order selection is described by minimisation methods). In L. & Z. (2016), the authors detail experiments about the influence on the performance of different mother wavelets and level of decomposition for wavelet packet

decomposition, type of entropy, and the number of base learners in a random forest classifier. The authors state that experimental results were superior to those of several state-of-the-art competing methods, showing that wavelet packet entropy had promising results for 1D signal classification, such as of ECG. In Leonarduzzi et al. (2010), the authors explain that such signals present complex irregular fluctuations. Hence, to extract information related with such fluctuations, the authors use multi-fractal analysis, specifically wavelet leader based multi-fractal analysis in short-time windows, which had already been proposed in Jaffard et al. (2006) and achieved superior results.

### 4.3.2 Proposed 3D Border-Line Classification Approach

In this approach, the main goal of the skin lesion classification is to distinguish between melanomas and nevi. To this end, as depicted in Fig. 4.7, the utilised methodology pipeline is comprised of three main steps: data preparation, comprised of extraction, preparation, and augmentation; feature extraction; and classification.

**Data Preparation** For each 4-channel image (RGB+Z) of the SKINL2 dataset (Fig. 4.8a), a lesion mask (Fig. 4.8b) is manually generated (using the colour RGB channels, Fig. 4.8c) so that the lesion perimeter pixels could be identified and their depth channel values (hereinafter the "Z" channel, Fig. 4.8d) sequentially extracted from a random starting position (hereinafter the "border-line" vector, Fig. 4.8e).The Z-pixel values might not be all within acceptable ranges (mainly due to errors caused by light reflection), therefore all border-line values higher than 10mm (chosen as empirically threshold) were replaced by previously valid values in the sequence. Afterwards, border-line Z-values are normalised to a range of $[-0.5, 0.5]$. Note that RGB is only used to produce the segmentation mask (being discarded discarded afterwards), as all data used to train the model is from the Z channel.

Additionally, three supplementary border-lines may be extracted depending on the experiment augmentation settings. If enabled, this step iteratively shrinks the lesion mask by 20 pixels until it produces 3 inner border-lines, which are inside the lesion region. The rationale behind this is that melanoma and nevus surfaces are different, hence more information would allow to compensate any model overfitting and also reinforce a better comprehension about the problem dimension to the classifier.

Since not all border-lines have the same length (as skin lesions come in all shapes and sizes), it was necessary to uniformise their size before the feature extraction process. Thus, four transformations were considered: **T1**) pad smaller lines with zeros; **T2**) repeat (by rotation) smaller lines; **T3**) linearly stretch smaller lines; or **T4**) cubicly interpolate smaller lines.

In addition to the already mentioned options, further data augmentation techniques were

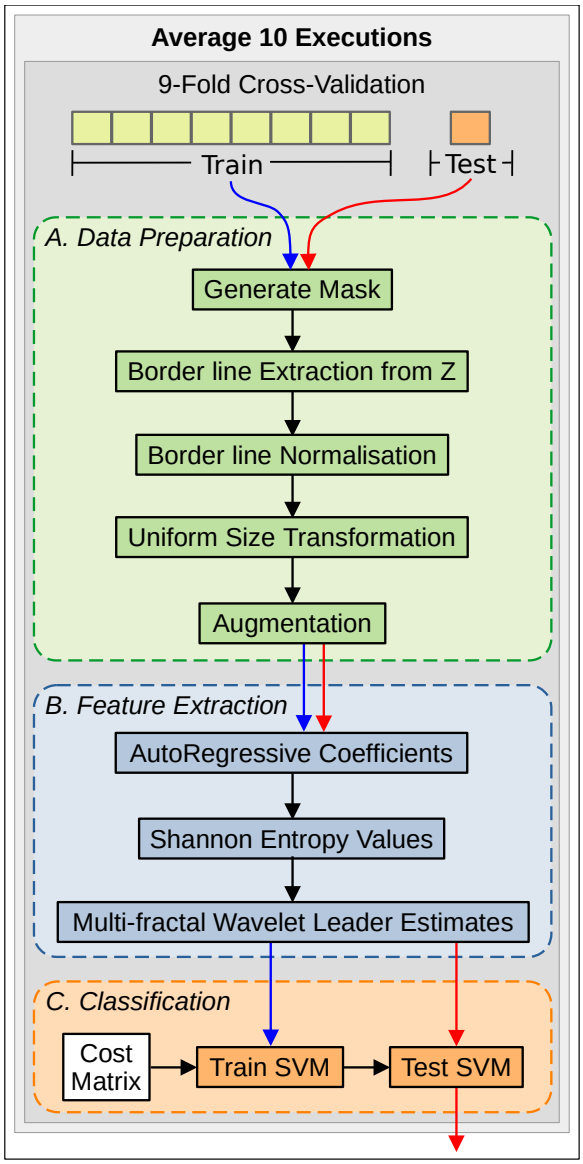
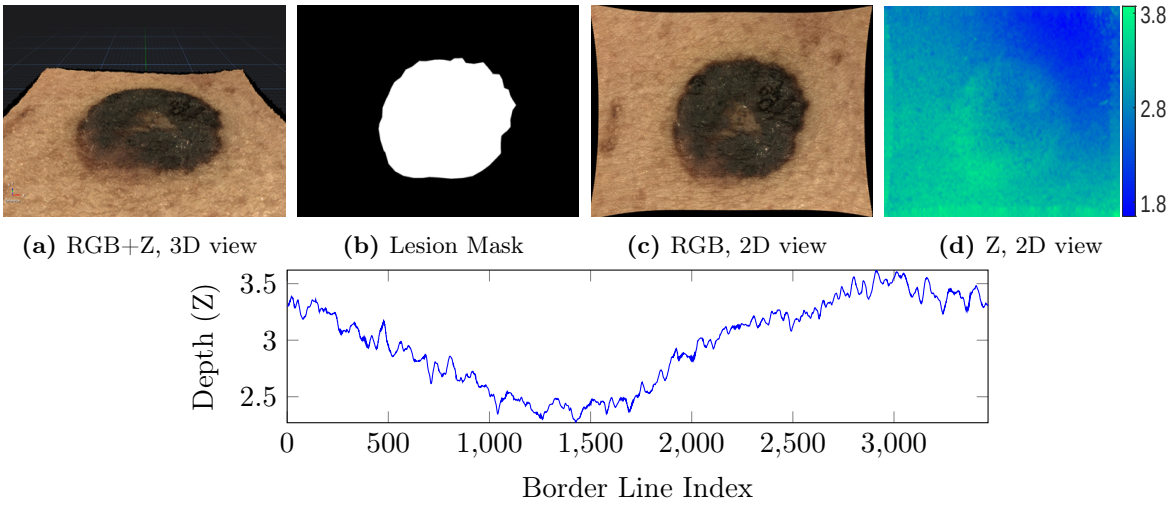



**Figure 4.7:** Proposed methodology pipeline comprises three main blocks, which are executed inside a 9-Fold Cross-Validation scheme that is executed 10×. Given train and test data, the first block prepares the data by: generating a lesion mask so that border lines may be extracted from the Z dimension; which are then normalised and transformed to a uniform size; and augmented in the end. Given the prepared data, the second block performs feature extraction by: computing named AutoRegressive Coefficients, Shannon Entropy Values, and Multi- fractal Wavelet Leader Estimates. Given the extracted features, the third block trains an SVM using features generated from the fold-dependent train data and a defined cost matrix, and later tests the SVM model using features generated from the fold-dependent test data. Blue arrows indicate the pipeline training sequence. Red arrows indicate the pipeline testing sequence. Black arrows indicate previous dependencies or common progressions through the pipeline.

added to evaluate the generalisation of the model: by flipping/inverting all border-lines (doubling the dataset size); and repeating data to balance-out the smaller class samples.

**Feature Extraction** After having a set of equal-sized border-lines, feature extraction takes place. Most of the features provide global representations of the border-line (which makes fine details along smaller sections become expressionless) or require continuous samples of the data to generate meaningful coefficients. Due to these cases, border-line vectors were either empirically split into chunks of eight equal-sized windows or observed with windows of size 8. The selected list of features (186 in total) corresponds to a set that includes the more relevant features, regarding their performance in classifying waveform signals with similarity to border-lines. Their description is as follows:

- AutoRegressive model coefficients of order 4 (Zhao & Zhang, 2005) over eight equal-sized windows (producing $4 \times 8$ features). For each window, the model coefficients are estimated using the Burg method (Kay, 1988), which estimates the reflection coefficients




**(a)** RGB+Z, 3D view **(b)** Lesion Mask **(c)** RGB, 2D view **(d)** Z, 2D view

**(e)** Border Line

**Figure 4.8:** SKINL2 Dataset melanoma data sample *0059* (a) 3D visualisation, (b) lesion segmentation mask, (c) RGB central view, (d) depth channel "Z" central view with blue-to-light-blue colour bar, and (e) extracted border line Z-values in millimetres.

and enables the reflection coefficients to estimate the AR parameters, recursively. Based on Zhao & Zhang (2005), where model order selection methods are used to determine the model order that provides the best fit in a similar classification problem, order 4 was selected for our model.

- Shannon entropy values (Walden & Cristan, 1998) for the maximal overlap discrete wavelet packet transform at level 4 (L. & Z., 2016) applied to the signal divided into eight windows, resulting in $2^4 \times 8$ features.
- Multi-fractal wavelet leader estimates (Jaffard et al., 2006) of the second cumulant of the scaling exponents and the range of its exponents, which quantify the local regularity, or singularity spectrum as in Leonarduzzi et al. (2010), creating $2 \times 8$ and 10 features, respectively. Wavelet variance measures the variability in a signal by scale (over octave-band frequency intervals), being extracted for each signal over the entire data length, as in Maharaj & Alonso (2014), based on Walden & Cristan (1998). In order to have an unbiased estimate of the wavelet variance, it is necessary to use only levels with at least one wavelet coefficient unaffected by boundary conditions. Our signal (border-line) length and the Daubechies 2 (db2) wavelet (Cohen, 1994) result in the usage of 10 levels. Similarly to Leonarduzzi et al. (2010), the width of the singularity spectrum obtained from multi-fractal 1D wavelet leader estimates was selected as a measure of the multi-fractal nature of the border-line signal. Note that the second order cumulants were selected because they broadly represent the departure of the scaling exponents from linearity – scaling exponents are scale-based exponents describing power-law behaviour in the signal at different resolutions.

**Classification** The classification of the lesions is performed resorting to a SVM model with a polynomial kernel of second order and a box constraint of 1. Given that this is a binary

classification problem, the SVM solver is the Iterative Single Data Algorithm, (which is optimised through a series of one-point minimisations and neither respects the linear constraint nor explicitly includes the bias term) (Kecman et al., 2005a).

In addition, due to the unbalanced nature of the dataset, adjustments to the classifier's cost matrix were also tested. In this matrix, each element consists of the cost of guessing that a sample belongs to class X (lines) when it belongs to class Y (columns), leaving all elements of its main diagonal equal to zero. Therefore, since the melanoma class is three times smaller than the nevus class, the experimental cost matrix adjustments were made to accommodate this discrepancy in unit steps: making a mistake in the melanoma classification have an importance equal to that of a nevus ($[0\,1; 1\,0]$), or have twice ($[0\,1; 2\,0]$), or thrice ($[0\,1; 3\,0]$) the said importance.

### 4.3.3 Results and Discussion

The proposed classification methodology was applied to the publicly available SKINL2 dataset. Like in the previous Section (4.2), the second version of this dataset was used due to its increase in lens magnification of $\approx 30\%$ (which means more detail) in comparison to its first version. For the target classification labels, this dataset currently comprises 9 melanomas and 27 nevi images (MvsN), which undergo the pre-processing, feature extraction, and classification processes described in Section 4.3.2. All experiments were performed using Leave-One-Out Cross-Validation (which effectively results in a 9-fold CV because of the dataset size) and were executed 10 times to mitigate any biased or stochastic decision in the model.

The results achieved in these assessments are evaluated in terms of percentage of classification ACC, SPE, SEN, BAC. To facilitate the reader's understanding, each experiment is first summarised into the BAC metric (Table 4.3) and then only the best results are detailed with the other three previous mentioned metrics (Table 4.4). Metric results above 75% are highlighted with a colour gradient white-to-grey.

Additionally, as mentioned in Section 4.3.2-Data Preparation, data augmentation variations are also shown in the results table "Augmentation" column. Four variations are tested: no augmentation ("—"), flipping the data ("Flip"), balancing-out ("BalanceOut"), or flipping and balancing-out data ("Flip+BalanceOut") during training. The experiments also include tests using a variable number of lines (column "#Lines"), as described in Section 4.3.2-Data Preparation (i.e. whether to use or not the 3 supplementary border-lines). As mentioned in Section 4.3.2-Classification, because of the unbalanced dataset distribution, balancing of the classifier's training was also performed resorting to its cost matrix, as expressed in the tables by column "CostMatrix". Furthermore, each combination of the previously mentioned variations is used four times, since there are four possible ways of adjusting the border-lines length (that is T1/T2/T3/T4, as expressed in Section 4.3.2-Data Preparation).

**Table 4.3:** Summary Results for each experiment using 3D Border-Line Features.

| | | | BAC | | | |
|---|---|---|---|---|---|---|
| #Lines | CostMatrix | Augmentation | T1 | T2 | T3 | T4 |
| 1 | [ 0 1 ; 1 0 ] | — | 93.55 | 50.00 | 90.91 | 72.73 |
| | | Flip | 87.69 | 52.22 | 78.13 | 85.00 |
| | | BalanceOut | 95.00 | 55.00 | 78.13 | 85.00 |
| | | Flip+BalanceOut | 87.69 | 55.00 | 81.94 | 69.95 |
| | [ 0 1 ; 2 0 ] | — | 95.00 | 65.00 | 72.73 | 81.94 |
| | | Flip | 82.15 | 45.00 | 90.91 | 78.82 |
| | | BalanceOut | 70.46 | 45.00 | 63.24 | 74.10 |
| | | Flip+BalanceOut | 72.58 | 65.00 | 69.95 | 70.38 |
| | [ 0 1 ; 3 0 ] | — | 93.55 | 55.00 | 89.71 | 78.82 |
| | | Flip | 82.15 | 58.71 | 70.33 | 81.94 |
| | | BalanceOut | 68.58 | 66.07 | 64.07 | 78.82 |
| | | Flip+BalanceOut | 70.46 | 62.96 | 66.07 | 69.95 |
| 1+3 | [ 0 1 ; 1 0 ] | — | 89.45 | 46.49 | 90.91 | 80.85 |
| | | Flip | 84.74 | 61.11 | 82.75 | 88.90 |
| | | BalanceOut | 70.09 | 36.26 | 71.35 | 71.37 |
| | | Flip+BalanceOut | 71.09 | 55.82 | 69.95 | 78.13 |
| | [ 0 1 ; 2 0 ] | — | 86.38 | 46.49 | 73.81 | 82.15 |
| | | Flip | 83.25 | 44.37 | 80.20 | 80.13 |
| | | BalanceOut | 67.90 | 48.83 | 68.76 | 81.85 |
| | | Flip+BalanceOut | 62.96 | 56.71 | 68.89 | 77.24 |
| | [ 0 1 ; 3 0 ] | — | 76.52 | 48.66 | 71.98 | 78.82 |
| | | Flip | 68.52 | 57.94 | 69.95 | 87.69 |
| | | BalanceOut | 68.52 | 36.16 | 68.81 | 68.33 |
| | | Flip+BalanceOut | 65.17 | 48.41 | 78.17 | 73.13 |

The experimental setup takes, on average, 17 minutes to process the dataset, extract features, and classify using an SVM with 9-Fold CV over 10 executions. There are $2 \times 3 \times 4 \times 4$ experiments of 9-Fold each in Table 4.3 (but only $1 \times 2 \times 4 \times 1$ in Table 4.4), which are run 10 times for variance verification. The highest standard deviation is 4.3, and most executions have 0.0 (zero).

From Table 4.3, the highest and lowest results are achieved with experiments involving line size transformations T1 and T2, respectively, with an average difference of 26.72*pp* BAC between both. As expected, looping the extraction of Z values (along the mask perimeter until the recorded vector reaches the same length of the largest mask perimeter – T2) did not provide proper results because the repetition of information/details introduced large variations on the extracted features (which take into account the signal's structure). Experiment T3 has a similar behaviour to T1 but the linear interpolation seems to degrade the results by 3.41*pp* BAC, on average. Experiment T4, on the other hand (with cubic interpolation), only degrades 0.73*pp*, on average. Therefore the scenario T1, which pads the remaining space with zeros, provides the overall best results.

**Table 4.4:** Detailed Metric Results for the best group.

| | | | T1 (Padding Zero) | | |
|---|---|---|---|---|---|
| #Lines | CostMatrix | Augmentation | ACC | SEN | SPE |
| 1 | [ 0 1 ; 1 0 ] | — | 88.89 | 100.00 | 87.10 |
| | | Flip | 88.89 | 85.71 | 89.66 |
| | | BalanceOut | 91.67 | 100.00 | 90.00 |
| | | Flip+BalanceOut | 88.89 | 85.71 | 89.66 |
| | [ 0 1 ; 2 0 ] | — | 91.67 | 100.00 | 90.00 |
| | | Flip | 86.11 | 75.00 | 89.29 |
| | | BalanceOut | 75.00 | 50.00 | 90.91 |
| | | Flip+BalanceOut | 77.78 | 53.85 | 91.30 |

Balancing the data by repeating samples of the smallest class (*BalanceOut* augmentation) provides a similar result to balancing the SVM cost matrix for the dataset unbalanced data ($[0\,1; 2\,0]$), which is also evident in Table 4.4 (91.67/100.00/90.00, rows 3 and 5).

Still in Table 4.3, the presence of "normal" and "flipped" vectors (*Flip* augmentation) almost always provides worst results. From the observed experiments, the presence of flipped vectors greatly induced homogeneity across the features space, which makes it harder, or impossible, for the classifier to find a proper separation in the data. Therefore, using the two augmentation settings (*Flip+BalanceOut*) had similar outcomes. Depending on the configuration, one of the augmentation modes always induces worse results. Although this could be a misinterpreted effect due to the size of the dataset.

Regarding the number of extracted lines, the use of extra border-lines generally degraded the results, except in T4. In 28 (out of 48) experiments, using more than one border-line, provided worse results in comparison to using a single border-line. This can be explained by the fact that, having used a small dataset, the classifier might not find the best hyper-plane separation and the added samples help nudge it, albeit the added samples are of a different origin thus can also decrease some of the data separability. Note, however, that in T4, in 9 out of 12 experiments the results improved in comparison to the use of only one border-line. Thus, it is more evident that, despite of the data interpolation, the results are improved by using the additional border-lines.

As can be seen in Table 4.4, SPE values range from 87.10% to 91.30%, which are the best results. SPE indicates the ability to correctly reject healthy patients without a condition. This (high SPE) makes it useful for ruling in disease. However, given that the SKINL2 dataset is very unbalanced, it is not easy to obtain balanced SPE and SEN results at the same time. Even so, manually balancing the data through augmentation or repetition seems to help in certain cases. The accuracy performance is 91.67% when forcing sample balancing either by replicating the smaller class or adding more weight to said class misclassifications. It must be noticed that the proposed method exploits only the lesion depth information for feature extraction and

classification (for the melanoma vs nevus – MvsN – problem). However, while showing results similar to other methods reported in the literature, it is not directly comparable because the other methods are based on RGB colour images and, consequently, in other datasets. Additionally, this approach relies on a less invasive technology (does not require physical contact with the patient) that uses light-field cameras instead of dermatoscopy devices.

Other top results, as the $89 - 90\%$ BAC in experiment T3, could also be worth noticing. Because of the behaviour of the BAC metric, factually, any result above 80% can only have SEN and SPE values in the $[61, 100]$ range. While results above 90% need SEN/SPE values in the $[81, 100]$ range. This means that other not-detailed results could compete with state-of-the-art skin lesion classification algorithms (if not for the data modality constraint, which makes direct comparison impossible). On a practical perspective, a passive mechanism even with results of 70% BAC could be appealing, since false positive detection of nevus as melanoma is not completely undesirable. In T1, cost matrix $[0\ 1; 2\ 0]$, the highest SPE of 91.30% is achieved (72.58% BAC), which means that a system would identify 7 out of 9 melanomas, albeit a misclassification rate of 6 out of 27 nevi.

As previously mentioned, this work focuses on showing that skin surface topology has potential discriminative information for melanoma classification and, as such, comparisons with other literature works that resort to RGB only are neither possible or relevant. Additionally, at the time of writing, to the authors best knowledge, there are no works published by other authors resorting to a dataset that provides RGB+Z images.

### 4.3.4 Conclusions

Automated melanoma detection is crucial to help dermatologists to improve their diagnostic accuracy. Among all skin lesion discriminations, classification of MvsN is considered the most difficult, therefore a computer expert system is of utmost importance. As an alternative to recent works, where skin lesion classification is based on dermoscopic images (2D), this section investigated other type of image information, which have been fairly unexplored, e.g., surface (3D) information; to find out whether it could potentially provide better discrimination of melanoma from nevus. Taking advantage of the recently introduced technology of light-field cameras, this work provides a new insight on this domain, being the first one to demonstrate that it is possible to achieve with quite good accuracy the classification of skin lesions, based on multi-dimensional imaging. To this end, the 3D border-lines of the lesion were used to perform a classification with high discrimination. Due to its characteristics, the extracted signal, obtained from the border-lines, can be classified using 1D features.

The achieved experimental results present a discrimination of melanomas against nevi of 95.00% BAC (100.00% SEN and 90.00% SPE). Since these results are comparable with others available in the literature, they provide evidence that skin lesion classification (of melanoma

and nevus) is possible using non-invasive techniques and avoiding the additional artifacts that the use of a dermatoscope (and gel) induces in algorithm pipelines.

Overall, this work provides insight for further research in the field of skin lesion image detection, segmentation, and classification to either improve existing methods/models that are lacking in performance or refine the existing top performers. It is also demonstrated that the extended 3D information enabled by the light-field cameras is useful, beyond conventional texture (2D), to improve lesion discrimination algorithms.

## 4.4 Summary

This chapter focused on creating and describing the SKINL2 dataset – a novel publicly available dataset providing depth information about skin surface – and two studies exploiting the 3D characteristics of the skin lesion surface. In one, a novel approach to this field is presented, exploiting the 3D characteristics of the skin lesion surface, thus advancing beyond common features such as shape, colour, and texture extracted from dermoscopic RGB images. These features were used to train a Bag-of-Features model to distinguish between malignant and benign lesions and discriminate melanoma from all other lesion types. The achieved experimental results indicate the existence of relevant discriminative characteristics in the 3D surface of skin lesions, which allow the improvement of existing classification methods based only on 2D image characteristics.

In a different approach, only the lesions' border-line characteristics are investigated. A selected group of features is extracted from the depth information of 3D images, which are then used for classification. Despite class imbalance often present in medical image datasets, the proposed algorithm achieves high performances while using only depth information for the detection of melanomas. Such results showed that potential gains can be achieved by extracting information from this often overlooked dimension, which provides more balanced results in terms of SEN and SPE than other settings.

The insight drawn from these experiments could foster further research that takes advantage of all the information provided by the light-field cameras, namely embracing depth information with texture information (2D) to improve lesion discrimination algorithms (as performed in Chapter 5).

# Chapter 5

# Towards Melanoma Classification

Content



Lesion identification by specialists is a labour intensive, time costly, and error prone process. Therefore, it could be improved with the use of automated methods. Fortunately, with the advent of DL, computer-aided diagnosis of cancers seems increasingly possible (Litjens et al., 2017). Indeed, automated DL techniques for skin lesion classification may automate future screening and enable early detection of skin cancer (Adegun & Viriri, 2020). However, as detailed in Yao et al. (2021), available skin lesion datasets are usually very small in comparison to what is normally used to train DL models. Therefore, many studies prefer to extract hand-crafted features in order to reduce the model learning space and, consequently, its natural capability to overfit (Yang et al., 2018; Satheesha et al., 2017).

Datasets used in skin lesion classification use the same type of information as dermatology experts, i.e. dermoscopic images (2D/colour). The resulting classification performances are yet to become sufficient to professionally help dermatologists. Despite the limited composition of current datasets, other type of image information could also be used for this end. This includes other data dimensions, which are fairly unexplored as they are not suited for direct human observation, but can still provide relevant information for computer systems. One of these modalities is 3D imaging (e.g., stereo), which has already proven to enhance skin lesion

discrimination performances due to the added depth information (McDonagh et al., 2008; Smith et al., 2011).

Thus, this chapter is dedicated to developments made by using both 2D/colour information and 3D/surface information (depth maps) for melanoma classification. With this aim, two classification approaches were created. The first, presented in Section 5.1, is a step in the direction of merging current 2D state-of-the-art results and evaluated 3D characteristics by using, respectively, an ensemble comprised of a DL model for colour classification and a Multiple Instance Learning (MIL) model for 3D surface classification. The second approach, described in Section 5.2, was created to be a single model capable of performing melanoma discrimination independently of the use of either colour, depth, or both; and allow DL classification even in the presence of small data quantities. Both classification approaches were designed to enable comparison of whether or not 3D information should be used and if features of such third dimension could be beneficial for the classification process. In either classification approach, the target classification labels are: binary discrimination of melanoma versus nevus samples (MvsN); or binary discrimination of melanoma versus all other skin lesion types (MvsAll). Finally, Section 5.3 summarises this chapter and highlights the most relevant conclusions.

## 5.1 Joining 2D Classification and 3D Characteristics

This section's contribution focuses on the proposal of an ensemble model that enables melanoma classification by resorting to 3D surface data when the initial colour classification is uncertain. Both the colour of 2D images and the corresponding depth information are used by resorting to a dataset of light-field skin lesions. Classification of colour or depth information is performed separately. For 2D information, a Transfer Learning (TL) approach (Hosny et al., 2019) comprising a DL model is used. While for the depth information, features extracted from the 3D surface feed a Multiple Instance Learning (MIL) approach when the DL model shows high uncertainty towards classification. Both local and global features are used to characterise the 3D depth surfaces. Feature selection also takes place and it is performed by an automatic feature reduction algorithm, which allows the model to cope with the dataset size.

The remainder of section is organised as follows: Section 5.1.1 presents the literature and background involved in this work. Section 5.1.2 describes the proposed approach and the corresponding pipeline, which comprises an ensemble of two models. Section 5.1.3 describes the first model, including relevant details about model training and classification, and Section 5.1.4 describes the second model, including relevant details about feature extraction and selection, and classification. Finally, Section 5.1.5 presents and discusses the attained results, while Section 5.1.6 highlights the conclusions and future work.

### 5.1.1 Relevant Background

In recent years, the DL paradigm has attracted research in several domains of medical image analysis, demonstrating that noticeable improvements are achieved beyond conventional approaches (Ravì et al., 2016; Shen et al., 2017; Li et al., 2018; Tang et al., 2020). In the field of skin lesion classification, Convolutional Neural Networks (CNN) have also produced promising results (Gonzalez-Diaz, 2018; Tang et al., 2020). In Kawahara et al. (2016), a CNN pre-trained on the ILSVRC is used as a feature extractor (rather than trained from scratch). This work demonstrated that the existing filters (used on the ILSVRC natural images) generalise well for a set of 10 classes using non-dermoscopic images. More recently, research with such pre-trained models reported the highest performance measurements ever published across multiple test datasets (Hosny et al., 2019). The use of pre-trained models is typically accompanied by a TL approach (Shin et al., 2016; Barata et al., 2018), which can be further aided by manually extracted features (e.g., as in Hagerty et al., 2019). In Hosny et al. (2019), classification of segmented colour skin lesions is performed using TL with the pre-trained AlexNet CNN (Krizhevsky, 2014) to achieve high accuracy performance values.

The research addressed in the remainder of this section comprises several concepts to which relevant information is provided. Therefore, the remainder of this section is structured into paragraphs that address such concepts.

**Uncertainty** Sometimes, DL classification results are enhanced with the model's inner statistics, namely the features' distribution that exists before a Softmax layer. If the model's values prior to this layer are not well separated, it might indicate that the model is uncertain about the target label that corresponds to the correct answer, or even if both are equally correct. For this reason, some researchers look for a better solution to replace the Softmax layer (Khan et al., 2019). These model's values can be used to determine network class uncertainties or, for example, the CNN belief in the classification of the segmented pixels (Sensoy et al., 2018; DeVries & Taylor, 2018; Abdar et al., 2020). Such uncertainty values, which exist before the Softmax layer, have been used to improve CNN models (Cho et al., 2020; Abdar et al., 2020; Sensoy et al., 2018). As highlighted in Leibig et al. (2017), further inspection of uncertain decisions results in better performance. Additional research on uncertainty can be found in Abdar et al. (2020).

**Multiple Instance Learning (MIL)** Assuming the calculation and usage of such uncertainties, other models can be used (or combined) to compensate a previous DL model that is uncertain of its classification output. These other models need not be of matching technique but a new combination of multiple models. When a new model depends on multiple outputs of a previous one, such composition is known as Multiple Instance Learning (MIL). This concept was introduced in Keeler et al. (1990), and was later used, in Maron & Ratan (1998), to solve a machine vision scene classification problem.

In Keeler et al. (1990), an instance is defined as one or more fixed-size sub-images of a given image, and the bag of instances is the image itself. An image is labelled positive if it contains a target scene related instance or negative otherwise. For this to work, it is assumed that a relationship between the instances within a bag and the class label of the bag exists, allowing the classification itself to be performed in several ways. For example, given each instance classification, a bag of instances can be given the final label by a thresholding model, by a count-based assumption, by the presence of a single positive or negative class, or by more complex models, like for example a multidimensional-polynomial-border created by a SVM model (Kecman et al., 2005a).

In Wang et al. (2020), the same concept is exploited. The authors proposed a weakly supervised DL framework with uncertainty estimation, in order to address a disease classification problem. Firstly, a CNN instance-level classifier is iteratively refined by using the proposed uncertainty-driven deep MIL scheme. Secondly, a Recurrent Neural Network takes each of the previous instances features (from the same bag/image) as input and generates the final prediction, considering each local instance and their global aggregated representation.

**Segmentation** Most methods dealing with skin lesion classification require some form of prior lesion segmentation or region identification (Hosny et al., 2019). Several previous works present some form of skin lesion segmentation to prepare the data for classification, such as Hagerty et al. (2019); Gonzalez-Diaz (2018); Tang et al. (2020); Li et al. (2018); Ravì et al. (2016); Barata et al. (2018). This preprocessing step is typically needed since skin information (or image acquisition artefacts) can produce outlier features or expand the dimension of the hyperspace in which the parameter search is performed by DL algorithms (as, for example, with CNN), urging for a preprocessing step in order to avoid undesirable outcomes. A relevant example of such method is described in Navarro et al. (2018), where the image is segmented into super-pixels using local features and then iteratively merged into regions to form two classes of regions (lesion and non-lesion), while considering a spatial continuity constraint on the super-pixels colour.

**Dataset** To the best of the authors' knowledge, all published literature works that use publicly available datasets operate on 2D datasets, either of dermoscopic or macro images. Hence, the most common type of existing features comes from the same 2D modality. Although significant performances have already been achieved using these single modality datasets (Pathan et al., 2018), the low granularity of the information might still pose limitations to the classification problem, as only planar lesion-information can be retrieved from such data.

To overcome this limitation, alternative modalities, such as using stereoscopy technology (McDonagh et al., 2008; Smith et al., 2011), have already shown to be efficient in identifying the type of skin lesion when a third dimension is present. Even so, literature on 3D surface of melanoma or related skin lesions is still very scarce. Nevertheless, existing research indicates that proper results arise when using depth information (3D), like the study in Satheesha et al. (2017), that artificially generated 3D information to enhance an existing 2D dataset. In order

to fill the void of 3D skin lesion data, a dataset named Skin Lesion Light-fields (SKINL2) was made public to enable research over skin lesions 3D surface information (Faria et al., 2019c).

**Hand-crafted Features** As mentioned before, most works in the literature rely on 2D datasets, that either extract hand-crafted features for melanoma classification or, more recently, use DL or TL to automate the process. Some of these hand-crafted features include: lesion type and configuration (primary and secondary morphology), colour, distribution, shape, texture, and border irregularity (Mahmouei et al., 2018; Korotkov & Garcia, 2012; Pathan et al., 2018). After the feature extraction step more automated ML methods such as K-Nearest Neighbours, ANN, Logistic Regression, Decision Trees, and SVMs are used to perform classification, typically with no more than moderate success (Korotkov & Garcia, 2012; Pathan et al., 2018). Hence, the literature transition in recent years to more rewarding DL methods which relieve the research on new features. Examples of related work using 2D hand-crafted features and known classifiers can be found in Korotkov & Garcia (2012); Barata et al. (2018).

So far, there are no 3D features specifically studied for melanoma classification. Thus, a primary approach towards defining a relevant set of such features is to look at other research fields, where 3D features have been used. Depending on the target recognition task, several 3D features have been developed and generalised across multiple 3D datasets and tasks. This type of generalisation is performed to propose a set of features that capture a broad spectrum of 3D characteristics – typically applied to key regions. In general, an algorithm responsible for extracting the designed features is called feature extractor and the key regions where these feature extractors are applied are determined by a keypoint detector. In the scope of this work, the Normal Aligned Radial Features (NARF, Steder et al., 2011) is used as both a keypoint detector and feature extractor. Other relevant feature extractors are the following:

- Radius-based Surface Descriptor (RSD, Marton et al., 2010),
- Global RSD (GRSD, Kanezaki et al., 2011),
- Globally Aligned Spatial Distribution (GASD, Lima & Teichrieb, 2016),
- Rotation Invariant Feature Transform (RIFT, Lazebnik et al., 2005),
- Point Feature Histogram (PFH, Rusu et al., 2008),
- Fast PFH (FPFH, Rusu et al., 2009a),
- Signature of Histograms of OrienTations (SHOT, Tombari et al., 2010b),
- Ensemble of Shape Functions (ESF, Wohlkinger & Vincze, 2011),
- 3D Shape Context (SC3D, Frome et al., 2004),
- Unique Shape Context (USC, Tombari et al., 2010a).

**Feature Selection** In many cases, the initial number of features can be overwhelming for the classification algorithm, particularly when the number of data samples is not enough to enable a correct understanding of all feature space combinations. Thus, feature reduction is necessary to select the most meaningful ones, which can be done by using several methods, such as, a diagonal adaptation of Neighborhood Component Analysis (NCA, Yang et al., 2012). NCA is a non-parametric algorithm that enables feature selection with the goal of

maximising the prediction accuracy of regression and classification algorithms. The algorithm performs better when estimating feature importance for distance-based supervised models that use pairwise distances between observations to predict the response. NCA can be understood as a pre-processing step before the classification step, as in Jiménez et al. (2019); Artzi et al. (2019); Rahman et al. (2019), allowing the removal of similar or noisy features from the feature space. But it can also be used between models (Akram et al., 2018a), namely when initial DL models produce too many latent features in comparison with the amount of available data samples (Akram et al., 2018b).

### 5.1.2 Proposed Multi-Instance Learning Classification Approach

As pointed out before, the proposed approach explores depth information (Z), in addition to conventional colour (RGB), to improve beyond current classification results. To this end, a new pipeline was devised (as summarised in Section 5.1.2-Overview), to operate over a dataset with lesion segmentation masks (generated as described in Section 5.1.2-Segmentation). This pipeline utilises both a DL process, as a baseline 2D classification model (Section 5.1.3), as well as a two-step model scheme that resorts to hand-crafted features from the 3D surface (Section 5.1.4). This is an ensemble classification approach, where the objective is to collectively obtain better predictive performances than those from any of the individual learning algorithms on its own.

**Overview** An overview of the pipeline is depicted in Fig. 5.1. Given a 4D dataset, with its lesion segmentation masks, at any 10-fold CV partition $k$, a $\text{Train}_k$ and $\text{Test}_k$ datasets are received by the ensemble pipeline. As training precedes the test step, the $\text{Train}_k$-set is first used to train both a TL model and a MIL model, prior to the use of the $\text{Test}_k$-set.

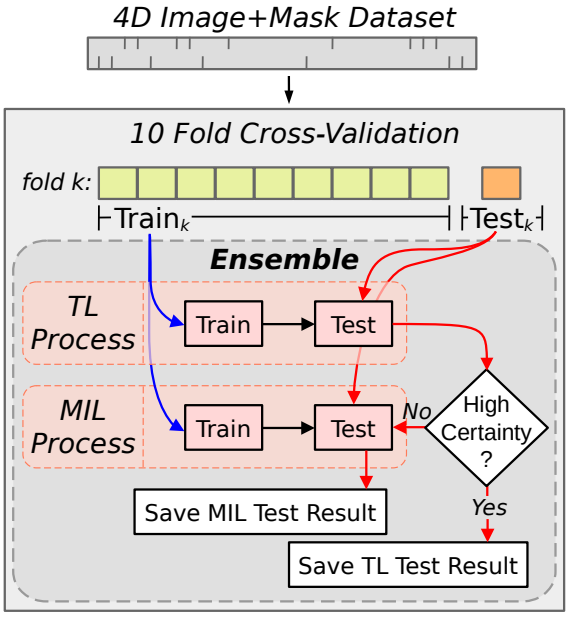



**Figure 5.1:** Proposed Ensemble Pipeline: a given dataset is partitioned into 10 folds for cross-validation; at any stage, both TL and MIL are trained on 9 training folds and later tested on 1 test fold; afterwards, if the TL process certainty is high, the TL test image classification result is recorded, otherwise the MIL classification result is recorded. Blue arrows indicate the pipeline training sequence. Red arrows indicate the pipeline testing sequence. Black arrows indicate previous dependencies and/or abstract progressions through the pipeline.

TL is performed with a DL model to update its weights to the classification problem at hand. The other part of the ensemble classifier (MIL) comprises a two-step learning approach. The Softmax layer present in the CNN model allows to predict the level of confidence the CNN has in its prediction, which is known as the model certainty. It can be asserted either naively or by imposing alternative computations. Therefore, if the CNN 2D classification model is certain of its prediction it is set as the ensemble prediction, otherwise, the MIL 3D-classification model is preferred.

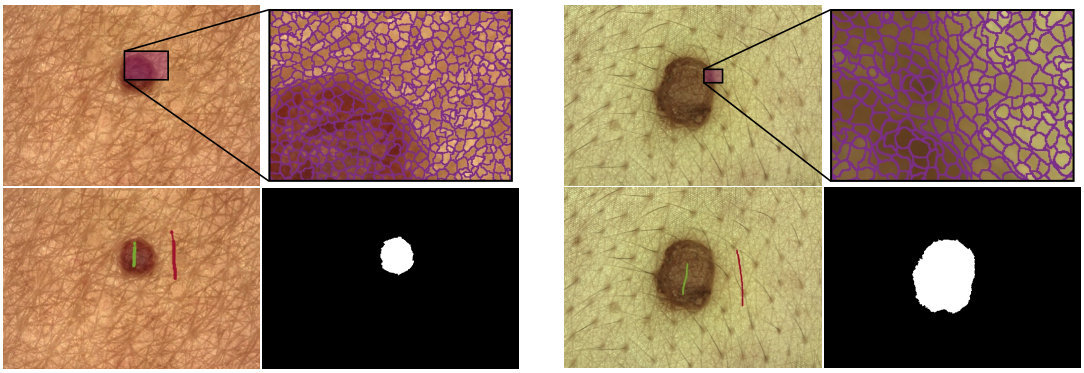
**Segmentation** The employed segmentation method is based on a modified version of the algorithm described in Li et al. (2004), dubbed Lazy Snapping, which resorts to an internal method to group similar pixels. However, in this work, such method is replaced by a more recent approach named Simple Linear Iterative Clustering (SLIC), described in Achanta et al. (2012), which has been observed to achieve good performance in coloured images of skin lesions in Navarro et al. (2018).

Given an RGB coloured image (Fig. 5.2, top-left), pixels are first grouped into super-pixels (Fig. 5.2, top-right) using the SLIC method. This method serves as a pre-processing step for the Lazy Snapping algorithm, as it compacts the problem dimension to less samples (super-pixels). In this work, the SLIC compactness is set to 10 and its clustering phase is performed for 10 iterations. Then, the Lazy Snapping algorithm constructs a graph of the image super-pixels, where each super-pixel is a node connected by weighted edges. The higher the probability that pixels are related, the higher the weighted edge. The algorithm cuts along weak edges, achieving the object segmentation by maximising the colour similarity within the object. To generate the necessary binary segmentation mask, that separates foreground from background, the graph-cut is guided with user provided information (Fig. 5.2, bottom-left) about pixels belonging to the lesion (foreground, green points in the figure) and pixels belonging to the non-lesion skin (background, red points in the figure).

Given the user input, the separation between foreground object and background elements is generated by the Lazy Snapping algorithm, as a segmentation mask (Fig. 5.2, bottom-right).

### 5.1.3 TL Process

A model named AlexNet is created with the ILSVRC pre-trained weights and, as such, it only receives 2D RGB images. The process is performed for 32 epochs of batch size 10, with a learning rate of 0.001. The colour images first undergo a segmentation process (described in Section 5.1.2-Segmentation), where the non-lesion skin is coloured black, effectively removing colour information and forcing the CNN to focus on the RGB characteristics of the target lesion area. Data augmentation (by rotation) is also performed, as described in Hosny et al. (2019), but only during training and not in the testing stage.



**(a)** Angioma, SKINL2 v2 **(b)** Nevus, SKINL2 v3

**Figure 5.2:** SKINL2 Lesion Segmentation Method: given a coloured central-view image (a) or (b) dataset (top-left); the image pixels are grouped through superpixel over-segmentation (top-right); then, visually, some pixels regarding the lesion (in green) and skin region (in red) are marked to help guide the segmentation process (bottom-left); lastly, a skin lesion segmentation mask is generated (bottom-right).

Additionally, in the present work an uncertainty-enhancement is used. Instead of naively using the internal Softmax probabilities for the ensemble model uncertainty, the model is reinforced with the capability to generate its internal classification certainty during training. Thus, the loss function is changed from the default Softmax cross entropy to the sum of two components (Sensoy et al., 2018), as expressed by

$$\text{loss} = \frac{\|\text{v}_1\|_2^2 + \|\text{v}_2\|_2^2}{2} \times 0.005 + \text{UIF}(\text{e}^{\text{o}}, \text{a}), \tag{5.1}$$

where $\|X\|_2^2$ represents the L2-norm, defined as $\frac{1}{2}\sum x_i^2$, where $x_i$ are the elements of the vector $X$, $v_1$ and $v_2$ are the outputs of the first and last classification layers, $o$ are the values at the end of the network, and $a$ the target classification one-hot label probabilities. The mean-square-error uncertainty-infused function (UIF), is expressed as

$$\text{UIF}(\text{b}, \text{a}) = (\text{b} - \frac{\text{a}}{\text{s}})^2 + \frac{\text{a} \times (\text{s} - \text{a})}{\text{s}^2 \times (\text{s} + 1)} + \text{KL}(\text{P}(\text{b}, \text{a}) \| \text{Q}), \tag{5.2}$$

where $s$ is the sum of all one-hot exponential values and KL is the Kullback-Leibler divergence term, defined as $\text{KL}(\text{P}' \| \text{Q})$, where $P'$ is the result obtained from applying Eq. (5.3) and $Q$ is the one-hot distribution.

$$\text{P}(\text{b}, \text{a}) = (\text{a} - 1) \times (1 - \text{b}) + 1 \tag{5.3}$$

The KL divergence is used in this context to regularise the predictive distribution by penalising predictions that diverge from the desired uncertainty, which is known as Learned Loss Attenuation (Sensoy et al., 2018).

At the end of the Alexnet uncertainty-infused-model training stage, a classification uncertainty for each class can be obtained by dividing the number of possible output classes by the exponential of the values outputted by the network. For the proposed ensemble, a classification certainty above 50% is considered high.

Both the classification labels and uncertainties are output to the ensemble definition described in Section 5.1.2-Overview.

### 5.1.4 MIL Process

The MIL process performs skin lesion classification using only 3D surface information. A detailed pipeline of this process is depicted in Fig. 5.3. Note that the correct dependency-flow starts with the training stage (blue-arrows), which might initiate black-arrow flows. Any procedure is only executed if all input training flows (arrows) are present or if it has already been executed for training.

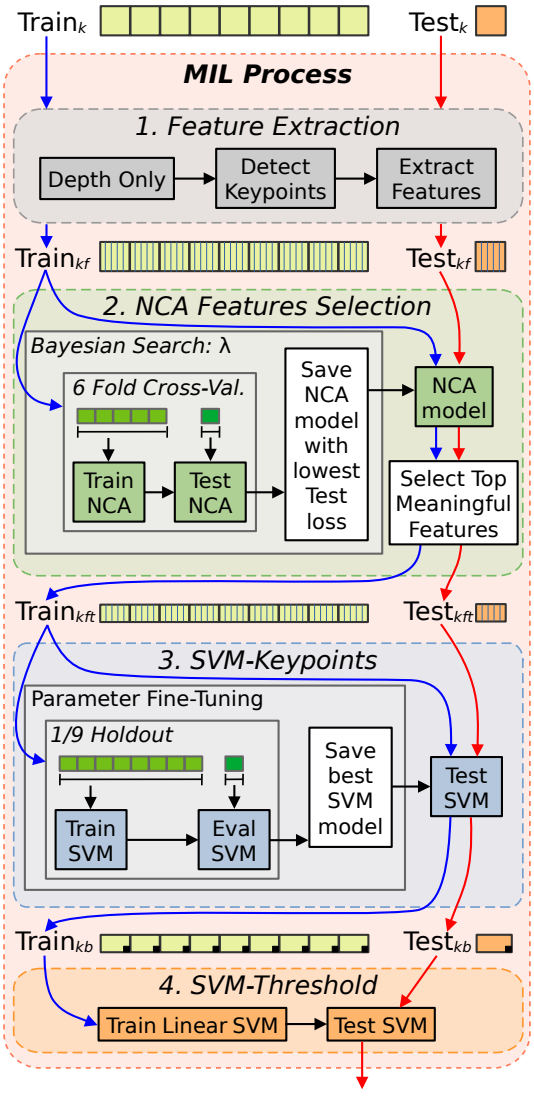



**Figure 5.3:** MIL Process Pipeline, comprising four blocks: *1)* given an image dataset, only depth information is kept and features from detected image keypoints are extracted; *2)* given a keypoints training-set, an NCA model is created with the lowest possible loss. This is obtained by performing a Bayesian search over a 6-fold cross-validation of the train data to find the NCA optimal $\lambda$ value. Given the NCA model, the algorithm advances to next block with the top meaningful features; *3)* given the selected features, a fine-tuned SVM model for keypoint classification is created. This tuning occurs through a parameter search using 1-out-of-9 folds for evaluation of said SVM model, and the SVM keypoint classification labels bagged by image advance to the next block; finally, *4)* given a dataset with bags of labels, a linear SVM is trained to provide the grouped image final classification label. Blue arrows indicate the pipeline training sequence. Red arrows indicate the pipeline testing sequence. Black arrows indicate previous dependencies and/or abstract progressions through the pipeline.

The process comprises four main blocks, each being executed only after the previous one's completion. Blocks named 1 and 2 comprise the dataset pre-processing stage with feature extraction and selection, while blocks 3 and 4 comprise the actual MIL aspect of the process. Detailed information about each block is provided in the following four subsections.

**Feature Extraction** Given either a training or a test-set of 4-channel images (RGB+Z), pixel values in the RGB channels of all input images are replaced with zeros. This operation is performed to guarantee that no colour related feature is generated, meaning that further processing only uses depth. Having only the 3D surface, the NARF keypoint detector elects several key locations in each image. Using the lesion masks (as described in Section 5.1.2-Segmentation), after a dilation process to extend each mask by 25 pixels, keypoints not belonging to the new lesion region are discarded. Then, feature extractors are applied to each of the remaining keypoint locations. In essence, this block generates a new $\text{dataset}_{kf}$ from the input set of images, where each image is now represented by multiple instances (keypoints) of multiple features.

The NARF keypoint detector was selected because it seems specially suited for skin lesion images, since it selects the surface locations where abrupt changes occur and takes object borders into account, such as skin-to-lesion borders, which have already been recognised as relevant information (Pereira et al., 2020b). The keypoint detector has two major characteristics. First, keypoints are extracted in areas where the direct underlying surface is smooth and the neighbourhood contains major surface changes. The resulting keypoints are located in the vicinity of significant geometric structures and not directly on them. Second, NARF takes object borders into account. Such objects are detected when non-continuous transitions from the foreground to the background arise. Thus, the silhouette of an object has a strong influence on the resulting keypoints.

As for the feature extractor methods, 11 are utilised, generating a total of 5726 features per keypoint. These extractors were selected based on the relevance of their characteristics for the type of input signal in use, i.e., 3D information. RIFT (32 features) was selected because it provides invariance to illumination, viewpoint, scale, and rotation. Like RIFT, NARF (42 features) and PFH/FPFH (125/33 features) also possess some of these characteristics, PFH/FPFH, in particular, provides robustness against outliers and noise. Other features extractors as SHOT (361 features), SC3D (1989 features), and USC (1969 features) also provide robustness against noise. Additionally, both SHOT and USC are reported to provide uniqueness amongst detection, as well as unambiguous representations. Finally, ESF (640 features), RSD (2 features), GASD (512 features), and GRSD (21 features) were selected for being descriptive, simple, and intuitive shape descriptors. ESF has proven to be efficient and expressive, while GRSD adds expressiveness to the simple RSD by partitioning the image point cloud into several voxel-surfaces of understandable shapes.

**NCA Feature Selection** Given a (training) feature dataset, feature reduction is performed resorting to an NCA model. This is done because some of the extracted features might not contribute for the adequate label separation during later classification process.

Since NCA uses internal mechanisms similar neural networks as part of its creation process, it is possible that the generated feature's meaningfulness-weight is overfitted to the training data. To overcome this problem, NCA includes a regularisation parameter $\lambda$ that helps to

prevent overfitting. Since this parameter has to be predefined, a Bayesian search had to be performed to find the $\lambda$ value that yields the lowest average test loss of a six-fold CV partitioning scheme of the given (training) features dataset. This inner CV is implemented to further prevent data overfit.

Having found the NCA model with the optimal $\lambda$, the (training) features dataset can now be reduced to the most meaningful features. Meaningfulness-weights obtained from the training data can be applied to later testing-sets. In this model, only features with a normalised absolute meaningfulness greater than 0.02 are selected – meaning that features with meaningfulness-weights below $2pp$ are discarded. In essence, this block generates a new $\text{dataset}_{kft}$ from the feature dataset, where only features relevant to classification are maintained.

The implemented Bayesian search is performed by constraining $\lambda$ values to the range $[0.00001, 0.1]$, using four initial seeds randomly chosen from the $\lambda$ search range. This search is executed for 50 steps, comprising 24 evaluations each. To promote a balance between the search exploitation and exploration (Gelbart et al., 2014), the Bayesian propensity to explore is 0.5. In addition, to avoid over-exploiting, the acquisition function in Gelbart et al. (2014) is modified as suggested in Bull (2011).

As for the NCA model parameters, the inner network is optimised using Stochastic Gradient Descent and an initial learning rate is determined by selecting 200 random dataset samples and training a temporary model on increasing learning rates for 15 epochs. The learning rate providing the lowest loss is selected as the initial learning rate (on average, the initial learning rate is 51.20000). With the initial learning rate defined, the network is trained using all training data (five-folds) over 10 epochs with a mini-batch size that enables at least 40 iterations per epoch. At each epoch, the learning rate is decreased when a convergence tolerance of step size 0.000005 is met.

Since there is only one regularisation parameter ($\lambda$) for all weights, and the weight magnitudes must be comparable, i.e. within the same range, any dataset data entering the model is normalised with zero mean and unitary standard deviation.

**Keypoint-level Classification** Given a (training) dataset of meaningful-features, which comprises multiple instance data (keypoints) for each image, label classification of each image keypoint takes place resorting to a SVM. As SVMs have several hyper-parameters, parameter fine-tuning is necessary at this stage. Due to the complexity of the pipeline, the theoretical determination of the optimal SVM implementation is not feasible. Assuming an initial data partitioning into 10 folds CV (as detailed in Section 5.1.2-Overview), the training data comprises nine folds. Therefore, the last fold is hold-out from the SVM classification training process, so that it can be later used for the SVM selection during the parameterisation fine-tuning. This single fold is called evaluation fold. Having found the SVM model with the best performance in the evaluation fold, the same model can be applied to later testing-

sets. In essence, this block generates a new dataset$_{kb}$ comprising a bag of classified keypoints, that is, a label classification for each keypoint of each image. During pipeline training stage, classification label results from both training and evaluation folds advance to the next block as one training-set – i.e., maintaining the original dataset data sample counts. Evaluation results will not be perfect, but this is helpful during the next pipeline block training stage as it provides behavioural insight of how the model operates on unseen data.

As for the SVM parameter fine-tuning, instead of using a full Bayesian search, several predefined parameters were evaluated for simplicity. The SVM kernel function can be either linear, polynomial, or Gaussian. In the case of polynomial, it can be either of order 2 or 3. In the case of Gaussian, it can be either of kernel scale 0.9, 3.6, or 14. Data normalisation always takes place and the box constraint is set to 1. This enables the evaluation of six different SVM models in total. The quadratic kernel SVM is typically the top performing.

The SVM solver is the Iterative Single Data Algorithm (ISDA, Kecman et al., 2005a), given that this is a binary classification problem. In addition, the SVM also comprises a custom cost matrix, which is set to $[0\ 1; 2\ 0]$ to enforce a double penalty when missclassifying the melanoma class. In this matrix, each element consists of the cost of guessing that a sample belongs to class X (lines) when it belongs to class Y (columns), leaving all elements of its main diagonal equal to zero.

**Image-level Classification** Finally, given bags of labels, a last SVM model provides the image-level label classification. Since the objective is to reduce a variable-sized list of keypoint-level labels to a single image-level label, the data is summarised to enable thresholding. That is, given an arbitrary number of data samples belonging to an image, the data is transformed into two sums: the number of melanoma labels and the number of non-melanoma labels. Then, these sums are normalised to the $[0, 1]$ range, while making their sum 1, producing a probability distribution over predicted output classes, as occurs in a Softmax layer. Furthermore, these probabilities are given as features to a SVM model of linear function and ISDA solver, with box constraint set to 1, and without implicit data standardisation. In a training pipeline, this effectively produces a threshold along the probability distribution that attempts to separate the target class labels. A SVM is used rather than a common thresholding technique, due to its capability for better forming the threshold boundary and also because it would enable future work beyond binary classification. As in the keypoint-level classification, the SVM cost matrix is adjusted to enforce a double penalty when missclassifying the melanoma class ($[0\ 1; 2\ 0]$).

In a testing pipeline, the linear SVM model image-level labels are sent to the ensemble, as described in Section 5.1.2-Overview.

### 5.1.5 Results and Discussion

The performance of the proposed method is evaluated and discussed in this section, encompassing two classification experiments, both executed applying 10-fold CV, as previously mentioned. The first experiment, consists in MvsN, i.e., a more difficult task, and the second experiment covers classification of MvsAll.

**Dataset** The proposed pipeline was applied to the publicly available SKINL2 dataset. Particularly in this work, the second and third versions of this dataset were used. Both versions are present due to their increase in lens magnification of approximately 30% (which means more detail) in comparison to its first version. At the time of writing, the third version was still under development and the available data was used as an extension of the second version. In total, 98 images were used (70 from the second dataset and 28 from the third). These images comprise 14 melanomas, 36 nevi, and 48 other lesion types (16 angiomas, six basal cell carcinomas, one dermatofibroma, 24 seborrheic keratoses, and one verruca). All images undergo the pipeline described in Section 5.1.2. Therefore, experiment MvsN comprises 14 melanoma samples against 36 nevus samples, while experiment MvsAll comprises 14 melanoma samples against all other 84 non-melanoma samples.

**Feature Selection** In the pipeline described in Section 5.1.2, the MIL process is responsible for performing the classification when the TL process does not have enough certainty. The feature selection performed within fold samples in this step is a key component of the former process. Depending on the fold, different dataset samples arrive at *NCA Feature Selection* block (Section 5.1.4-NCA Feature Selection), which in turn will induce different features to be marked as meaningfully in different folds for the classification objective.

Table 5.1, comprising five major columns, provides some statistics regarding feature selection. For each feature extractor in the first column, the number of inner features comprising said extractor is shown in the second column. Subsequent columns are sub-divided to provide information for either the MvsN or the MvsAll experiment, respectively. Across the 10-fold execution, the number of unique features that are selected at least once are defined in the third column, while the total amount of features (regardless of repetition) selected across folds is presented in the fourth column. Finally, the fifth column indicates how many times a feature extractor is used, that is, if any of its features were used in any given fold.

This table shows that most literature features considered potentially relevant for melanoma surface discrimination are not selected. This can be considered a normal behaviour since features with higher discriminative power overshadow the lesser ones, making the NCA model algorithm reduce their meaningfulness to marginal values. This occurs as they do not present added information to the higher representative features.

**Table 5.1:** Features Outputted in Feature Selection block.

| Feature Extractor Name | Features Inside Extractor | Unique Features Selected†‡ | | Total # Features Selected†‡ | | # Times Extractor is Used†‡ | |
|---|---|---|---|---|---|---|---|
| ESF | 640 | 28 | 26 | 126 | 92 | 10 | 10 |
| FPFH | 33 | 0 | 0 | 0 | 0 | 0 | 0 |
| GASD | 512 | 35 | 32 | 200 | 149 | 10 | 10 |
| GRSD | 21 | 2 | 2 | 14 | 13 | 9 | 7 |
| NARF | 42 | 0 | 2 | 0 | 2 | 0 | 1 |
| PFH | 125 | 0 | 0 | 0 | 0 | 0 | 0 |
| RIFT | 32 | 2 | 3 | 7 | 17 | 5 | 9 |
| RSD | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| SC3D | 1989 | 0 | 0 | 0 | 0 | 0 | 0 |
| SHOT | 361 | 0 | 0 | 0 | 0 | 0 | 0 |
| USC | 1969 | 0 | 0 | 0 | 0 | 0 | 0 |

† Values resulting from the aggregation of 10-fold cross-validation.

‡ Values for *MvsN* and *MvsAll* results, respectively.

It is also possible to observe that only the ESF, GASD, GRSD, and RIFT feature extractors are selected across the two experiments, with NARF being used in only one fold of the second experiment (MvsAll). Concerning the feature extractors, it can be seen that, if the uniquely selected features were always the same across folds (third column), then the total amount of features selected (fourth column) would be 10× that value, which is never the case. However, this does not mean that no feature is meaningful enough to be selected across folds.

For the five selected feature extractors, Fig. 5.4 presents the number of times each feature extractor is used across folds (bar plot representing the same information as in Table 5.1), as well as the number of times each feature extractor's feature is used (scatter plot). From this figure, it is possible to observe that: in the MvsN experiment, four ESF and six GASD features are always selected (i.e. having usage count equal to 10) independently of the fold data, while in the MvsAll experiment, only two ESF and two GASD features are always selected. This suggests that discrimination between melanoma and nevus is possible in more ways than in melanoma versus every other class, as evidenced by the scatter plot's data-points spread. Also, in the second experiment, the NCA model algorithm excluded some features while adding others, namely including two features from NARF in one fold, as previous mentioned. All in all, from one experiment to the other, a total of 50 features change from either being or not being used in the experiment pipeline, while 40 remain in usage at least once. On average across folds, the feature selection block chooses $33.8 \pm 4.8488$ features in the MvsN experiment, and $28.1 \pm 4.5080$ features in the MvsAll experiment.

**Results** In this subsection, the results are presented in terms of percentage of classification ACC, SPE, SEN, and BAC, where SEN represents the successful melanoma identification rate and SPE the successful identification of the other class.

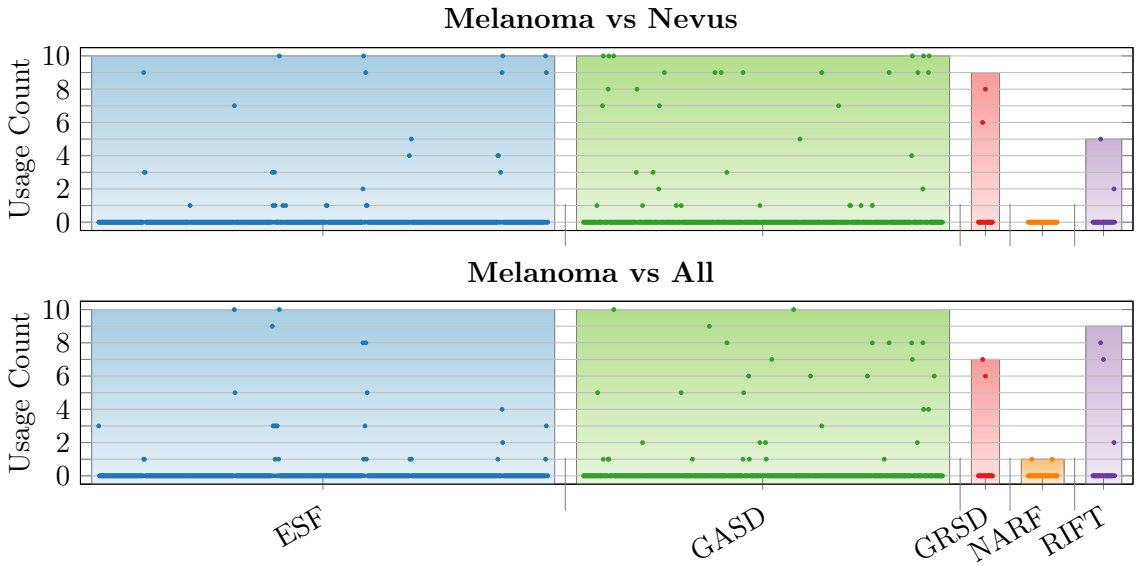



**Figure 5.4:** Number of times that features (or feature extractors) are selected during the 10-fold CV process. Scatter plot values indicate how many times a given feature (from a feature extractor) is marked meaningful for classification during the feature selection block. Bar plot bars indicate how many times a feature extractor is meaningful for classification (.i.e having any of its features selected for usage in a fold) during the feature selection block. Only features extractors which had any meaningful features for classification are displayed.

**Table 5.2:** Ensemble Experimental Results.

| Experiment | Model | ACC | SEN | SPE | BAC |
|---|---|---|---|---|---|
| MvsN | TL-naive | 68.00 | 21.43 | 86.11 | 53.77 |
| MvsN | Proposed Ensemble | **84.00** | **71.43** | **88.89** | **80.16** |
| MvsAll | TL-naive | 73.47 | 14.29 | 83.33 | 48.81 |
| MvsAll | Proposed Ensemble | **90.82** | **78.57** | **92.86** | **85.71** |

naive: as in Hosny et al. (2019), without explicit uncertainty.

As detailed in Section 5.1.2, the proposed ensemble model is comprised of two processes: TL and MIL. The TL process includes alterations to enable the classification uncertainty to be determined in a non-naive manner. Therefore, the TL model without the mentioned uncertainty calculations is referred to as "TL-naive", which corresponds to the effective implementation of the method described in Hosny et al. (2019), providing the literature baseline classification result in both experiments. As can be observed in Table 5.2 the ACC performance of the TL-naive is of 68.00% and 73.47% for the MvsN and MvsAll experiments, respectively. While the ACC increases in the experiment with more data (which has 48 additional samples in comparison with MvsN), it is important to point-out that the SEN metric decreases by 7.14*pp*, even though the number of melanoma samples is the same (14) in both experiments. This decrease represents one melanoma misclassification. The SPE metric is not comparable between both experiments, since the amount of samples differs between experiments. Across folds, TL-naive identifies 31 out of 36 nevus in the first experiment, and 70 out of 84 non-melanoma lesions in the second experiment.

In the MvsN experiment, TL-naive incorrectly classifies 16 samples in the testing stage. Per-

forming naive uncertainty calculations with the TL-naive model (using the internal Softmax probabilities) enables the identification of nine potential misclassifications. Among these, only six are actually misclassifications, while three were originally correct. If the uncertainty-enhancement is performed during training, the (TL) model incorrectly classifies 17 samples in the testing stage, but enables the correct identification of 11 (instead of six) of these misclassifications (while also incorrectly identifying one sample that was actually correct). This improvement to the TL uncertainty identification enables MIL, which has a 72.00% ACC and 78.57% SEN, to potentially correct or disregard said misclassifications (as described in the proposed ensemble pipeline, Section 5.1.2-Overview). From the TL-uncertain-classifications, MIL corrects 10 out of 11 misclassifications and only wrongly changes one sample that was originally correctly identified, although with low certainty, improving from the TL initial performance to 84.00% ACC, as shown in Table 5.2 for the "Proposed Ensemble".

In MvsAll, the detailed observations are similar to the previous experiment. The TL-naive incorrectly classifies 26 samples in the testing stage from which the naive uncertainty calculations enable the identification of 12 potential misclassifications – 10 comprising actual misclassifications and two originally correct. If trained with the uncertainty-enhancement, the (TL) model incorrectly classifies 28 samples – but potentially enables the correct identification of 21 (instead of 10) misclassifications (while also incorrectly identifying three samples which were actually correct). As with the previously detailed-experiment results, this improvement to the TL uncertainty identification enables MIL, which has a 51.00% ACC and 71.43% SEN, to potentially correct or disregard the misclassifications. From the TL-uncertain-classifications, MIL corrects 20 out of the 21 misclassifications and incorrectly classifies one of the three uncertain (but correctly classified) samples – improving from the TL initial performance to 90.82% ACC, as shown in Table 5.2 for the "Proposed Ensemble".

In this section, all comparisons with the baseline classification results obtained with TL-naive have shown that the proposed ensemble method provides superior performance results. This can be interpreted as an indirect comparison with the works considered in Hosny et al. (2019) and other works that resorted to the same dataset and metrics as Hosny et al. (2019). In essence, since TL-naive (Hosny et al., 2019) reports results superior to 11 other methods, it indicates that the proposed ensemble method should present a higher performance than these methods as well. It is possible to apply this bias in relation to other published works, as Pereira et al. (2020b); Tang et al. (2020); Barata et al. (2018); Pathan et al. (2018); Hagerty et al. (2019), which are applied to the same datasets and use the same metrics as Hosny et al. (2019).

### 5.1.6 Conclusions

Automated melanoma detection is crucial to help dermatologists improve their diagnostic accuracy. Still, even with DL methods, current systems are yet to achieve satisfactory sensitivity performances. Instead of continuously attempting to improve algorithms with available

colour (2D) datasets, which are commonly used by dermatology experts, new dimensions and modalities may be explored as, for example, surface (3D) information; which can potentially provide new melanoma discrimination capabilities. In order to advance beyond current state-of-the-art results, more reliable solutions might depend on the joint exploitation of both 2D and 3D information. Taking advantage of the recently introduced technology of light-field cameras, the main contribution of this work is to exploit depth information, in addition to colour, for classification of skin lesions using a recent dataset of multi-dimensional imaging, which was specifically acquired to provide richer information for image classification. Accordingly, this research aims to build a model that takes advantage of the recent scientific advances in both 2D and 3D modalities. As a result, this work is the first to incorporate DL uncertainty evaluation mechanisms with Multiple Instance Learning for the training of a robust synergistic ensemble classifier with the intent of performing skin lesion classification using light-field imaging.

Targeting the melanoma class with this model, despite the large class imbalance – often present in medical image datasets – and limited data samples, the ensemble model achieves a cross-validation ACC of 84.00%, with 71.43% SEN and 88.89% SPE. These results account for the classification against nevus lesions and show an ACC increase of 16.00*pp* (supported by a SEN increase of 50.00*pp*) from the baseline method applied to the SKINL2 dataset. In a more challenging setting, discrimination of melanomas against all other available skin lesions was achieved with 90.82% ACC, 78.57% SEN, and 92.86% SPE, with a similar ACC increase of 17.35*pp* from the baseline, also supported by a SEN increase of 64.28*pp*. The performed experimental assessment allows to extrapolate that melanoma skin lesion classification can be improved by including 3D information, such as surface depth.

In the presence of untrustworthy 2D features, the achieved results indicate that the 3D surface provides redeeming results, showing that improvement of existing methods is still possible when looking beyond 2D image characteristics.

## 5.2 Melanoma Classification with Morlet Scattering Transform

In general, image classification requires the use of representations that reduce non-informative intra-class variability and yet preserve discriminative information across classes. In DL, deep neural networks (DNN) build hierarchical invariant representations learned by applying linear and non-linear operators in succession during training. These are learned in a dataset-dependent basis, however most image classification problems have generic learnable representations that are common across fields. When multiple instances of the same element are present in a dataset, translations, rotations, and scaling are common sources of variability for most images. Changes in the object view point and perspective projections of three dimensional surfaces correlate many of the dataset samples. With the use of Wavelet Scattering (WS) (Bruna & Mallat, 2013), it is possible to build neural networks invariant to these trans-

lations and rotations (Sifre & Mallat, 2014). These can be implemented as a convolutional neural network (CNN) with successive spatial wavelet convolutions at each layer.

This section explores the use of depth data from skin lesions combined with colour information by resorting to light-field images and semi-automated segmentation masks. Based on a publicly available dataset named SKINL2 (Section 4.1), a DL-based classifier is developed. The DL model relies on Morlet Wavelet-based features that greatly reduce the dimensionality problem by performing Wavelet Scattering Transforms on the input data (Andén & Mallat, 2014; Bruna & Mallat, 2013; Sifre & Mallat, 2013). These features are used as an alternative to a deeper model by providing unique features invariant to translation, rotation, scale, and frequency shifting – a transformation bearing similarities to Gabor filters in initial CNN convolutions (Springenberg et al., 2015; Yosinski et al., 2015). The contribution of these new depth features is shown in comparison to the classification of 2D colour images. Additionally, the extent to which depth information can improve current state-of-the-art skin lesion classification systems that only resort to the traditional 2D imagery is also assessed.

The main contribution of this section is the exploitation of 3D surface skin data as an alternative data modality for melanoma discrimination. Additionally, the Morlet Wavelet-based features are also introduced for this type of data and compared to the current state-of-the-art results. Since this data is originated from light-field imagery, a comparison to typical colour based classification is possible, as the used dataset provides both colour image and 3D information for every image-pixel data.

The remainder of section is organised as follows: Section 5.2.1 presents the literature involved in this work, including the concept of wavelet scattering. Section 5.2.2 describes the proposed approach pipeline, including relevant details about the experiment parameters, segmentation, data pre-processing, augmentation, normalisation, model feature extraction, and the DL model. Section 5.2.3 performs parameter selection and discusses the attained results. Finally, Section 5.2.4 highlights the conclusions.

### 5.2.1 Relevant Background

Image recognition and classification using ML has become a major topic in a wide range of research fields, specially with DL. For instance, in the field of skin lesion classification, CNNs have produced promising results (Gonzalez-Diaz, 2018; Tang et al., 2020) when operating on 2D/colour information. Yet, recent research based on data-driven models have reported the highest performance measurements ever published across multiple 2D test datasets (Hosny et al., 2019). The use of these pre-trained models is typically accompanied by a Transfer Learning (TL) method (Shin et al., 2016; Barata et al., 2018), which can be additionally aided by manually extracted features (e.g., as in Hagerty et al., 2019).

In order to properly address various concepts or areas necessary to support this work, the remainder of this section is structured into four paragraphs, namely: Deep Learning, segmentation, datasets, and Wavelet Scattering (WS).

**Deep learning** Deep CNN-based models (DCNN) automate many aspects of skin cancer classification (Gessert et al., 2019; Xie et al., 2020; Yuan et al., 2017; Esteva et al., 2017; Liu et al., 2020). However, diagnostic performance is still hindered by several already mentioned factors, making large sets of data necessary for adequately training, as exemplified by the use of millions of images in the ILSVRC (Deng et al., 2009). Recently, some works have shown that TL can enable significant classification results, comparable to those of professional dermatologists diagnostics (Esteva et al., 2017; Liu et al., 2020).

On large-scale image classification tasks (e.g., ILSVRC), improving the DCNN structure from an initial AlexNet (Krizhevsky et al., 2012) to the recent RegNet (Radosavovic et al., 2020), or increasing the model parameter capacity (Radosavovic et al., 2020; Tan & Le, 2019; He et al., 2016), enables better performances. However, in small-scale image datasets it is very difficult to increase the performance, as increasing the number of parameters may induce the model to transition from an under-fitting space to space where the over-fitting probability is higher (Belkin et al., 2018). In some works described in the literature, DCNNs are selected without taking into account the new dataset size and the new intra- or inter-class variations, avoiding such issues by using mechanisms that mitigate the over-fitting problem (Esteva et al., 2017; Liu et al., 2020; Yu et al., 2018; Han et al., 2018; Brinker et al., 2019; Hosny et al., 2019). Some of these mechanisms are TL (Hosny et al., 2019), data augmentation (Bisla et al., 2019; Hosny et al., 2019), and multi-target weighted loss functions (Fernando & Tsokos, 2021; Hosny et al., 2019). Alternatively, other data and model adjustments that have been learned from large-scale image classification, such as data normalisation into certain ranges or residual connections between distant layers, can be used (Radosavovic et al., 2020; He et al., 2019; Wu & He, 2018; Ioffe & Szegedy, 2015; Mishkin et al., 2017). A combination of such mechanisms is used in Hosny et al. (2019) using TL with a pre-trained AlexNet CNN model.

**Datasets** The research developed in this section relies on the SKINL2 dataset. In this context, it is relevant to note that this dataset is even smaller than other available 2D datasets like the one used in Yao et al. (2021).

**Wavelet scattering** Also relevant for this work is the concept of Scattering Transform and its early use in CNN architectures as alternative to initial convolution layers, since it provides unique features invariant to translation, rotation, scale, and frequency shifting, allowing the creation of lesser deep models.

In Mallat (2012), the concept of Lipschitz-continuous translation- and rotation-invariant operators for wavelets is presented, where differentiable manifolds are smoothly mapped with invertible functions – diffeomorphism. Lipschitz continuity is the central condition to guarantee the existence and uniqueness of a solution to an optimisation problem. This condition is discarded by CNNs when matching patterns during the training process, allowing similar patterns to exist (even if only initially) and match identical solutions (Bruna & Mallat, 2013). This wavelet-propagating operator is a path-ordered product of nonlinear and not-comparable operators, each one computing the modulus of a wavelet transform. The scattering transform window is generated by a Lipschitz-continuous local integration, which converges to a translation-invariant wavelet scattering transform as the window size increases. The scattering coefficients also provide representations of stationary processes (Mallat, 2012; Waldspurger, 2017).

Based on this core concept is the Wavelet Scattering (WS) framework, which is used as convolution layers for NNs. The convolutions obtained from WS – whose filters are fixed to be wavelet and low-pass averaging filters coupled with modulus non-linearities – compute translation invariant image representations, which are invariant to deformations while preserving high frequency information for classification. In Bruna & Mallat (2013), the mathematical analysis of wavelet scattering networks explain important properties of DCNN classification, presenting results for handwritten digits and texture discrimination.

Some degree of invariance to translation and diffeomorphism is necessary for many classification or regression tasks. In DL, using CNNs for example, the use of the WS framework can create an initial model that includes one or more layers responsible for transforming the non-linear input into representations invariant to geometric transformations (translations, rotation, scale and frequency shifting), while preserving a high degree of discriminability (Bruna & Mallat, 2013; Waldspurger, 2015; Sifre & Mallat, 2013). These transformations have two main advantages. First, they perform dimensionality reduction to the data, while allowing a structured feature representation to be captured for a given task. Second, the geometric-invariant representation that is mapped into a smaller dimension space allows for simpler model building, especially in the presence of small training sets (Adel et al., 2017; Bruna & Mallat, 2013; Chudáček et al., 2013).

Across several fields, replacement or augmentation of learnable convolutions is being performed with this WS framework. In summary, the scattering transform is defined as a complex-valued CNN whose filters are fixed to be wavelets and the non-linearity is a complex modulus. Because wavelet transform is contractive, as is the complex modulus, so is the whole network, resulting in a reduction of variance and added stability relative to additive noise. Also, since each layer is a wavelet transform that separates the scales of the incoming signal, invariability to deformation of the original signal is also attained. All these aforementioned properties enable the representation of structured signals such as natural images, textures, audio recordings, biomedical signals, and molecular density functions, among others.

### 5.2.2 Proposed Wavelet Scattering-based Classification Approach

As pointed out before, the aim of this work is to improve the accuracy of melanoma discrimination of conventional methods that only use colour (RGB) information, by including an additional dimension (depth) that characterises the skin surface rugosity. To achieve this goal, a pre-processing and classification pipeline is proposed to enable the use of RGB and corresponding depth (Z), which are referred to as *image components* along with a segmentation mask to be computed at the first stage of the pipeline. The influence of depth information in melanoma discrimination is also evaluated when both types of data are simultaneous used (i.e., RGBZ), in comparison with the use of RGB information only. The classification pipeline, in particular, comprises two main stages: a Morlet Scattering Transform, which mimics initial DL convolutions by computing initial features with high discrimination capacity and enabling the use of a shallower model; followed by the actual DL model, comprised of learnable convolutions and a softmax output.

Overall, the proposed pipeline has three types of configurations in this study: target classes; target dimensions; and model extensions. The *target classes* configuration, which will be further detailed in Section 5.2.3, sets the classification spectrum as either: binary discrimination of melanoma versus nevus samples (MvsN); or binary discrimination of melanoma versus all other skin lesion types (MvsAll). The *target dimensions* configuration sets the data dimensions (e.g. image size after resize) at a given classification study (Section 5.2.2-Data Pre-Processing). Finally, the *model extensions* defines a set of training configurations, which provide extended results to the *target dimensions* and help the interpretation of the model capabilities (Section 5.2.2-Morlet Scattering and Section 5.2.2-Classification). Both *target dimensions* and *model extensions* are defined in this section and later exploited in Section 5.2.3-Parameter Selection to define the final configuration of the proposed method for the selected dataset classification targets.

The processing pipeline comprises six stages, as depicted in Fig. 5.5, which are detailed in the following six paragraphs. Given a RGBZ dataset, where each pixel consists in colour (RGB) and depth (Z) information, a lesion segmentation mask is firstly generated, as described in Section 5.1.2-Segmentation. After extraction of the lesion segmentation mask, a given dataset sample is comprised of an RGB image, its depth map Z, and the segmentation mask, i.e. a total of five components at the pixel level (RGBZ plus segmentation mask). This dataset undergoes a process of data augmentation by means of random rotations in order to reduce the overfitting probability, as described in Section 5.2.2-Augmentation. Using the segmentation mask, the minimal lesion-bounding-box is determined and the pixels beyond such box are removed from the data, effectively making the new data a rectangular crop of the segmented lesion area. Concurrently, pixel values belonging to healthy skin in this crop area are set to zero. At this point, as described in Section 5.2.2-Data Pre-Processing, the parameters defined by *target dimensions* configuration define which image components to maintain and to what shape resize the data sample. Then, data is normalised into a defined range (Section 5.2.2-Normalisation) to

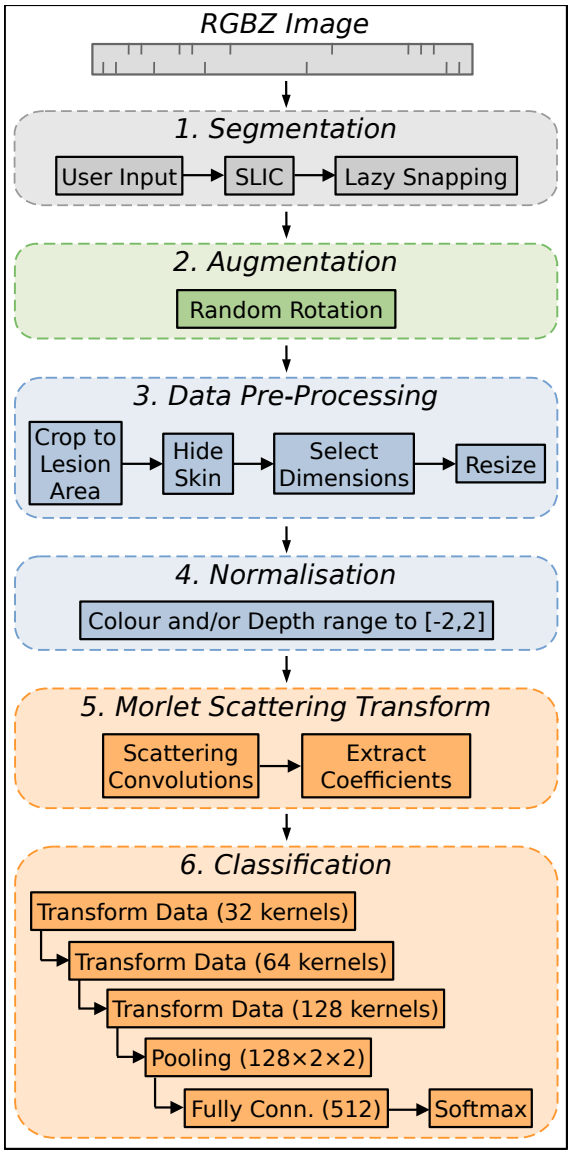



**Figure 5.5:** Proposed Pipeline: a given light-field (RGBZ) dataset *1)* ask for user input and perform segmentation using Lazy Snapping empowered by SLIC; *2)* apply augmentation by a random rotation to both RGBZ data and segmentation; *3)* pre-process data by cropping around segmentation lesion area, hide skin information, select which image components to maintain, and resize the cropped image to the target experiment size; *4)* apply normalisation by transforming the cropped image values into a range between $[-2, 2]$; *5)* create the scattering convolutions and extract a set of scattering coefficients; and finally *6)* apply a classification model that sequentially transforms said set into a larger one, which is then reduced through pooling for a final fully connected layer to provide the softmax discrimination label.

feed the Morlet Scattering Transform (Section 5.2.2-Morlet Scattering), which extracts features to fuel the DL model (Section 5.2.2-Classification). This model increasingly expands the data sample analysis, before feeding the final fully connected layer that provides the softmax discriminative label. Detailed information about each stage is provided in the following six subsections.

**Augmentation** Classification algorithms, as is the case of DCNN, usually require large amounts of data to yield proper performance (class separation) and convergence (feature discovery). The dataset used in this work has a small number of images, therefore it is necessary to expand it by augmenting the existing images. To this aim, all input training image samples are randomly rotated from 0 to 360 degrees prior to be used in the training phase. Additionally, each epoch comprises 72 passes through the training dataset, which allows each image to be analysed at 72 potentially-different angles before a new epoch begins with another set of 72 random rotations. This is a similar approach to that implemented in Hosny et al. (2019), but in this case the rotation-degree is not restricted and augmentation is not used during the test phase.

At this stage, each input training image (i.e., dataset sample) comprises five components ($C$): an RGB image, the corresponding depth data, and the lesion segmentation mask. All image channels are geometrically transformed by the same rotation, keeping the information aligned, such that the segmentation mask still provides the correct lesion location in both the RGB and depth information.

**Data Pre-Processing** Given an RGBZ dataset sample and its lesion segmentation mask, the pre-processing stage sets the *target dimensions* configuration parameters for the experimental setup. There are two parameterisations: *i)* selection of the image components; and *ii)* model input image size. Besides these options, the image data entering the pre-processing stage is cropped to the bounding limits defined by the lesion segmentation mask. Concurrently, the healthy skin region in this cropped area is removed by setting the corresponding pixel values to 0 (zero). The removal of the surrounding healthy skin region is intended to focus the model on the lesion, not allowing speculations about possible patterns or features of regions outside the lesion area.

In regard to the *image components*, the pipeline can operate in different modes by exploiting either only colour (RGB) data, only depth (Z) data, or both colour and depth (RGBZ) data. Only the selected components are used by the proposed algorithm. The selection of such different operational modes, has obvious impact on the learning process and consequently on the model, allowing to compare the performance between models obtained by learning with different image components.

In regard to the *image size* parameter, three possible re-scaling factors are considered, where the image is resized: to $32 \times 32$, to $64 \times 64$, or to $128 \times 128$ pixels. This image resize is necessary because the crop of the lesion region generates different area sizes for different images, creating conflicts of input data sizes for the model along the proposed pipeline. Additionally, considering that the original image size may be too large, depending on the number of images available in the dataset, the model resources may be inadequate, for instance, accelerating the model overfit. Therefore, the last step of the pre-processing stage is to resize the existing images to a fixed (smaller) size using bilinear interpolation.

**Normalisation** Given the *image components* entering in this stage, the respective data is normalised to improve the model convergence. This is a usual procedure due to the fact that CNNs, or NN in general, perform better if the input data is constrained to certain ranges.

For the colour components, the normalisation transforms the data to the approximate range $[-2, 2]$ as in other DCNN applications (namely Hosny et al., 2019). This is, as traditionally applied in ImageNet, normalisation is carried out by subtracting the values of $(0.485, 0.456, 0.406)$ and dividing by the values of $(0.229, 0.224, 0.225)$ for the R, G, and B components, respectively, so that the value range is comprised between $[-2, 2]$. For the depth component,

the same operation is performed by subtracting 6.26 and dividing 3.03, in order to constrain it to the range of $[-2, 2]$. This normalisation stage operates on either colour, depth or both components according to the selection made in the previous data pre-processing stage.

**Morlet Scattering** At this stage, a dataset sample is represented by either 3, 1, or 4 channels ($C$) (only RGB, only Z, or RGBZ, respectively). Prior to be processed by the classification model (Section 5.2.2-Classification), unique features invariant to rotation, translation, and scale are extracted using a WS framework with a Morlet wavelet as the mother wavelet (Sifre & Mallat, 2013). In addition to the extraction of unique features, this process also reduces the data volume and, consequently, further prevents model overfitting. This extraction of features can be performed either by calculating only first-order coefficients or by extending to second-order calculations, which are considered as part of the *model extensions* parameters.

The mother wavelet ($\psi(t)$) used in this work is the Morlet wavelet and, to speed-up the process, the convolutions are performed in the Fourier domain. The corresponding family of wavelets is generated by dilation and translation from the mother wavelet as in Eq. 5.4, where $a$ is a scale factor and $b$ is the time index, while the factor $|a|^{1/2}$ is used to ensure energy preservation. In this work, the input data is represented as 2D matrices of $N \times N$ values, where $N$ can only assume the values 32, 64, or 128. Let $x[\mathbf{n}]$ be any signal on this $N \times N$ grid, as $x[n, m]$. The periodic convolution with another signal $y[\mathbf{n}]$ is denoted by $x \circledast y[\mathbf{n}]$. The scattering transform uses a wavelet filter bank for each order greater than zero, that is $\psi^{(1)}_{\lambda_1}[\mathbf{n}]$ for the first-order and $\psi^{(2)}_{\lambda_2}[\mathbf{n}]$ for second-order respectively, where $\lambda_1$ and $\lambda_2$ are frequency indices in the sets $\Lambda_1$ and $\Lambda_2$. The low-pass filters are represented by $\phi_J[\mathbf{n}]$, specifying an averaging log-scaling filter of $2^J$ (which nearly linearises the variations of scattering coefficients), where $J$ is a regulator variable. Input data partitioning is also computed in relation to $J$ as non-overlapping patches of size $2^J$, thus producing $N/2^J$ partitions. This logarithmic non-linearity is first applied to invariant scattering coefficients to linearise their power law behaviour across scales. This is similar to the normalisation strategies used with bag of words (Lazebnik et al., 2005) and deep NNs (LeCun et al., 2010). Together with a non-linear function $p(t)$, the filters comprise the scattering transform. The non-linear function employed in this work is the complex modulus $p(t) = |t|$, as in Andén & Mallat (2014); Bruna & Mallat (2013).

$$\psi_{a,b}(t) = |a|^{1/2} \psi\left(\frac{t-b}{a}\right) \tag{5.4}$$

The zeroth-order scattering coefficient $S_0(x[\mathbf{n}])$ is the local average as given by Eq. 5.5. The first-order scattering coefficients are obtained from convolution of $x[\mathbf{n}]$ with the first-order wavelet filter bank, as defined in Eq. 5.6. These are the least computationally expensive coefficients to be used in the classification process. Second-order coefficients are obtained as an extension of the first-order ones, as defined in Eq. 5.7, where further data structures are captured by decomposing the $p(\cdot)$ results using the second filter bank $\psi^{(2)}_{\lambda_2}$. Note that this is only performed for a subset $\Lambda_{2,*} \subset \Lambda_2$ defined only for the elements of $\Lambda_2$ corresponding

to elements of $\Lambda_1$, since results from the first-order $p$ represent low-frequencies. The Morlet filters are similar to normalised zero-mean Gabor functions and are, therefore, computed as such for simplicity. To reduce computational load, data obtained from $p(t)$ is down-sampled as in Sifre & Mallat (2014).

$$S_0(x[\mathbf{n}]) = (x \circledast \phi_J)[\mathbf{n}] \tag{5.5}$$

$$S_1(x[\mathbf{n}, \lambda_1]) = \left(p\left((x \circledast \psi^1_{\lambda_1})[\mathbf{n}]\right) \circledast \phi_J\right)[\mathbf{n}], \quad \lambda_1 \epsilon \Lambda_1 \tag{5.6}$$

$$S_2(x[\mathbf{n}, \lambda_1, \lambda_2]) = \left(p\left(\left(p\left((x \circledast \psi^1_{\lambda_1})[\mathbf{n}]\right) \circledast \psi^2_{\lambda_2}\right)[\mathbf{n}]\right) \circledast \phi_J\right)[\mathbf{n}], \quad \lambda_1 \epsilon \Lambda_1, \lambda_2 \epsilon \Lambda_2(\lambda_1) \tag{5.7}$$

In this work, the $J$ regulariser is always set to 2 and a rotation parameter $r$, which defines how many filter rotations are performed to induce rotation-invariance, is set to 8. Assuming the already mentioned $N \times N$ pixel-grid, the Scattering Transform of the WS framework with a scale $J$ and $r$ angles will generate a 3D set of features $V_S$, as expressed in Eq. 5.8, for methods configured to use only first-order coefficients, or as expressed in Eq. 5.9, for methods including second-order coefficients. An input dataset image generates a one-fourth-sized grid $\hat{N}$ of either $8 \times 8$, $16 \times 16$, or $32 \times 32$, with either $K = 17$ or $K = 81$ feature values in each cell, depending if they are configured to use only first-order or both first-order and second-order coefficients. For example, if the experiment is configured to run RGB components (i.e. three pixel-grids, $C = 3$) with first-order coefficients, then three sets are generated, each with $K = 17$ features per cell – a total of three $17 \times \hat{N}$ feature sets per dataset image.

$$V_{S_1 K} = 1 + rJ \ , \ V_{S_1 x} = \frac{N}{2^J} \ , \ V_{S_1 y} = \frac{N}{2^J} \tag{5.8}$$

$$V_{S_2 K} = 1 + rJ + \frac{r^2 J(J-1)}{2} \ , \ V_{S_2 x} = \frac{N}{2^J} \ , \ V_{S_2 y} = \frac{N}{2^J} \tag{5.9}$$

Prior to the next stage, feature sets are stacked along the $V_{SK}$ dimension to generate a single feature set $\hat{V}$ of size $KC \times \hat{N}$. This means, for example, if three blocks are produced (as occurs when processing RGB data), then the new set $\hat{V}$ will maintain the second and third dimensions, while the first dimension grows to three times the size – assembling a $\hat{V}$ of $3K \times \hat{N}$ features. Stacking is performed on the first dimension ($K$), in opposition to other dimensions of size $N$, so that features regarding the same image location but of different components remain grouped together. That is, when working with the image components, vectors of $K$ features that are extracted from each individual component (in a particular region) are stacked together in order to simplify the visualisation of the feature-information by the subsequent CNN classification model convolutions.

**Classification** As depicted in Fig. 5.6, given a set of features $\hat{V}$, the classification is performed by a CNN model that, apart from the first convolutions, is a fixed-size network for the whole experiment. The model comprises three main parts: *i)* a fitting part, where input features are convolved with a kernel designed to fit the data to the fixed network dimensions; *ii)* an expanding part, where two repeating blocks process and expand the data; and *iii)* a classification part, where a fully connected layer provides the classification output.

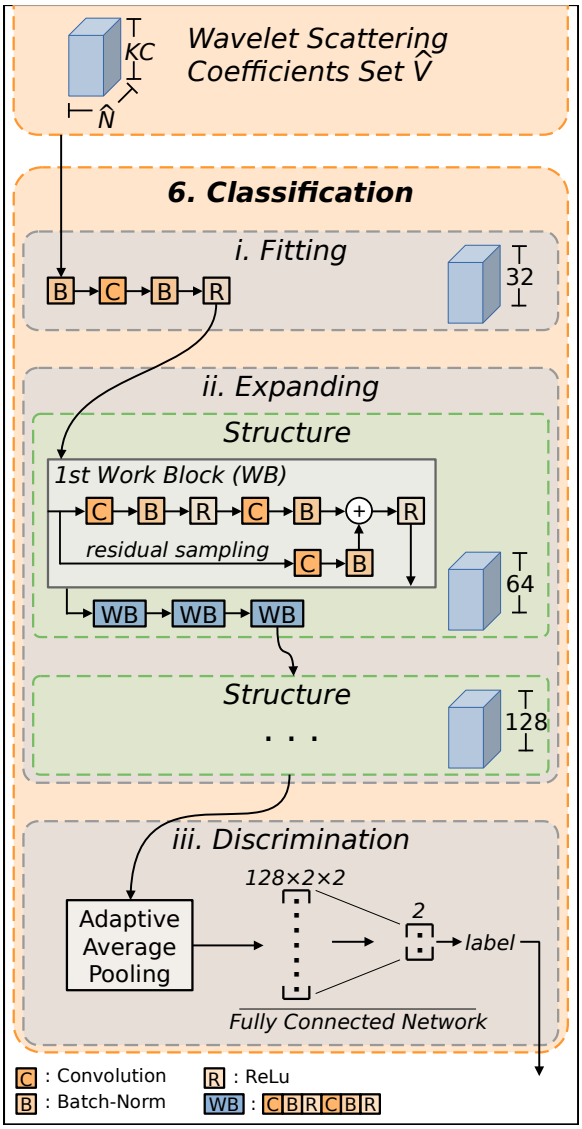



**Figure 5.6:** Classification Model Pipeline. Receiving a feature-set $\hat{V}$ of scattering coefficients, train a deep learning model comprised of three main parts: (i) a initial data fitting, (ii) a main processing part with convolutions that expand a given data volume, and finally (iii) a fully connected layer.

For all convolutions, the value of the stride is 1 pixel, the kernel size is $3 \times 3$ unless stated otherwise, and the value of the bias is set to zero. In all batch-normalisation layers (Ioffe & Szegedy, 2015), the running estimates parameter is set to 0.1 and possess learnable affine-transformation parameters, unless stated otherwise. For the remainder of this section, every convolution layer is followed by a batch-normalisation and a Relu activation function, unless stated otherwise.

The fitting part of the network ($i$) comprises batch-normalisation and a convolution layer of 32 kernels. In this first convolution layer, the feature-set $\hat{V}$, which has an experiment-variable size $KC \times \hat{N}$, is transformed to a fixed size of $32 \times \hat{N}$. The first part of the network has $K \times C \times 288 + 64$ trainable parameters. Additionally, the initial batch-normalisation has no learnable affine-transformation parameters and only exists to further regularise the input data range for the model.

The expanding part of the network ($ii$) is a structure that repeats twice, each comprising four working blocks (WB) with a residual connection. The only difference from one structure to

the next is the target number of kernels in every convolutional layer, which are 64 and 128 for the first and second structure, respectively. Each of the four mentioned working blocks comprises two convolutions. The first working block of each structure has an additional third convolution, which receives the same data as the first convolution (performing the same operations). However, this block's kernels are of size $1 \times 1$ and there is no Relu at the end. The output of this third convolution is added to the second convolution batch-normalisation output, before Relu, as residual information. These two-parts of the network structure have 279,680 and 1,116,416 trainable parameters, respectively.

The classification part of the network (*iii*) performs a binary softmax classification with the result of a biased fully connected layer of 512 inputs to two neurons. This layer is adopted, with the traditional sigmoid activations, as it is an universal approximator (Csáji, 2001) for classification problems. Since the set $\hat{V}$ entering the network has size $KC \times \hat{N}$, at this point, after all convolutions, it will have $128 \times \hat{N}$. This means that it will have a variable size in the second and third dimensions, represented by $\hat{N}$. In order to encapsulate this information into a fixed size, so that models compiled for different input sizes remain comparable, an average pooling layer is added before the fully connected layer to adaptably reduce the data volume into a fixed sized $128 \times 2 \times 2$ volume (i.e. the referred 512 input values of the fully connected layer). This last part of the network has 1,026 trainable parameters.

The fully described network is trained using Stochastic Gradient Descent with Nesterov momentum (Sutskever et al., 2013). The learning rate is fixed at 0.001 and the momentum at 0.9. Additionally, weight decay (L2 penalisation) is also performed at 0.0005, in order to exponentially decay weights to zero, limiting the number of free parameters in the model and avoiding rapid over-fitting.

In this work, instead of having the learning rate influencing the new momentum velocity by scaling the gradients, the velocity does not depend on the learning rate. Rather, the learning rate is used when updating the model parameters, scaling the whole velocity equation result (meaning that it also scales the previous momentum-ed velocity). This choice was made to smooth the model learning, as to further challenge early overfitting.

Finally, to promote balanced classification-error corrections in the network during training, the model softmax-cross-entropy loss function is weighted (via cost matrix) for a given class, as the number of training samples in the largest class divided by the given class number of training samples. Effectively, this makes one error in the smaller class more significant than one error in the larger class, implicitly balancing the dataset.

### 5.2.3 Results and Discussion

The experimental results presented in this section are expressed in terms of percentage of classification ACC, SEN, and SPE, aligned with most of the cited works, where SEN represents the successful melanoma identification rate and SPE the successful identification of the other class. Since this is an unbalanced problem, BAC is also used.

These results encompass two main classification experiments (*target classes*), both executed applying 10-fold Cross Validation (CV). The first experiment, refers to melanoma classification against nevus samples (MvsN), while the second experiment performs the classification of melanoma versus all other skin lesion types (MvsAll).

The learning process was run for seven epochs in all executions, aiming for approximately 500 dataset passes through the model, as each epoch comprises 72 random rotation of each sample. In the *model extensions* configuration the following four different batch sizes were used for the model: 20, 40, 60 and 80.

The remainder content of this section is organized as follows: Firstly, the used data and the used *target classes* partitioning is described. Secondly, a description of the parametrisation selection of the final model is performed. Finally, the achieved results with the selected parameters are discussed and compared to the current state-of-the-art.

**Dataset** The pipeline was evaluated using the publicly available SKINL2 dataset, as previously described in Section 5.1.5-Dataset.

#### Parameter Selection

Before any final results can be extracted and compared with the state-of-the-art, the model parameters must be adjusted to the dataset and image data. Three parameter configurations need to be discussed: coefficients order (i.e. either first or second order coefficients), target size of resized images and model batch size. To understand the influence of the coefficient order and the image size on the data components, the results for the more balanced MvsN dataset is first analysed. Table 5.3 depicts the average BAC results for the image size parameter in each data dimension when using either the first- or second-order coefficients. BAC values are the averaged results obtained by the different batch sizes.

As can be seen in Table 5.3, the best average BAC performance in each coefficient order (marked in boldface) is achieved by the intermediate image size of $64 \times 64$ pixels, with 81.25% and 68.30% BAC performance in the first-order, for RGB and depth respectively, and 82.89% and 67.01% in the second order results. The higher performance in the intermediate image size is expected because using the smaller $32 \times 32$ image size removes too much information

**Table 5.3:** Average BAC for each image resize and for 1st and 2nd order coefficients, over the possible data components - MvsN experiment.

| Order | Image Size | Data Components | |
|---|---|---|---|
| | | RGB | Z |
| 1st | $32 \times 32$ | 78.08 | 64.29 |
| | $64 \times 64$ | **81.25** | **68.30** |
| | $128 \times 128$ | 79.32 | 65.67 |
| | *average* | *79.55* | *66.09* |
| 2nd | $32 \times 32$ | 77.48 | 58.28 |
| | $64 \times 64$ | **82.89** | **67.01** |
| | $128 \times 128$ | 76.36 | 59.87 |
| | *average* | *78.24* | *61.72* |

due to the down-sampling. However, using a larger $128 \times 128$ image size slightly decreases the classification performance as the model quickly overfits on more detailed features provided by the WS framework during the training on this small dataset.

Table 5.3 also allows to analyse the average performance, for all images sizes and batch sizes, (marked in italics) for the two different coefficient orders. The best average result is obtained for the first-order coefficients with 79.55% and 66.09%, for RGB and depth respectively, against 78.24% and 61.72% BAC when using second-order coefficients.

The results shown in Table 5.4 for the MvsAll experiment were obtained under the same test conditions. In this case, the best average BAC is not achieved for the same image size. Yet, the different results obtained for each image size allows to observe that $64 \times 64$ offers the best compromise in both coefficient orders. For example, in the first-order coefficient results, selecting $128 \times 128$ instead of $64 \times 64$, causes an improvement of 1.3*pp* in the Z average BAC, while for the RGB the performance drops 3.28*pp*. Therefore, $64 \times 64$ is preferred, favouring the RGB classification. This analysis also works for the second-order coefficients. If the $32 \times 32$ image size is selected instead of the $64 \times 64$, the average BAC for Z improved by 0.15*pp*, while for RGB it drops 7.88*pp*. Therefore, $64 \times 64$ is preferred, also favouring of the RGB classification.

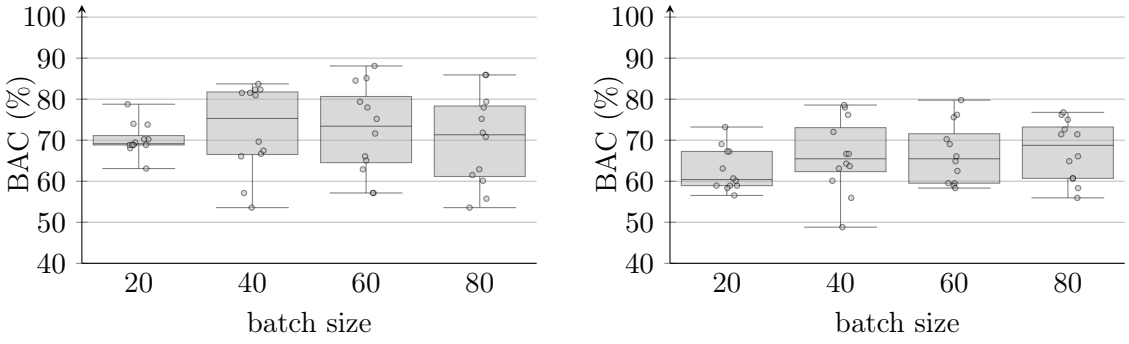
Similarly to the image size, the best coefficient order for the MvsAll experiment is not an obvious choice. Resorting to the same rationale as in MvsN, in Table 5.4 the best average BAC across image and batch size (marked in italics) is obtained by the second-order coefficients with 71.73% and 61.86%, for RGB and Z respectively, against 69.59% and 59.92% for the first-order coefficients. This can be partially explained due to the added variability in the dataset comprising the MvsAll experiment. In this case, there are seven different skin lesion types, instead of only two, creating a broader view of the classification problem and, consequently, requiring more detailed features, as present in second-order coefficients. The difficulty in selecting the best parameters in the case of the MvsAll experiment is probably due to the fact that in this experiment classes are even more imbalanced than in MvsN.

**Table 5.4:** Average BAC for each image resize and for 1st and 2nd order coefficients over the possible data components - MvsAll experiment.

| Order | Image Size | Data Components RGB | Z |
|---|---|---|---|
| 1st | $32 \times 32$ | 68.60 | 56.99 |
| | $64 \times 64$ | **71.73** | 60.71 |
| | $128 \times 128$ | 68.45 | **62.05** |
| | *average* | *69.59* | *59.92* |
| 2nd | $32 \times 32$ | 66.67 | **62.50** |
| | $64 \times 64$ | **74.55** | 62.35 |
| | $128 \times 128$ | 73.96 | 60.71 |
| | *average* | *71.73* | *61.86* |

From these comparisons, it is safe to conclude that a good compromise in terms of the average BAC metric performance is achieved when configuring the image size as $64 \times 64$, using first-order coefficients for the MvsN experiment and second-order coefficients for the MvsAll experiment.

The batch size for each experiment can also be determined following the same approach. Since these experiments contain different amounts of data samples, 50 for MvsN and 98 for MvsAll, it is expected that the preferred batch size will also differ in a similar ratio. Resorting to a box and whisker plot, Fig. 5.7a depicts the average BAC metric-value for the different batch sizes in MvsN independently of the image size, the coefficient order, and the use of either RGB or depth data. As before, looking for the average best performing metric value sets the best batch size as 40 with 75.30% BAC average performance.




**(a)** *MvsN* **(b)** *MvsAll*

**Figure 5.7:** Box-plot of BAC (with data points) for the different batch sizes in (a) MvsN and (b) MvsAll across the remaining parametrisation options.

A similar analysis is performed for the MvsAll experiment, as depicted in Fig. 5.7b. In this figure, box-plot data-dispersion appears smaller than in Fig. 5.7a, most likely due to the increase in the dataset size. Starting from the left, MvsAll results appear initially similar to MvsN: a compact spread at batch size 20; an average improvement peaking at 40 with some data points polling down the average performance; then starting to lose performance

**Table 5.5:** Proposed Morlet-based Method Results.

| Dataset | Method | ACC | SEN | SPE | BAC |
|---|---|---|---|---|---|
| MvsN | Baseline (RGB) | 68.00 | 21.43 | 86.11 | 53.77 |
| | Proposed (RGB) | 84.00 | 78.57 | 86.11 | 82.34 |
| | Proposed (Z) | 74.00 | 50.00 | 83.33 | 66.67 |
| | Proposed (RGBZ) | **94.00** | **92.86** | **94.44** | **93.65** |
| MvsAll | Baseline (RGB) | 73.47 | 14.29 | 83.33 | 48.81 |
| | Proposed (RGB) | 86.73 | 50.00 | 92.86 | 71.43 |
| | Proposed (Z) | 85.71 | 14.29 | **97.62** | 55.95 |
| | Proposed (RGBZ) | **89.80** | **78.57** | 91.67 | **85.12** |

Baseline: as in Hosny et al. (2019)

at batch size 60. In opposition, the average BAC performances rise again to a new peak at batch size 80, providing a even better average performance as well as a more compact behaviour than with 40. This is expected since the amount of data samples is almost twice in the MvsAll experiment than in the MvsN. Thus, the selected batch size for the MvsN and MvsAll experiments are 40 and 80, respectively.

### Results

Using the parameters defined just above, that is: image size of $64 \times 64$; first-order coefficients and batch size 40 for the MvsN experiment; and second-order coefficients and batch size 80 for MvsAll – the proposed model achieves the results depicted in Table 5.5. These results were obtained using RGB and Z (depth) components individually, *Proposed (RGB)* and *Proposed (Z)* respectively, and with all components, *Proposed (RGBZ)*. The results are also compared to the state-of-the-art method in Hosny et al. (2019), named *Baseline (RGB)*, providing classification results for both experiments (MvsN and MvsAll). This classification method was selected as baseline since it performs comparisons with three well-known 2D datasets and outperforms other 11 state-of-the-art algorithms. Averaging across the three datasets mentioned in its' work, this method reports a 96.8% ACC performance when using data augmentation and 88.9% when not using it. At the time of writing this thesis, to the author's knowledge, there are no other works published by other authors resorting to the SKINL2 dataset, which could be used for comparison.

Using the dataset employed in this work (SKINL2), the baseline method provides a 68.00% and 73.47% ACC performance with 53.77% and 48.81% BAC for the MvsN and MvsAll experiments, respectively. While the ACC increases in the MvsAll experiment (which has 48 additional samples in comparison with MvsN), it is important to point-out that the SEN metric decreases by 7.14*pp* even though the number of melanoma samples is the same (14) in both experiments. This decrease represents the misclassification of one additional melanoma, identifying only 3 out of 14 in the MvsN experiment and 2 out of 14 in MvsAll. The SPE

metric is not comparable since the amount of samples differs between these experiments. Across the 10-fold CV, the baseline method correctly identifies 31 out of 36 nevus in the first experiment, and 70 out of 84 non-melanoma lesions in the second experiment.

As can be seen in Table 5.5 for the MvsN experiment, the proposed approach (*Proposed (RGBZ)*) achieves 94.00% ACC and 93.65% BAC, an increase of 26.00*pp* and 39.88*pp*, respectively, when compared to the *Baseline (RGB)* method. This improvement comprises the utilisation of both RGB and depth components. If only the RGB data dimension is used, the proposed pipeline achieves only 84.00% ACC and 82.34% BAC, 10.00*pp* and 11.31*pp* lower than the results achieved when using both components, respectively. Also, the use of only the depth component does not perform as well as using RGB component, however its performance is still superior than the baseline method for all metrics, except for SPE.

As expected, the combined use of both RGB and depth components surpasses the individual usage of only one of them, allowing to infer that the depth component owns discriminative power not present in RGB. For instance, exploring the label predictions performed by the separate RGB and Z models, it is clear that two melanoma samples, which are correctly classified using depth, are not correctly classified when using RGB only. Getting the two components together in the new model (RGBZ) also allows the prediction of the other two melanoma samples, which were wrongly classified using only RGB components, thus supporting the assumption that the skin lesion surface has potential to improve the discrimination between melanoma and nevus.

For the experiment MvsAll, the results achieved by the proposed pipeline are also shown in Table 5.5, where *Proposed (RGBZ)* attains 89.80% ACC and 85.12% BAC, an increase of 16.33*pp* and 36.31*pp* respectively, when compared to the *Baseline (RGB)* method. Like in the MvsN experiment, this increase corresponds to the use of both RGB and depth components. When using the RGB component alone, the proposed approach achieves only 86.73% and 71.43%, that is 3.07*pp* and 13.69*pp* lower than the *Proposed (RGBZ)* results, although still superior than using the *Baseline (RGB)* method.

If the method uses only the depth component, similarly to the case of MvsN, the results are also lower than the *Proposed (RGBZ)* results, yet still superior to the *Baseline (RGB)* results for all metrics. In this MvsAll experiment, however, the data imbalance is greater than in MvsN. Incorrect melanoma classifications almost go unnoticed by the ACC metric since, for instance, a classification of all data as non-melanoma image samples immediately achieves 85.71% ACC. Nevertheless, this would be noticeable because the BAC metric would only achieve 50.00%. This means that, although the proposed RGBZ method achieves a similar ACC performance, the total number of melanoma-misclassification is lower, because the BAC performance is 85.12%, accounting for 78.57% SEN in this case. This corresponds to the correct classification of 11 out of 14 melanoma samples, nine more than the *Baseline (RGB)*.

In this section, all comparisons with the baseline classification method have shown that the proposed approach provides superior performance results. Accordingly, this can be seen as an indirect benchmark comparison of the proposed method with the works compared in Hosny et al. (2019) and other works that resorted to the same dataset and metrics. In essence, since the baseline method reports results superior to 10 other works, it is expected that the proposed approach could also show results superior to the mentioned works, if they were to be applied to the SKINL2 dataset. This hypothesis may be further extended to other works like Pereira et al. (2020b); Tang et al. (2020); Barata et al. (2018); Pathan et al. (2018); Hagerty et al. (2019)), that use the same datasets and metrics as the baseline method.

In addition to the discussed results, it is worthwhile to mention some studies that compare the results of computational models with human classification of skin lesions performed by specialists, i.e. dermatologists. This is the case, for instance of Esteva et al. (2017); Marchetti et al. (2018); Haenssle et al. (2018); Brinker et al. (2019), where the SEN and SPE are evaluated and compared. In Brinker et al. (2019), these comparisons were carried out in regard to the task of performing MvsN classification, involving 157 dermatologists that span across 12 German university hospitals. The test dataset used in this experiment comprises 20 melanomas and 80 nevi randomly selected from the ISIC dataset. Indirectly, this enables the comparison of the proposed approach with the results obtained from the 157 dermatologists. A mean of 74.1% for SEN and 60% for SPE was achieved by the dermatologists with dermoscopic images. This is inferior to the performance reported in Table 5.5 for the proposed RGBZ approach, which provides an additional 18.76*pp* in SEN and 34.44*pp* in SPE. Furthermore, in Marchetti et al. (2018) and Haenssle et al. (2018), respectively, 8 and 58 dermatologists have also participated in a similar study on another set of 100 images, obtaining 82% and 86.6% for SEN, and 59% and 71.3% for SPE. Again, on average, the proposed approach outperforms these classification results obtained by specialists.

Although the results obtained in Table 5.5 cannot be directly compared with the studies cited above, they establish a valuable reference for the expected classification performance made by specialists in the same MvsN dataset. Therefore, it is possible to infer that, on average, the proposed Morlet Scattering approach would outperform the human-based classification.

### 5.2.4 Conclusions

Currently, even with DL methods, discrimination of melanoma remains a challenging problem and current systems are yet to achieve satisfactory sensitivity performances. Rather than continuously attempting to improve algorithms by using the same data as commonly used by dermatology experts, other dimensions and modalities (as the skin lesion 3D surface) should be explored. In order to go beyond current state-of-the-art results, more reliable solutions might include merging 2D data together with other dimensional aspects, such as surface, which has potential to provide extended melanoma discrimination capabilities.

Taking advantage of the recently introduced technology of the light-field cameras, the main contribution of this section, apart from the proposed pipeline, is the evaluation of the skins' 3D surface data as an alternative data modality when performing melanoma classification, as well as its comparison to current state-of-the-art results. This is done resorting to a recent dataset of multi-dimensional imaging, which was specifically acquired for this goal. Since the data originates from light-field imaging, every image-pixel data comprises both dimensions, enabling the creation of the proposed pipeline, which operates in a comparable setting.

Despite the large class imbalance (often present in medical image datasets) and limited data samples, the attained classification results appear to surpass the sensitivity and specificity to discriminate melanomas from nevi, not only of the state-of-the-art algorithms, but also of human specialists. In the proposed approach pipeline ($RGBZ$), the melanoma discrimination against nevus was achieved with 94.00% ACC (comprising 92.86% SEN and 94.44% SPE) when combining 2D data with depth, a 26.00$pp$ ACC increase in relation to the state-of-the-art baseline method. In a similar setting, for the discrimination of melanomas against all other available skin lesions, the proposed approach achieved 89.80% ACC (comprising 78.57% SEN and 91.67% SPE), an increase of 16.33$pp$ relative to the state-of-the-art baseline method.

The experimental assessment allows to conclude that image classification problems, including melanoma skin lesion classification, can be further improved by including 3D information, such as surface depth data.

## 5.3 Summary

This chapter focused on using the SKINL2 dataset to propose two approaches that go beyond the current state-of-the-art 2D results by jointly exploiting the characteristics of both texture/colour and surface of skin lesions, while defining a user-accessible segmentation method.

The first approach comprises a model of two competing classification methods, which were combined using an uncertainty-aware decision function. The methods either classify 2D/colour information or, if uncertain of the classification label, classify 3D/surface data. Despite the large class imbalance, the ensemble model achieves high cross-validated melanoma classification accuracy. The results showed that, in the absence of discriminative 2D characteristics, the 3D surface provides redeeming results, demonstrating that improvement of the existing methods is possible when looking beyond 2D image characteristics.

The second approach proposes a processing pipeline which uses a Morlet Scattering Transform for feature extraction and a CNN model for classification, while also allowing to perform a comparison between using only 2D information, only 3D information, or both. Results showed that discrimination of melanoma achieved higher results when both 2D and 3D are used together. Overall the results of this section demonstrate significant improvements over

conventional approaches.

Expanding on the presented concepts, further research can be done in the field of skin lesion image classification to either improve existing methods that lack in performance or refine existing top performers, as shown in this research. Thus, future works should aim to enlarge existing datasets and acquisition modalities to enable the emergence of features specifically tailored for skin lesion detection and classification.

# Chapter 6

# Conclusion

Content



Skin lesion detection and decision making are very sensitive matters that require the intervention of experts such as dermatologists and pathologist. However, with recent advances on computer vision applications and systems, new tools have been developed to aid the decision-making process and even perform screening where the attention of a medical professional cannot be easily reached.

Early detection of suspicious PSL has a determinant role in clinical prognosis. One of the main issues with melanoma (skin cancer) detection is that it can also be visually similar to other skin abnormalities. Therefore, several techniques have been studied to develop reliable systems that aim to assist in the clinical diagnosis decision. This thesis addresses this issue by exploring existing methods from the literature and by creating new methods and approaches to advance this research field.

This chapter recaps the covered issues and the highlights addressed in this document.

## 6.1 Synthesis

The first set of contributions was presented in Chapter 3 describing two segmentation algorithms developed for different objectives. Works on skin lesion segmentation generally aim at round-like dermatology-expert delineation, since these are the available ground-truths that accompany public datasets. Hence, the first segmentation algorithm was created to tackle this problem by assuming that skin lesions can be represented by bi-model histograms of two dominant peaks and by finding the optimal gradient in which the segmentation would provide such peak separation. Then, in order to extract more detailed information of the skin border, a second segmentation algorithm was developed with the aim of finding more realistic borders. These two algorithms were compared against several methods described in the literature, and showed that, overall, the proposed segmentation methods were capable of out-

performing the state-of-the-art ones. Additionally, in preparation for the following research tasks, an experimental approach was created to validate the importance of the previously created segmentation masks by extracting features from them for the purpose of melanoma classification.

A second set of contributions was introduced in Chapter 4, which extended the previous knowledge on 2D images to the acquired 3D (light-field) dataset. The proposed approaches focus on providing evidence that images comprising the created dataset have relevant depth information for melanoma discrimination and classification by exploiting the 3D characteristics of the skin lesion surface. This was performed by combining existing algorithms with new techniques that exploit the new 3D information, clearly improving the classification performance. As an extension of the previous contributions, depth information along detailed lesion borders was also evaluated, showing that such information is still relevant for the classification process.

Finally in Chapter 5, the previous contributions were gathered to create two approaches for melanoma discrimination in light-field images of skin lesions. These methods exploit the characteristics of both texture/colour and the surface of skin lesions. The achieved results showed that, in the absence of discriminative 2D characteristics, the 3D surface provides redeeming results. In fact, discrimination of melanoma achieved higher results when both 2D and 3D are used together, demonstrating significant improvements over conventional 2D approaches.

## 6.2 Summary of Scientific Contributions

The core of this thesis is based on the developments associated to the identification and classification of skin lesions, with the particular scope of discriminating melanomas using light-field images. The major objectives included the research and development of new feature extraction algorithms and classification approaches, which would improve the state-of-the-art associated with melanoma skin lesions. To this end, a dataset of light-field images featuring melanocytic skin lesions was created and made publicly available aiming to increase the research in this area, at national and international level. Additionally, research and evaluation of different melanoma detection approaches using computer-based algorithms capable of differentiating melanoma from non-melanoma were proposed using the created dataset.

Working towards the goals established in Chapter 1, the work presented in this thesis resulted in the following contributions:

**Contribution 1: Segmentation approaches**

This contribution is presented in Sections 3.1 and 3.2 with the proposal of the GHT and LBPC segmentation methods (published in Pereira et al., 2019b,a). These proposals

target the generation of segmentation masks as dermatologists do (using visual colour gradients), as well as the extraction of detailed border-line definitions of the lesion border. In addition, Section 3.2 presents a comparison of the LBPC algorithm with other 39 segmentation algorithms, across three datasets and five performance metrics, setting a baseline of comparisons for future studies. This algorithm and study was published in Pereira et al. (2020a).

**Contribution 2: Assessment of segmentation-detail importance for classification**
This contribution is presented in Sections 3.4 and 4.3, and published in Pereira et al. (2020b) and in Pereira et al. (2021d). Previously, lesion border detail was not well expressed in the literature as a means of classification. With this research, evidence emerged confirming that segmentation-details can contribute to melanoma discrimination.

**Contribution 3: Creation of a light-field dataset of skin lesions**
The creation of a new dataset, with a new acquisition method and new data dimension, is also a relevant contribution, as presented in Section 4.1 (and published in Faria et al., 2019c,a), since it enables further advancements in the field of skin lesion classification by using light-field technology (Faria et al., 2019b).

**Contribution 4: Show the discriminative power of skin surface for classification**
This contribution is related with the approaches proposed in Sections 4.2 and 4.3. It is also evident in Sections 5.1 and 5.2, where the inclusion of features extracted from the depth dimension increase the classification metric performance (as expressed in Pereira et al., 2021c,d,b,a).

Finally, in addition to the experience gained, various journal articles and conference papers were submitted and published during this research as a result of the targeted goals. The following is a list of all scientific publications and presentations performed during the development of this work:

- **Pereira, P. M. M.**, Fonseca-Pinto, R., Paiva, R. P., Assuncao, P. A. A., Tavora, L. M. N., Thomaz, L. A., and Faria, S. M. M., Dermoscopic skin lesion image segmentation based on Local Binary Pattern Clustering: Comparative study, Biomedical Signal Processing and Control, vol.59, pp.1–12, 2020.
- **Pereira, P. M. M.**, Fonseca-Pinto, R., Paiva, R. P., Assuncao, P. A. A., Tavora, L. M. N., Thomaz, L. A., and Faria, S. M. M., Skin lesion classification enhancement using border-line features – The melanoma vs nevus problem, Biomedical Signal Processing and Control, vol.57, pp.1-8, 2020.
- **Pereira, P. M. M.**, Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., and Faria, S. M. M, Multiple Instance Learning using 3D Features for Melanoma Detection, IEEE Journal of Biomedical and Health Informatics, –, pp.–, Q1, 2021 *(submitted)*.
- **Pereira, P. M. M.**, Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., and Faria, S. M. M., Melanoma Classification using Light-Fields

with Morlet Scattering Transform and CNN: surface depth as a valuable tool to increase detection rate, Medical Image Analysis, Special Issue on Image Analysis in Dermatology, –, pp.–, 2021 *(submitted)*.

- **Pereira, P. M. M.**, Tavora, L. M. N., Fonseca-Pinto, R., Paiva, R. P., Assuncao, P. A. A., and Faria, S. M. M., Image Segmentation using Gradient-based Histogram Thresholding for Skin Lesion Delineation, 12th International Joint Conference on Biomedical Engineering Systems and Technologies, vol.2, pp.84-91, Prague, Czech Republic, February, 2019.

- **Pereira, P. M. M.**, Fonseca-Pinto, R., Paiva, R. P., Tavora, L. M. N., Assuncao, P. A. A., and Faria, S. M. M., Accurate Segmentation of Dermoscopic Images based on Local Binary Pattern Clustering, 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, pp.314-319, Opatija, Croatia, February, 2019.

- Faria, S. M. M., Santos, M., Assuncao, P. A. A., Tavora, L. M. N., Thomaz, L. A., **Pereira, P. M. M.**, Fonseca-Pinto, R., Santiago, F., Dominguez, V., and Henrique, M., Dermatological Imaging using a Focused Plenoptic Camera: the SKINL2 Light Field Dataset., Conference on Telecommunications, pp.1-4, Lisbon, Portugal, June, 2019.

- Faria, S. M. M., Filipe, J. N., **Pereira, P. M. M.**, Tavora, L. M. N., Assuncao, P. A. A., San-tos, M. O., Fonseca-Pinto, R., Santiago, F., Dominguez, V., and Henrique, M., Light Field Image Dataset of Skin Lesions, 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp.3905-3908, Berlin, Germany, July, 2019.

- **Pereira, P. M. M.**, Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., and Faria, S. M. M., Skin Lesion Classification using Bag-of-3D-Features, Conference on Telecommunications, pp. 1-6, Leiria, Portugal, February, 2021.

- **Pereira, P. M. M.**, Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., and Faria, S. M. M. Skin lesion classification using features of 3D border lines, 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp.1-6, Guadalajara, Mexico, October, 2021.

- **Pereira, P. M. M.**, Dermo-Plenoptic Imaging for Skin Surface Assessment, Presented at Encontro Ciência, Lisbon, Portugal, 2018.

- **Pereira, P. M. M.**, Melanoma Detection based on Light-Field Imaging, Presented at Encontro Ciência, Lisbon, Portugal, 2018.

- **Pereira, P. M. M.**, Skin Lesion Classification using Light-Field Imaging, Presented at Medical Imaging Summer School - Medical Imaging Meets Deep Learning, Favignana, Sicily, 2018.

## 6.3 Future Directions

Many different experiments and validations have been left for the future due to constraints of time and data availability. Future work can be focused on the expansion of the current dataset and the development of acquisition methods for other untapped dimensions.

### Short-term Perspectives

Continuing with the research made in this thesis, future work can look for more detailed methods for skin surface extraction from light-field imaging. This is possibility since extracted depth originated from the proprietary Raytrix software. This software is not open source, thus hindering possible improvements and adaptations focused on the skin lesion domain characteristics. The added detail provided from such improvement might unveil new skin characteristics and enhance the currently proposed methods. As an example of an improvement, future work could take into account the existence of hairs and, therefore, exclude such information from depth maps (by filling the hair regions with surrounding skin projected from other view angles), while developing a specialised skin surface extraction method.

The SKINL2 dataset, introduced in this thesis, must remain publicly available, such that future research can be built upon the proposed methods and new proposals can be developed. It is also relevant that the dataset continues to grow, not only to improve future classification results, but also to ease the current implementation constraints (in terms of data imbalance and quantity), and thus allow better modelling of testing and validation.

### Long-term Perspectives

Having concluded that the skin surface has relevant information when aiming to perform skin lesion classification, future work can consider looking deeper than the skin surface whilst still maintaining the portability and simplicity of the acquisition setup. This generally overlooked dimension would help dermatologists, as it provides a new source of information that, included in CAD systems, would allow to perform better informed judgements. Additionally, such dimensional information would enable research on deep-skin structures, with the potential of allowing better discrimination of existing types of skin conditions.

Construction of 3D models of deeper layers of the skin might enable the analysis of the lesions' 3D structures and the mapping of skin lesion structures to particular layers of the skin. This would automatically exclude certain diagnosis and allow to focus algorithms in particular skin lesion cases that can only occurs at such depths.

# Bibliography


Abbasi, N. R., Shaw, H. M., Rigel, D. S., Friedman, R. J., McCarthy, W. H., Osman, I., Kopf, A. W., & Polsky, D. (2004). Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria. *JAMA*, *292*(22), 2771–2776.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Khosravi, A., Acharya, U. R., Makarenkov, V., et al. (2020). A review of uncertainty quantification in deep learning: techniques, applications and challenges. *arXiv e-prints*, (pp. 1–61).

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(11), 2274–2282.

Adegun, A., & Viriri, S. (2020). Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. *Artifical Intelligence Review*, (pp. 1–31).

Adel, T., Cohen, T., Caan, M., Welling, M., study group, A., Initiative, A. D. N., et al. (2017). 3D scattering transforms for disease classification in neuroimaging. *NeuroImage: Clinical*, *14*, 506–517.

Adelson, E. H., & Wang, J. Y. (1992). Single lens stereo with a plenoptic camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *14*(2), 99–106.

Ahmad, W., Palmieri, L., Koch, R., & Sjöström, M. (2018). Matching light field datasets from plenoptic cameras 1.0 and 2.0. In *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video*, (pp. 1–4). Helsinki, Finland: IEEE.

Ahn, E., Bi, L., Jung, Y. H., Kim, J., Li, C., Fulham, M., & Feng, D. D. (2015). Automated saliency-based lesion segmentation in dermoscopic images. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 3009–3012). Milan, Italy: IEEE.

AJ, S., TB, F., MC, M., & et al (1979). Early recognition of cutaneous melanoma. *JAMA*, *242*(25), 2795–2799.

Akram, T., Khan, M. A., Sharif, M., & Yasmin, M. (2018a). Skin lesion segmentation and recognition using multichannel saliency estimation and M-SVM on selected serially fused features. *Journal of Ambient Intelligence and Humanized Computing*, (pp. 1–20).

Akram, T., Laurent, B., Naqvi, S. R., Alex, M. M., Muhammad, N., et al. (2018b). A deep heterogeneous feature fusion approach for automatic land-use classification. *Information Sciences*, *467*, 199–218.

Alcón, J. F., Ciuhu, C., Ten Kate, W., Heinrich, A., Uzunbajakava, N., Krekels, G., Siem, D., & De Haan, G. (2009). Automatic imaging system with decision support for inspection of pigmented skin lesions and melanoma diagnosis. *IEEE Journal of Selected Topics in Signal Processing*, *3*(1), 14–25.

Alliance, M. R. (2020). Melanoma statistics. https://www.curemelanoma.org/about-melanoma/melanoma-statistics-2/. Accessed: 2020-02-10.

Andén, J., & Mallat, S. (2014). Deep scattering spectrum. *IEEE Transactions on Signal Processing*, *62*(16), 4114–4128.

Anwar, S., Smith, L. N., & Smith, M. L. (2012). 3D Skin texture analysis: A neural network and photometric stereo perspective. In *International Conference on 3D Body Scanning Technologies*, (pp. 30–38). Lugano, Switzerland.

Ares Rodríguez, M., Royo Royo, S., Vilaseca Ricart, M., Herrera Ramírez, J. A., Delpueyo Español, X., & Sanàbria Ortega, F. (2014). Handheld 3D scanning system for in-vivo imaging of skin cancer. In *International Conference on 3D Body Scanning Technologies*, (pp. 231–236). Lugano, Switzerland.

Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., & Delfino, M. (1998). 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. *Archives of Dermatology*, *134*(12), 1563–1570.

Argenziano, G., Soyer, H., De Giorgi, V., Piccolo, D., Carli, P., & Delfino, M. (2000). *Interactive atlas of dermoscopy (Book and CD-ROM)*. Milan: EDRA Medical Publishing & New Media.

Argenziano, G., Soyer, H. P., Chimenti, S., Talamini, R., Corona, R., Sera, F., Binder, M., Cerroni, L., De Rosa, G., Ferrara, G., et al. (2003). Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet. *Journal of the American Academy of Dermatology*, *48*(5), 679–693.

Argenziano, G., Zalaudek, I., Ferrara, G., Hofmann-Wellenhof, R., & Soyer, H. (2007). Proposal of a new classification system for melanocytic naevi. *British Journal of Dermatology*, *157*(2), 217–227.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, (pp. 1027–1035). New Orleans, USA: Society for Industrial and Applied Mathematics.

Artzi, M., Bressler, I., & Ben Bashat, D. (2019). Differentiation between glioblastoma, brain metastasis and subtypes using radiomics analysis. *Journal of Magnetic Resonance Imaging*, *50*(2), 519–528.

Auger, F., & Flandrin, P. (1995). Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, *43*(5), 1068–1089.

Baghdadchi, S., Liu, K., Knapp, J., Prager, G., Graves, S., Akrami, K., Manuel, R., Bastos, R., Reid, E., Carson, D., et al. (2014). An innovative system for 3D clinical photography in the resource-limited settings. *Journal of Translational Medicine*, *12*(1), 169.

Baghersalimi, S., Bozorgtabar, B., Schmid-Saugeon, P., Ekenel, H. K., & Thiran, J.-P. (2019). Dermonet: densely linked convolutional neural network for efficient skin lesion segmentation. *Journal on Image and Video Processing*, *2019*(1), 71.

Baig, R., Bibi, M., Hamid, A., Kausar, S., & Khalid, S. (2020). Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images - a review. *Current Medical Imaging*, *16*(5), 513–533.

Ballerini, L., Fisher, R. B., Aldridge, B., & Rees, J. (2013). A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, vol. 6, (pp. 63–86). Springer. https://licensing.edinburgh-innovations.ed.ac.uk/i/software/dermofit-image-library.html.

Barata, C., Celebi, M. E., & Marques, J. S. (2015). Melanoma detection algorithm based on feature fusion. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 2653–2656). Milan, Italy: IEEE.

Barata, C., Celebi, M. E., & Marques, J. S. (2018). A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE Journal of Biomedical and Health Informatics*, *23*(3), 1096–1109.

Barata, C., Ruela, M., Francisco, M., Mendonça, T., & Marques, J. S. (2013). Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Systems Journal*, *8*(3), 965–979.

Barata, C., Ruela, M., Francisco, M., Mendonça, T., & Marques, J. S. (2014). Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Systems Journal*, *8*(3), 965–979.

Baratloo, A., Hosseini, M., Negida, A., & Ashal, G. (2015). Part 1: simple definition and calculation of accuracy, sensitivity and specificity. *Emergency*.

Bayraktar, M., Kockara, S., Halic, T., Mete, M., Wong, H. K., & Iqbal, K. (2019). Local edge-enhanced active contour for accurate skin lesion border detection. *BMC Bioinformatics*, *20*(2), 91.

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2018). Reconciling modern machine learning practice and the bias-variance trade-off. *arXiv preprint arXiv:1403.1687*, (pp. 1–23).

Berseth, M. (2017). ISIC 2017-Skin lesion analysis towards melanoma detection. *arXiv preprint arXiv:1703.00523*, (pp. 1–4).

Bezdek, J. (1981). *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York: Springer.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), 191–203.

Bi, L., Kim, J., Ahn, E., Feng, D., & Fulham, M. (2016). Automated skin lesion segmentation via image-wise supervised learning and multi-scale superpixel based cellular automata. In *IEEE International Symposium on Biomedical Imaging*, (pp. 1059–1062). Prague, Czech Republic: IEEE.

Bi, L., Kim, J., Ahn, E., Kumar, A., Feng, D., & Fulham, M. (2019). Step-wise integration of deep class-specific learning for dermoscopic image segmentation. *Pattern Recognition*, *85*, 78–89.

Bi, L., Kim, J., Ahn, E., Kumar, A., Fulham, M., & Feng, D. (2017). Dermoscopic image segmentation via multi-stage fully convolutional networks. *IEEE Transactions on Biomedical Engineering*, *64*(9), 2065–2074.

Bisla, D., Choromanska, A., Berman, R. S., Stein, J. A., & Polsky, D. (2019). Towards automated melanoma detection with deep learning: Data purification and augmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition. Workshops*, (pp. 2720–2728). Long Beach, CA, USA: IEEE.

Blum, A., Hofmann-Wellenhof, R., Luedtke, H., Ellwanger, U., Steins, A., Roehm, S., Garbe, C., & Soyer, H. (2004). Value of the clinical history for different users of dermoscopy compared with results of digital image analysis. *Journal of the European Academy of Dermatology and Venereology*, *18*(6), 665–669.

Bozorgtabar, B., Abedini, M., & Garnavi, R. (2016). Sparse coding based skin lesion segmentation using dynamic rule-based refinement. In *International Workshop on Machine Learning in Medical Imaging*, (pp. 254–261). Athens, Greece.

Bradley, D., & Roth, G. (2007). Adaptive thresholding using the integral image. *Journal of Graphics Tools*, *12*(2), 13–21.

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T., et al. (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, *113*, 47–54.

Bruna, J., & Mallat, S. (2013). Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1872–1886.

Bull, A. D. (2011). Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, *12*(10), 2879–2904.

Burdick, J., Marques, O., Romero-Lopez, A., Giró Nieto, X., & Weinthal, J. (2017). The impact of segmentation on the accuracy and sensitivity of a melanoma classifier based on skin lesion images. In *Society for Imaging Informatics in Medicine*, (pp. 1–6). Pittsburgh, PA, USA.

Carli, P., Chiarugi, A., & Giorgi, V. (2005). Examination of lesions (including dermoscopy) without contact with the patient is associated with improper management in about 30% of equivocal melanomas. *Dermatologic Surgery*, *31*(2), 169–172.

Celebi, M. E., Hwang, S., Iyatomi, H., & Schaefer, G. (2010). Robust border detection in dermoscopy images using threshold fusion. In *IEEE International Conference on Image Processing*, (pp. 2541–2544). Hong Kong, Hong Kong: IEEE.

Celebi, M. E., Iyatomi, H., Schaefer, G., & Stoecker, W. V. (2009a). Lesion border detection in dermoscopy images. *Computerized Medical Imaging and Graphics*, *33*(2), 148–153.

Celebi, M. E., Kingravi, H. A., Iyatomi, H., Alp Aslandogan, Y., Stoecker, W. V., Moss, R. H., Malters, J. M., Grichnik, J. M., Marghoob, A. A., Rabinovitz, H. S., et al. (2008). Border detection in dermoscopy images using statistical region merging. *Skin Research and Technology*, *14*(3), 347–353.

Celebi, M. E., Schaefer, G., Iyatomi, H., Stoecker, W. V., Malters, J. M., & Grichnik, J. M. (2009b). An improved objective evaluation measure for border detection in dermoscopy images. *Skin Research and Technology*, *15*(4), 444–450.

Chan, T. F., Sandberg, B. Y., & Vese, L. A. (2000). Active contours without edges for vector-valued images. *Journal of Visual Communication and Image Representation*, *11*(2), 130–141.

Chan, T. F., Vese, L. A., et al. (2001). Active contours without edges. *IEEE Transactions on Image Processing*, *10*(2), 266–277.

Chen, E. Z., Dong, X., Li, X., Jiang, H., Rong, R., & Wu, J. (2019). Lesion attributes segmentation for melanoma detection with multi-task U-Net. In *International Symposium on Biomedical Imaging*, (pp. 485–488). Venice, Italy: IEEE.

Cheng, I., Sun, X., Alsufyani, N., Xiong, Z., Major, P., & Basu, A. (2015). Ground truth delineation for medical image segmentation based on local consistency and distribution map analysis. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 3073–3076). Milan, Italy: IEEE.

Cho, C., Choi, W., & Kim, T. (2020). Leveraging uncertainties in softmax decision-making models for low-power iot devices. *Sensors*, *20*(16), 1–32.

Chudáček, V., Andén, J., Mallat, S., Abry, P., & Doret, M. (2013). Scattering transform for intrapartum fetal heart rate variability fractal analysis: A case-control study. *IEEE Transactions on Biomedical Engineering*, *61*(4), 1100–1108.

Cichorek, M., Wachulska, M., Stasiewicz, A., & Tymińska, A. (2013). Skin melanocytes: biology and development. *Advances in Dermatology and Allergology/Postępy Dermatologii i Alergologii*, *30*, 30–41.

Claridge, E., & Orun, A. (2002). Modelling of edge profiles in pigmented skin lesions. In *Medical Image Understanding and Analysis*, (pp. 53–56). Portsmouth, United Kingdom.

Clemente, C., Cochran, A. J., Elder, D. E., Levene, A., Mackie, R. M., Mihm, M. C., Rilke, F., Cascinelli, N., Fitzpatrick, T. B., & Sober, A. J. (1991). Histopathologic diagnosis of dysplastic nevi: concordance among pathologists convened by the world health organization melanoma programme. *Human Pathology*, *22*(4), 313–319.

Cohen, A. (1994). Ten Lectures on Wavelets, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 61, I. Daubechies, SIAM, 1992, xix+ 357 pp. *Journal of Approximation Theory*, *78*(3), 460–461.

Collaboration, I. (2017). International skin imaging collaboration: Melanoma project website. https://isic-archive.com/. Accessed: 01.12.2017.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603–619.

Csáji, B. C. (2001). Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, *24*, 48.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision*, (pp. 1–2). Prague, Czech Republic.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314.

Dansereau, D. G., Pizarro, O., & Williams, S. B. (2015). Linear volumetric focus for light field cameras. *ACM Transactions on Graphics*, *34*(2), 15–1.

Day, G., & Barbour, R. (2001). Automated skin lesion screening–a new approach. *Melanoma Research*, *11*(1), 31–35.

Dekker, A. H. (1994). Kohonen neural networks for optimal colour quantization. *Network: Computation in Neural Systems*, *5*(3), 351–367.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 248–255). Miami, FL, USA: IEEE.

DeVries, T., & Taylor, G. W. (2018). Leveraging uncertainty estimates for predicting segmentation quality. *arXiv e-prints*, (pp. 1–9).

Dhawan, A. P., Gordon, R., & Rangayyan, R. M. (1984). Nevoscopy: three-dimensional computed tomography of nevi and melanomas in situ by transillumination. *IEEE Transactions on Medical Imaging*, *3*(2), 54–61.

Ding, Y., John, N. W., Smith, L., Sun, J., & Smith, M. (2015). Combination of 3D skin surface texture features and 2D ABCD features for improved melanoma diagnosis. *Medical & Biological Engineering & Computing*, *53*(10), 961–974.

Domínguez-Espinosa, A. E. (2014). History of dermoscopy. *Dermatología Revista mexicana*, *58*(2), 165–172.

Donatsch, D., Bigdeli, S. A., Robert, P., & Zwicker, M. (2014). Hand-held 3D light field photography and applications. *The Visual Computer*, *30*(6), 897–907.

Ebrahimi, T., Foessel, S., Pereira, F., & Schelkens, P. (2016). JPEG Pleno: Toward an efficient representation of visual reality. *IEEE MultiMedia*, *23*(4), 14–20.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

Faria, S., Santos, M., Assuncao, P., Tavora, L., Thomaz, L., Pereira, P., Fonseca-Pinto, R., Santiago, F., Dominguez, V., & Henrique, M. (2019a). Dermatological imaging using a focused plenoptic camera: the SKINL2 light field dataset. In *Conference on Telecommunications*, (pp. 1–4). Lisbon, Portugal. http://on.ipleiria.pt/plenoisla.

Faria, S. M. M., Filipe, J. N., Assuncao, P. A. A., Santos, M. O., Fonseca-Pinto, R., Pereira, P. M. M., Tavora, L. M. N., Santiago, F., Dominguez, V., & Henrique, M. (2019b). Coding of still pictures. Doc. ISO/IEC JTC1/SC29/WG1 M82037.

Faria, S. M. M., Filipe, J. N., Pereira, P. M. M., Tavora, L. M. N., Assuncao, P. A. A., Santos, M. O., Fonseca-Pinto, R., Santiago, F., Dominguez, V., & Henrique, M. (2019c). Light field image dataset of skin lesions. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 3905–3908). Berlin, Germany: IEEE. http://on.ipleiria.pt/plenoisla.

Feng, H., Berk-Krauss, J., Feng, P. W., & Stein, J. A. (2018). Comparison of dermatologist density between urban and rural counties in the united states. *JAMA Dermatology*, *154*(11), 1265–1271.

Fernando, K. R. M., & Tsokos, C. P. (2021). Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *Early Access*, 1–12.

Fleming, M. G., Steger, C., Zhang, J., Gao, J., Cognetta, A. B., Dyer, C. R., et al. (1998). Techniques for a structural analysis of dermatoscopic imagery. *Computerized Medical Imaging and Graphics*, *22*(5), 375–389.

Fox, G. (2005). ABCD-EFG for diagnosis of melanoma. *Clinical and Experimental Dermatology*, *30*(6), 707–707.

Friedman, R. J., Rigel, D. S., & Kopf, A. W. (1985). Early detection of malignant melanoma: The role of physician examination and self-examination of the skin. *CA: A Cancer Journal for Clinicians*, *35*(3), 130–151.

Frome, A., Huber, D., Kolluri, R., Bülow, T., & Malik, J. (2004). Recognizing objects in range data using regional point descriptors. In *European Conference on Computer Vision*, (pp. 224–237). Prague, Czech Republic.

Fukunaga, K., & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, *21*(1), 32–40.

Garnavi, R., Aldeen, M., & Celebi, M. (2011). Weighted performance index for objective evaluation of border detection methods in dermoscopy images. *Skin Research and Technology*, *17*(1), 35–44.

Gelbart, M. A., Snoek, J., & Adams, R. P. (2014). Bayesian optimization with unknown constraints. *arXiv e-prints*.

Georgiev, T. (2009). New results on the plenoptic 2.0 camera. In *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, (pp. 1243–1247). Pacific Grove, CA, USA: IEEE.

Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., Werner, R., & Schlaefer, A. (2019). Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. *IEEE Transactions on Biomedical Engineering*, *67*(2), 495–503.

Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M. F., & Petkov, N. (2015). MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Systems with Applications*, *42*(19), 6578–6585.

Goldsmith, L. A., Askin, F. B., Chang, A. E., Cohen, C., Dutcher, J. P., Gilgor, R. S., Green, S., Harris, E. L., Havas, S., Robinson, J. K., et al. (1992). Diagnosis and treatment of early melanoma: NIH consensus development panel on early melanoma. *JAMA*, *268*(10), 1314–1319.

Gonzalez-Diaz, I. (2018). DermaKNet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 547–559.

Grob, J., & Bonerandi, J. (1998). The 'ugly duckling' sign: Identification of the common characteristics of nevi in an individual as a basis for melanoma screening. *Archives of Dermatology*, *134*(1), 103–104.

Guillo, L., Jiang, X., Lafruit, G., & Guillemot, C. (2018). Light field video dataset captured by a R8 raytrix camera. Tech. rep., ISO/IEC JTC1/SC29/WG1 & WG11.

Gustafson, D. E., & Kessel, W. C. (1979). Fuzzy clustering with a fuzzy covariance matrix. In *IEEE Conference on Decision and Control including Symposium on Adaptive Processes*, (pp. 761–766). San Diego: IEEE.

H., K., Johannsen, O., Kondermann, D., & Goldlücke, B. (2016). A dataset and evaluation methodology for depth estimation on 4D light fields. In *Asian Conference on Computer Vision*, vol. 10113, (pp. 19–34). Taipei, Taiwan: Springer.

Haddad, R. A., & Akansu, A. N. (1991). A class of fast gaussian binomial filters for speech and image processing. *IEEE Transactions on Signal Processing*, *39*(3), 723–727.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., et al. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, *29*(8), 1836–1842.

Hagerty, J. R., Stanley, R. J., Almubarak, H. A., Lama, N., Kasmi, R., Guo, P., Drugge, R. J., Rabinovitz, H. S., Oliviero, M., & Stoecker, W. V. (2019). Deep learning and handcrafted method fusion: higher diagnostic accuracy for melanoma dermoscopy images. *IEEE Journal of Biomedical and Health Informatics*, *23*(4), 1385–1391.

Hameed, N., Shabut, A., & Hossain, M. A. (2018). A computer-aided diagnosis system for classifying prominent skin lesions using machine learning. In *Computer Science and Electronic Engineering*, (pp. 186–191). Colchester, UK.

Han, S. S., Kim, M. S., Lim, W., Park, G. H., Park, I., & Chang, S. E. (2018). Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology*, *138*(7), 1529–1538.

Hance, G. A., Umbaugh, S. E., Moss, R. H., & Stoecker, W. V. (1996). Unsupervised color image segmentation: with application to skin tumor borders. *IEEE Engineering in Medicine and Biology Magazine*, *15*(1), 104–111.

Haralick, R. M., & Shapiro, L. G. (1992). *Computer and robot vision*. Addison-wesley.

Harangi, B. (2017). Skin lesion detection based on an ensemble of deep convolutional neural network. *arXiv preprint arXiv:1705.03360*, (pp. 1–4).

Harris, M. (2012). Focusing on everything. *IEEE Spectrum*, *49*(5), 44–50.

He, K., Girshick, R., & Dollár, P. (2019). Rethinking ImageNet pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 4918–4927). Seoul, South Korea: IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 770–778). Las Vegas, NV, USA: IEEE.

Heckbert, P. (1982). *Color image quantization for frame buffer display*, vol. 16. ACM.

Henning, J. S., Dusza, S. W., Wang, S. Q., Marghoob, A. A., Rabinovitz, H. S., Polsky, D., & Kopf, A. W. (2007). The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy. *Journal of the American Academy of Dermatology*, *56*(1), 45–52.

Hosny, K. M., Kassem, M. A., & Foaud, M. M. (2019). Classification of skin lesions using transfer learning and augmentation with Alex-net. *PLOS ONE*, *14*(5), 1–17.

Hu, K., Niu, X., Liu, S., Zhang, Y., Cao, C., Xiao, F., Yang, W., & Gao, X. (2019). Classification of melanoma based on feature similarity measurement for codebook learning in the bag-of-features model. *Biomedical Signal Processing and Control*, *51*, 200–209.

Huang, L.-K., & Wang, M.-J. J. (1995). Image thresholding by minimizing the measures of fuzziness. *Pattern Recognition*, *28*(1), 41–51.

Ilea, D. E., & Whelan, P. F. (2006a). Automatic segmentation of skin cancer images using adaptive color clustering. In *China-Ireland International Conference on Information and Communications Technologies*. Hangzhou, China.

Ilea, D. E., & Whelan, P. F. (2006b). Color image segmentation using a spatial k-means clustering algorithm. In *International Machine Vision and Image Processing Conference*. Dublin, Ireland.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *PMLR International Conference on Machine Learning*, (pp. 448–456). Lille, France.

ITU-R (2011). Studio encoding parameters of digital television for standard 4: 3 and widescreen 16: 9 aspect ratios. Recommendation BT.601-7, International Telecommunication Union, Geneva, CH.

Iyatomi, H., Norton, K.-A., Celebi, M. E., Schaefer, G., Tanaka, M., & Ogawa, K. (2010). Classification of melanocytic skin lesions from non-melanocytic lesions. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 5407–5410). Buenos Aires, Argentina: IEEE.

Iyatomi, H., Oka, H., Celebi, M. E., Ogawa, K., Argenziano, G., Soyer, H. P., Koga, H., Saida, T., Ohara, K., & Tanaka, M. (2008). Computer-based classification of dermoscopy images of melanocytic lesions on acral volar skin. *Journal of Investigative Dermatology*, *128*(8), 2049–2054.

Iyatomi, H., Oka, H., Saito, M., Miyake, A., Kimoto, M., Yamagami, J., Kobayashi, S., Tanikawa, A., Hagiwara, M., Ogawa, K., et al. (2006). Quantitative assessment of tumour extraction from dermoscopy images and evaluation of computer-based extraction methods for an automatic melanoma diagnostic system. *Melanoma Research*, *16*(2), 183–190.

Jafari, M. H., Samavi, S., Karimi, N., Soroushmehr, S. M. R., Ward, K., & Najarian, K. (2016). Automatic detection of melanoma using broad extraction of features from digital images. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 1357–1360). Orlando, FL, USA: IEEE.

Jaffard, S., Lashermes, B., & Abry, P. (2006). Wavelet leaders in multifractal analysis. In *Wavelet Analysis and Applications*, (pp. 201–246). Springer.

Jeon, H.-G., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.-W., & So Kweon, I. (2015). Accurate depth map estimation from a lenslet light field camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1547–1555). Boston, MA, USA: IEEE.

Jiménez, A. A., Márquez, F. P. G., Moraleda, V. B., & Muñoz, C. Q. G. (2019). Linear and nonlinear features and machine learning for wind turbine blade ice detection and diagnosis. *Renewable Energy*, *132*, 1034–1048.

Joel, G., Philippe, S.-S., David, G., Jean Philippe, C., Ralph, B., Joakim, K., Jean-Hilaire, S., & Murat, K. (2002). Validation of segmentation techniques for digital dermoscopy. *Skin Research and Technology*, *8*(4), 240–249.

Johannsen, O., Heinze, C., Goldluecke, B., & Perwaß, C. (2013). *On the Calibration of Focused Plenoptic Cameras*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Joint ISO/CIE Standard, ISO 11664-4:2008(E)/CIE S 014-4/E:2007 (2007). Colorimetry-Part 4: CIE 1976 L* a* b* Colour space.

Kanezaki, A., Marton, Z., Pangercic, D., Harada, T., Kuniyoshi, Y., & Beetz, M. (2011). Voxelized shape and color histograms for RGB-D. In *IROS Workshop on Active Semantic Perception*, (pp. 1–6). San Francisco, CA, USA.

Kapur, J. N., Sahoo, P. K., & Wong, A. K. (1985). A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing*, *29*(3), 273–285.

Kaufman, H. L. (2005). *The melanoma book: a complete guide to prevention and treatment*, vol. 1. New York, NY, USA: Gotham Books, 1st ed.

Kawahara, J., BenTaieb, A., & Hamarneh, G. (2016). Deep features to classify skin lesions. In *IEEE International Symposium on Biomedical Imaging*, (pp. 1397–1400). Prague, Czech Republic: IEEE.

Kay, S. (1988). *Modern spectral estimation: theory and application*. Pearson Education India.

Kéchichian, R., Gong, H., Revenu, M., Lezoray, O., & Desvignes, M. (2014). New data model for graph-cut segmentation: Application to automatic melanoma delineation. In *International Conference on Image Processing*, (pp. 892–896). IEEE.

Kecman, V., Huang, T., & Vogt, M. (2005a). Iterative single data algorithm for training kernel machines from huge data sets: Theory and performance. In *Support Vector Machines: Theory and Applications*, (pp. 255–274). Springer.

Kecman, V., Huang, T.-M., & Vogt, M. (2005b). *Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory and Performance*, (pp. 255–274). Berlin, Heidelberg: Springer.

Keeler, J. D., Rumelhart, D. E., & Leow, W. K. (1990). Integrated segmentation and recognition of hand-printed numerals. In *International Conference on Neural Information Processing Systems*, (pp. 557–563). Denver, CO, USA.

Ker, J., Wang, L., Rao, J., & Lim, T. (2018). Deep learning applications in medical image analysis. *IEEE Access*, *6*, 9375–9389.

Khan, S., Hayat, M., Zamir, S. W., Shen, J., & Shao, L. (2019). Striking the right balance with uncertainty. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 103–112). Long Beach, CA, USA: IEEE.

Kiefer, J. (2007). Effects of ultraviolet radiation on DNA. In *Chromosomal Alterations*, (pp. 39–53). Springer.

Kini, P., & Dhawan, A. P. (1992). Three-dimensional imaging and reconstruction of skin lesions. *Computerized Medical Imaging and Graphics*, *16*(3), 153–161.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine Learning Proceedings*, (pp. 249–256). Elsevier.

Kittler, H., Pehamberger, H., Wolff, K., & Binder, M. (2002). Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, *3*(3), 159–165.

Koehoorn, J., Sobiecki, A. C., Boda, D., Diaconeasa, A., Doshi, S., Paisey, S., Jalba, A., & Telea, A. (2015). Automated digital hair removal by threshold decomposition and morphological analysis. In *International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing*, (pp. 15–26). Reykjavik, Iceland.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (pp. 171–182). Springer.

Korotkov, K., & Garcia, R. (2012). Computerized analysis of pigmented skin lesions: A review. *Artificial Intelligence in Medicine*, *56*(2), 69–90.

Kreusch, J., & Rassner, G. (1990). Structural analysis of melanocytic pigment nevi using epiluminescence microscopy. review and personal experiences. *Der Hautarzt; Zeitschrift fur Dermatologie, Venerologie, und verwandte Gebiete*, *41*(1), 27–33.

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. *arXiv e-prints*, (pp. 1–7).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

L., T., & Z., M. (2016). ECG classification using wavelet packet entropy and random forests. *Entropy*, *18*(8), 285.

Lalitha, V., & Geetha, G. (2014). Automated melanoma detection-a review. *International Journal of Advanced Information Science and Technology*, *27*(27), 1–6.

Lankton, S., & Tannenbaum, A. (2008). Localizing region-based active contours. *IEEE Transactions on Image Processing*, *17*(11), 2029–2039.

Lazebnik, S., Schmid, C., & Ponce, J. (2005). A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1265–1278.

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. In *IEEE International Symposium on Circuits and Systems*, (pp. 253–256). Paris, France: IEEE.

Lee, H. D., Mendes, A. I., Spolaor, N., Oliva, J. T., Parmezan, A. R. S., Wu, F. C., & Fonseca-Pinto, R. (2018). Dermoscopic assisted diagnosis in melanoma: Reviewing results, optimizing methodologies and quantifying empirical guidelines. *Knowledge-Based Systems*, *158*, 9–24.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, *7*(1), 1–14.

Lemley, J., Bazrafkan, S., & Corcoran, P. (2017). Smart augmentation learning an optimal data augmentation strategy. *IEEE Access*, *5*, 5858–5869.

Leonarduzzi, R., Schlotthauer, G., & Torres, M. (2010). Wavelet leader based multifractal analysis of heart rate variability during myocardial ischaemia. In *Annual International Conference of the IEEE Engineering in Medicine and Biology*, (pp. 110–113). Buenos Aires, Argentina: IEEE.

Levoy, M. (2006). Light fields and computational imaging. *Computer*, *39*(8), 46–55.

Levoy, M., & Hanrahan, P. (1996). Light field rendering. In *Annual Conference on Computer Graphics and Interactive Techniques*, (pp. 31–42). New York, NY, USA: ACM.

Li, B. N., Chui, C. K., Chang, S., & Ong, S. H. (2011). Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation. *Computers in Biology and Medicine*, *41*(1), 1–10.

Li, C. H., & Lee, C. (1993). Minimum cross entropy thresholding. *Pattern Recognition*, *26*(4), 617–625.

Li, Y., Sun, J., Tang, C.-K., & Shum, H.-Y. (2004). Lazy snapping. *Transactions on Graphics*, *23*(3), 303–308.

Li, Z., Zhang, X., Müller, H., & Zhang, S. (2018). Large-scale retrieval for medical image analytics: a comprehensive review. *Medical Image Analysis*, *43*, 66–84.

Liao, H., Li, Y., & Luo, J. (2016). Skin disease classification versus skin lesion characterization: Achieving robust diagnosis using multi-label deep neural networks. In *International Conference on Pattern Recognition*, (pp. 355–360). Cancun, Mexico: IEEE.

Lima, J. P. S. d. M., & Teichrieb, V. (2016). An efficient global point cloud descriptor for object recognition and pose estimation. In *Conference on Graphics, Patterns and Images*, (pp. 56–63). Sao Paulo, Brazil.

Linsangan, N. B., Adtoon, J. J., & Torres, J. L. (2018). Geometric analysis of skin lesion for skin cancer using image processing. In *International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management*, (pp. 1–5). Baguio, Philippines: IEEE.

Lippmann, G. (1908a). Épreuves rréversibles donnant la sensation du relief. *Journal of Physics: Theories and Applications*, *7*(1), 821–825.

Lippmann, G. (1908b). La photographic integrale. *French Academie des Sciences*, *146*, 446–451.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., et al. (2020). A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, *26*(6), 900–908.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, *28*(2), 129–137.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 3431–3440). Boston, United States: IEEE.

Machado, M., Pereira, J., & Fonseca-Pinto, R. (2016). Reticular pattern detection in dermoscopy: an approach using curvelet transform. *Research on Biomedical Engineering*, *32*(2), 129–136.

MacKIE, R. (1971). An aid to the preoperative assessment of pigmented lesions of the skin. *British Journal of Dermatology*, *85*(3), 232–238.

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, (pp. 281–297). Oakland, CA, USA.

Maglogiannis, I., & Doukas, C. N. (2009). Overview of advanced computer vision systems for skin lesions characterization. *IEEE Transactions on Information Technology in Biomedicine*, *13*(5), 721–733.

Maharaj, E. A., & Alonso, A. (2014). Discriminant analysis of multivariate time series: Application to diagnosis based on ECG signals. *Computational Statistics & Data Analysis*, *70*, 67–87.

Mahdiraji, S. A., Baleghi, Y., & Sakhaei, S. M. (2018). BIBS, a new descriptor for melanoma/non-melanoma discrimination. In *Iranian Conference on Electrical Engineering*, (pp. 1397–1402). Iran, Mashhad.

Mahmouei, S. S., Aldeen, M., Stoecker, W. V., & Garnavi, R. (2018). Biologically inspired quadtree color detection in dermoscopy images of melanoma. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 570–577.

Majumder, S., & Ullah, M. A. (2018). Feature extraction from dermoscopy images for an effective diagnosis of melanoma skin cancer. In *International Conference on Electrical and Computer Engineering*, (pp. 185–188). Dhaka, Bangladesh.

Makanjuola, J. K., Aggoun, A., Swash, M., Grange, P. C., Challacombe, B., & Dasgupta, P. (2013). 3D-holoscopic imaging: A new dimension to enhance imaging in minimally invasive therapy in urologic oncology. *Journal of Endourology*, *27*(5), 535–539.

Mallat, S. (2012). Group invariant scattering. *Communications on Pure and Applied Mathematics*, *65*(10), 1331–1398.

Malvehy, J., Puig, S., Argenziano, G., Marghoob, A. A., Soyer, H. P., et al. (2007). Dermoscopy report: proposal for standardization: results of a consensus meeting of the international dermoscopy society. *Journal of the American Academy of Dermatology*, *57*(1), 84–95.

Mane, S., & Shinde, S. (2018). A method for melanoma skin cancer detection using dermoscopy images. In *International Conference on Computing Communication Control and Automation*, (pp. 1–6). Maharashtra, India.

Marchetti, M. A., Codella, N. C., Dusza, S. W., Gutman, D. A., Helba, B., Kalloo, A., Mishra, N., Carrera, C., Celebi, M. E., DeFazio, J. L., et al. (2018). Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *Journal of the American Academy of Dermatology*, *78*(2), 270–277.

Maron, O., & Ratan, A. L. (1998). Multiple-instance learning for natural scene classification. In *International Conference on Machine Learning*, vol. 98, (pp. 341–349). Madison, WI, USA.

Marshall, R., Styles, I., Claridge, E., & Bongs, K. (2014). Plenoptic imaging of the retina: can it resolve depth in scattering tissues? In *Biomedical Optics*, (pp. BM3A–60). Miami, FL, USA: Optical Society of America.

Marton, Z., Pangercic, D., Blodow, N., & Beetz, M. (2011). Combined 2D-3D categorization and classification for multimodal perception systems. *The International Journal of Robotics Research*, *30*(11), 1378–1402.

Marton, Z., Pangercic, D., Blodow, N., Kleinehellefort, J., & Beetz, M. (2010). General 3D modelling of novel objects from a single view. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (pp. 3700–3705). Taipei, Taiwan: IEEE.

Materka, A., & Strzelecki, M. (1998). Texture analysis methods - a review. Tech. rep., Institute of Electronics, Technical University of Lodz.

Matsumura, Y., & Ananthaswamy, H. N. (2004). Toxic effects of ultraviolet radiation on the skin. *Toxicology and applied pharmacology*, *195*(3), 298–308.

McDonagh, S., Fisher, R., & Rees, J. (2008). Using 3D information for classification of non-melanoma skin lesions. In *Medical Image Understanding and Analysis*, (pp. 164–168). Dundee, United Kingdom: BMVA Press.

McGrath, J., & Uitto, J. (2010). Anatomy and organization of human skin. *Rook's Textbook of Dermatology*, *8*, 1–53.

Mendes, A. I., Nogueira, C., Pereira, J., & Fonseca-Pinto, R. (2016). On the geometric modulation of skin lesion growth: a mathematical model for melanoma. *Research on Biomedical Engineering*, *32*(1), 44–54.

Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R., & Rozeira, J. (2013). Ph2-a dermoscopic image database for research and benchmarking. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 5437–5440). Osaka, Japan: IEEE. http://www.fc.up.pt/addi/ph2%20database.html.

Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F. V., Avila, S., & Valle, E. (2017). Knowledge transfer for melanoma screening with deep learning. *arXiv preprint arXiv:1703.07479*, (pp. 1–4).

Menzies, S. W., Ingvar, C., Crotty, K. A., & McCarthy, W. H. (1996). Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features. *Archives of Dermatology*, *132*(10), 1178–1182.

Mishkin, D., Sergievskiy, N., & Matas, J. (2017). Systematic evaluation of convolution neural network advances on the ImageNet. *Computer Vision and Image Understanding*, *161*, 11–19.

Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, *6*(4), 525–533.

Murphree, D. H., & Ngufor, C. (2017). Transfer learning for melanoma detection: Participation in ISIC 2017 skin lesion classification challenge. *arXiv preprint arXiv:1703.05235*, (pp. 1–3).

Nachbar, F., Stolz, W., Merkle, T., Cognetta, A. B., Vogt, T., Landthaler, M., Bilek, P., Braun-Falco, O., & Plewig, G. (1994). The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, *30*(4), 551–559.

Nadeau, C., & Bengio, Y. (2000). Inference for the generalization error. In *Advances in Neural Information Processing Systems*, (pp. 307–313).

Namozov, A., & Cho, Y. I. (2018). Convolutional neural network algorithm with parameterized activation function for melanoma classification. In *International Conference on Information and Communication Technology Convergence*, (pp. 417–419). Jeju Island, Korea: IEEE.

Navarro, F., Escudero-Viñolo, M., & Bescós, J. (2018). Accurate segmentation and registration of skin lesion images to evaluate lesion change. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 501–508.

Nehal, K., Oliveria, S., Marghoob, A., Christos, P., Dusza, S., Tromberg, J., & Halpern, A. (2002). Use of and beliefs about dermoscopy in the management of patients with pigmented lesions: a survey of dermatology residency programmes in the United States. *Melanoma Research*, *12*(6), 601–605.

Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., & Hanrahan, P. (2005). Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report*, *2*(11), 1–11.

Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, *29*(1), 51–59.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2000). Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision*, (pp. 404–420). Heidelberg, Germany: Springer.

Oliveira, R. B., Mercedes Filho, E., Ma, Z., Papa, J. P., Pereira, A. S., & Tavares, J. M. R. (2016). Computational methods for the image segmentation of pigmented skin lesions: a review. *Computer Methods and Programs in Biomedicine*, *131*, 127–141.

Oliveira, R. B., Papa, J. P., Pereira, A. S., & Tavares, J. M. R. (2018). Computational methods for pigmented skin lesion classification in images: review and future trends. *Neural Computing and Applications*, *29*(3), 613–636.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66.

Pampena, R., Kyrgidis, A., Lallas, A., Moscarella, E., Argenziano, G., & Longo, C. (2017). A meta-analysis of nevus-associated melanoma: Prevalence and practical implications. *Journal of the American Academy of Dermatology*, *77*(5), 938–945.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345–1359.

Pathan, S., Prabhu, K. G., & Siddalingaswamy, P. (2018). Techniques and algorithms for computer aided diagnosis of pigmented skin lesions – a review. *Biomedical Signal Processing and Control*, *39*, 237–262.

Peña Gutiérrez, S. (2016). *Novel melanoma diagnosis and prognosis methods based on 3D fringe projection*. B.S. thesis, Universitat Politècnica de Catalunya, Barcelona, Spain.

Pereira, J., Mendes, A., Nogueira, C., Baptista, D., & Fonseca-Pinto, R. (2015). An adaptive approach for skin lesion segmentation in dermoscopy images using a multiscale local normalization. In *Dynamics, Games and Science*, (pp. 537–545). Springer.

Pereira, P., Fonseca-Pinto, R., Paiva, R., Tavora, L., Assuncao, P., & Faria, S. (2018). Transfer learning of imagenet neural network for pigmented skin lesion detection. In *Portuguese Conference on Pattern Recognition*, (pp. 26–27). Coimbra, Portugal.

Pereira, P. M. M., Fonseca-Pinto, R., Paiva, R. P., Assuncao, P. A. A., Tavora, L. M. N., Thomaz, L. A., & Faria, S. M. M. (2020a). Dermoscopic skin lesion image segmentation based on local binary pattern clustering: Comparative study. *Biomedical Signal Processing and Control*, *59 (101924)*, 1–12.

Pereira, P. M. M., Fonseca-Pinto, R., Paiva, R. P., Assuncao, P. A. A., Tavora, L. M. N., Thomaz, L. A., & Faria, S. M. M. (2020b). Skin lesion classification enhancement using border-line features – The melanoma vs nevus problem. *Biomedical Signal Processing and Control*, *57*, 101765.

Pereira, P. M. M., Fonseca-Pinto, R., Paiva, R. P., Tavora, L. M. N., Assuncao, P. A. A., & Faria, S. M. M. (2019a). Accurate segmentation of dermoscopic images based on local binary pattern clustering. In *International Convention on Information and Communication Technology, Electronics and Microelectronics*, (pp. 314–319). Opatija, Croatia.

Pereira, P. M. M., Tavora, L. M. N., Fonseca-Pinto, R., Paiva, R. P., Assuncao, P. A. A., & Faria, S. M. M. (2019b). Image segmentation using gradient-based histogram thresholding for skin lesion delineation. In *International Joint Conference on Biomedical Engineering Systems and Technologies*, vol. 2, (pp. 84–91). Prague, Czech Republic.

Pereira, P. M. M., Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., & Faria, S. M. M. (2021a). Melanoma classification using light-fields with morlet scattering transform and CNN: surface depth as a valuable tool to increase detection rate. *IEEE Medical Image Analysis*, –, –. **Submitted to** Special Issue on Image Analysis in Dermatology.

Pereira, P. M. M., Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., & Faria, S. M. M. (2021b). Multiple instance learning using 3D features for melanoma detection. *IEEE Journal of Biomedical and Health Informatics*, –, –. **submitted to**.

Pereira, P. M. M., Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., & Faria, S. M. M. (2021c). Skin lesion classification using bag-of-3D-features. In *Conference on Telecommunications*, (pp. 1–6). Leiria, Portugal.

Pereira, P. M. M., Thomaz, L. A., Tavora, L. M. N., Assuncao, P. A. A., Fonseca-Pinto, R., Paiva, R. P., & Faria, S. M. M. (2021d). Skin lesion classification using features of 3D border lines. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 1–6). Guadalajara, Mexico: IEEE.

Pizzichetta, M. A., Talamini, R., Piccolo, D., Argenziano, G., Pagnanelli, G., Burgdorf, T., Lombardi, D., Trevisan, G., Veronesi, A., Carbone, A., et al. (2001). The ABCD rule of dermatoscopy does not apply to small melanocytic skin lesions. *Archives of Dermatology*, *137*(10), 1376–1378.

Platt, J. C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Tech. rep., Advances in kernel methods - support vector learning.

Psaty, E. L., & Halpern, A. C. (2009). Current and emerging technologies in melanoma diagnosis: the state of the art. *Clinics in Dermatology*, *27*(1), 35–45.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P. (2020). Designing network design spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 10428–10436). Virtual.

Rahman, M. M., Bhuiyan, M. I. H., & Das, A. B. (2019). Classification of focal and non-focal EEG signals in VMD-DWT domain using ensemble stacking. *Biomedical Signal Processing and Control*, *50*, 72–82.

Rajab, M., Woolfson, M., & Morgan, S. (2004). Application of region-based segmentation and neural network edge detection to skin lesions. *Computerized Medical Imaging and Graphics*, *28*(1), 61–68.

Rastgoo, M., Garcia, R., Morel, O., & Marzani, F. (2015). Automatic differentiation of melanoma from dysplastic nevi. *Computerized Medical Imaging and Graphics*, *43*, 44–52.

Ravanat, J.-L., Douki, T., & Cadet, J. (2001). Direct and indirect effects of UV radiation on DNA and its components. *Journal of Photochemistry and Photobiology B: Biology*, *63*(1), 88–102.

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G.-Z. (2016). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 4–21.

Raytrix (2018). Raytrix GmbH. https://raytrix.de/. Accessed: 24.09.2018.

Raytrix (2019). Raytrix downloads. https://raytrix.de/downloads/. Accessed: 05.04.2019.

Rerabek, M., & Ebrahimi, T. (2016). New light field image dataset. In *International Conference on Quality of Multimedia Experience*. Lisbon, Portugal.

Riaz, F., Hassan, A., Javed, M. Y., & Coimbra, M. T. (2014). Detecting melanoma in dermoscopy images using scale adaptive local binary patterns. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 6758–6761). Chicago, IL, USA: IEEE.

Riaz, F., Naeem, S., Nawaz, R., & Coimbra, M. (2019). Active contours based segmentation and lesion periphery analysis for characterization of skin lesions in dermoscopy images. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 489–500.

Rigel, D. S., Russak, J., & Friedman, R. (2010). The evolution of melanoma diagnosis: 25 years beyond the ABCDs. *CA: A Cancer Journal for Clinicians*, *60*(5), 301–316.

Robnik-Šikonja, M., & Kononenko, I. (1997). An adaptation of relief for attribute estimation in regression. In *International Conference on Machine Learning*, vol. 5, (pp. 296–304). Nashville, Tennessee, USA.

Roy, A., Pal, A., & Garain, U. (2017). JCLMM: A finite mixture model for clustering of circular-linear data and its application to psoriatic plaque segmentation. *Pattern Recognition*, *66*, 160–173.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Rusu, R., Blodow, N., & Beetz, M. (2009a). Fast point feature histograms (FPFH) for 3D registration. In *IEEE International Conference on Robotics and Automation*, (pp. 3212–3217). Kobe, Japan: IEEE.

Rusu, R., Blodow, N., Marton, Z., & Beetz, M. (2008). Aligning point cloud views using persistent feature histograms. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (pp. 3384–3391). Nice, France: IEEE.

Rusu, R., & Cousins, S. (2011). 3D is here: Point cloud library (pcl). In *IEEE International Conference on Robotics and Automation*, (pp. 1–4). Shanghai, China: IEEE.

Rusu, R., Holzbach, A., Blodow, N., & Beetz, M. (2009b). Fast geometric point labeling using conditional random fields. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (pp. 7–12). St. Louis, MO, USA: IEEE.

Sadeghi, M., Lee, T. K., McLean, D., Lui, H., & Atkins, M. S. (2013). Detection and analysis of irregular streaks in dermoscopic images of skin lesions. *IEEE Transactions on Medical Imaging*, *32*(5), 849–861.

Saha, S., Tahtali, M., Lambert, A., & Pickering, M. R. (2014). 3D x-ray reconstruction using lightfield imaging. In *SPIE Optical Engineering Applications*, (pp. 92090T–92090T). San Diego, CA, USA: International Society for Optics and Photonics.

Sahoo, P., Wilkins, C., & Yeager, J. (1997). Threshold selection using renyi's entropy. *Pattern Recognition*, *30*(1), 71–84.

Sahoo, P. K., Soltani, S., & Wong, A. K. (1988). A survey of thresholding techniques. *Computer Vision, Graphics, and Image Processing*, *41*(2), 233–260.

Saphier, J. (1921). Die dermatoskopie. *Archiv für dermatologie und Syphilis*, *128*(1), 1–19.

Sarkar, S., Paul, S., Burman, R., Das, S., & Chaudhuri, S. S. (2014). A fuzzy entropy based multi-level image thresholding using differential evolution. In *International Conference on Swarm, Evolutionary, and Memetic Computing*, (pp. 386–395). Bhubaneswar, India: Springer.

Satheesha, T., Satyanarayana, D., Prasad, M. G., & Dhruve, K. (2017). Melanoma is skin deep: a 3D reconstruction technique for computerized dermoscopic skin lesion classification. *IEEE Journal of Translational Engineering in Health and Medicine*, *5*, 1–17.

Seidenari, S., Pellacani, G., & Grana, C. (2006). Asymmetry in dermoscopic melanocytic lesion images: a computer description based on colour distribution. *Acta dermato-venereologica*, *86*(2), 123–128.

Senan, E. M., & Jadhav, M. E. (2019). Classification of dermoscopy images for early detection of skin cancer – a review. *International Journal of Computer Applications*, *975*, 8887.

Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, (pp. 3179–3189). Montreal, Canada.

Serratosa, F., & Sanfeliu, A. (2006). Signatures versus histograms: Definitions, distances and algorithms. *Pattern Recognition*, *39*(5), 921–934.

Shademan, A., Decker, R. S., Opfermann, J. D., Leonard, S., Krieger, A., & Kim, P. C. (2016). Supervised autonomous robotic soft tissue surgery. *Science Translational Medicine*, *8*(337), 1–9.

Shanbhag, A. G. (1994). Utilization of information measure as a means of image thresholding. *CVGIP: Graphical Models and Image Processing*, *56*(5), 414–419.

Sheha, M. A., Mabrouk, M. S., & Sharawy, A. (2012). Automatic detection of melanoma skin cancer using texture analysis. *International Journal of Computer Applications*, *42*(20), 22–26.

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, *19*, 221–248.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285–1298.

Shoieb, D. A., Youssef, S. M., & Aly, W. M. (2016). Computer-aided model for skin diagnosis using deep learning. *Journal of Image and Graphics*, *4*(2), 122–129.

Sifre, L., & Mallat, S. (2013). Rotation, scaling and deformation invariant scattering for texture discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1233–1240). Portland, OR, USA: IEEE.

Sifre, L., & Mallat, S. (2014). Rigid-motion scattering for texture classification. *arXiv preprint arXiv:1403.1687*, (pp. 1–9).

Silveira, M., Nascimento, J. C., Marques, J. S., Marçal, A. R., Mendonça, T., Yamauchi, S., Maeda, J., & Rozeira, J. (2009). Comparison of segmentation methods for melanoma

diagnosis in dermoscopy images. *IEEE Journal of Selected Topics in Signal Processing*, *3*(1), 35–45.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, (pp. 1–13).

Situ, N., Yuan, X., Chen, J., & Zouridakis, G. (2008). Malignant melanoma detection by bag-of-features classification. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 3110–3113). Vancouver, BC, Canada: IEEE.

Sivic, J., & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, (pp. 1470–1477). Nice, France: IEEE.

Skvara, H., Teban, L., Fiebiger, M., Binder, M., & Kittler, H. (2005). Limitations of dermoscopy in the recognition of melanoma. *Archives of Dermatology*, *141*(2), 155–160.

Smith, L., & MacNeil, S. (2011). State of the art in non-invasive imaging of cutaneous melanoma. *Skin Research and Technology*, *17*(3), 257–269.

Smith, L., Smith, M., Farooq, A., Sun, J., Ding, Y., & Warr, R. (2011). Machine vision 3D skin texture analysis for detection of melanoma. *Sensor Review*, *31*(2), 111–119.

Society, A. C. (2020). Melanoma skin cancer statistics. https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html. Accessed: 2020-02-10.

Soille, P. (2013). *Morphological image analysis: principles and applications*. Springer Science & Business Media.

Sousa, R. T., & de Moraes, L. V. (2017). Araguaia medical vision lab at ISIC 2017 skin lesion classification challenge. *arXiv preprint arXiv:1703.00856*, (pp. 1–2).

Soyer, H. P., Argenziano, G., Zalaudek, I., Corona, R., Sera, F., Talamini, R., Barbato, F., Baroni, A., Cicale, L., Di Stefani, A., et al. (2004). Three-point checklist of dermoscopy. *Dermatology*, *208*(1), 27–31.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, (pp. 1–14).

Steder, B., Rusu, R., Konolige, K., & Burgard, W. (2011). Point feature extraction on 3D range scans taking into account object boundaries. In *IEEE International Conference on Robotics and Automation*, (pp. 2601–2608). Shanghai, China: IEEE.

Stoecker, W. V., & Moss, R. H. (1992). Editorial: digital imaging in dermatology. *Computerized Medical Imaging and Graphics*, *16*(3), 145–150.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *PMLR International Conference on Machine Learning*, (pp. 1139–1147). Atlanta, GA, USA.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1–9). Boston, MA, USA: IEEE.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 2818–2826). Las Vegas, NV, USA.

Tan, M., & Le, Q. (2019). EfficientNet: rethinking model scaling for convolutional neural networks. In *PMLR International Conference on Machine Learning*, (pp. 6105–6114). Long Beach, CA, USA.

Tan, T. Y., Zhang, L., Lim, C. P., Fielding, B., Yu, Y., & Anderson, E. (2019). Evolving ensemble models for image segmentation using enhanced particle swarm optimization. *IEEE Access*, *7*, 34004–34019.

Tang, P., Liang, Q., Yan, X., Xiang, S., & Zhang, D. (2020). GP-CNN-DTEL: Global-part CNN model with data-transformed ensemble learning for skin lesion classification. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2870–2882.

Tombari, F., Salti, S., & Stefano, L. (2010a). Unique shape context for 3D data description. In *ACM Workshop on 3D Object Retrieval*, (pp. 57–62). New York, NY, USA.

Tombari, F., Salti, S., & Stefano, L. (2010b). Unique signatures of histograms for local surface description. In *European Conference on Computer Vision*, (pp. 356–369). Heraklion, Crete, Greece.

Tombari, F., Salti, S., & Stefano, L. (2011). A combined texture-shape descriptor for enhanced 3D feature matching. In *IEEE Int. Conf. on Image Process.*, (pp. 809–812). Brussels, Belgium: IEEE.

Tripp, J. M., Kopf, A. W., Marghoob, A. A., & Bart, R. S. (2002). Management of dysplastic nevi: a survey of fellows of the american academy of dermatology. *Journal of the American Academy of Dermatology*, *46*(5), 674–682.

Trussell, H. J. (1979). Comments on picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(5), 311.

Tsai, W.-H. (1985). Moment-preserving thresolding: A new approach. *Computer Vision, Graphics, and Image Processing*, *29*(3), 377–393.

Umbaugh, S. E., Moss, R. H., Stoecker, W. V., & Hance, G. A. (1993). Automatic color segmentation algorithms-with application to skin tumor feature identification. *IEEE Engineering in Medicine and Biology Magazine*, *12*(3), 75–82.

Vestergaard, M., Macaskill, P., Holt, P., & Menzies, S. (2008). Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *British Journal of Dermatology*, *159*(3), 669–676.

Vestergaard, M. E., & Menzies, S. W. (2008). Automated diagnostic instruments for cutaneous melanoma. In *Seminars in Cutaneous Medicine and Surgery*, vol. 27, (pp. 32–36). Frontline Medical Communications.

Walden, A., & Cristan, A. (1998). The phase–corrected undecimated discrete wavelet packet transform and its application to interpreting the timing of events. *Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *454*(1976), 2243–2266.

Waldspurger, I. (2015). *Wavelet transform modulus: phase retrieval and scattering*. Ph.D. thesis, Ecole doctorale 386: Sciences Mathématiques de Paris Centre.

Waldspurger, I. (2017). Exponential decay of scattering coefficients. In *International Conference on Sampling Theory and Applications*, (pp. 143–146). Tallin, Estonia.

Wan, S., Prusinkiewicz, P., & Wong, S. (1990). Variance-based color image quantization for frame buffer display. *Color Research & Application*, *15*(1), 52–58.

Wang, X., Tang, F., Chen, H., Luo, L., Tang, Z., Ran, A.-R., Cheung, C. Y., & Heng, P. A. (2020). UD-MIL: uncertainty-driven deep multiple instance learning for October image classification. *IEEE Journal of Biomedical and Health Informatics*, *24*(12), 3431–3442.

Wanner, S. (2014). *Orientation Analysis in 4D Light Fields*. Ph.D. thesis, IWR, Fakultät für Physik und Astronomie, Univ. Heidelberg.
URL http://www.ub.uni-heidelberg.de/archiv/16439

Wick, M. M., Sober, A. J., Fitzpatrick, T. B., Mihm, M. C., Kopf, A. W., Clark, W. H., & Blois, M. S. (1980). Clinical characteristics of early cutaneous melanoma. *Cancer*, *45*(10), 2684–2686.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2005). Practical machine learning tools and techniques. *Morgan Kaufmann*, (p. 578).

Wohlkinger, W., & Vincze, M. (2011). Ensemble of shape functions for 3D object classification. In *IEEE International Conference on Robotics and Biomimetics*, (pp. 2987–2992). Karon Beach, Phuket, Thailand: IEEE.

Wu, X. (1991). Efficient statistical computations for optimal color quantization. In *Graphics Gems II*, (pp. 126–133). Elsevier.

Wu, Y., & He, K. (2018). Group normalization. In *European Conference on Computer Vision*, (pp. 3–19). Munich, Germany.

Xie, Y., Zhang, J., Xia, Y., & Shen, C. (2019). Semi-and weakly supervised directional bootstrapping model for automated skin lesion segmentation. *arXiv preprint arXiv:1903.03313*, (pp. 1–8).

Xie, Y., Zhang, J., Xia, Y., & Shen, C. (2020). A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE Transactions on Medical Imaging*, *39*(7), 2482–2493.

Xu, L., Jackowski, M., Goshtasby, A., Roseman, D., Bines, S., Yu, C., Dhawan, A., & Huntley, A. (1999). Segmentation of skin cancer images. *Image and Vision Computing*, *17*(1), 65–74.

Yang, J., Sun, X., Liang, J., & Rosin, P. L. (2018). Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1258–1266). Salt Lake City, UT, USA: IEEE.

Yang, W., Wang, K., & Zuo, W. (2012). Neighborhood component feature selection for high-dimensional data. *Journal of Computers*, *7*(1), 161–168.

Yao, P., Shen, S., Xu, M., Liu, P., Zhang, F., Xing, J., Shao, P., Kaffenberger, B., & Xu, R. X. (2021). Single model deep learning on imbalanced small datasets for skin lesion classification. *arXiv preprint arXiv:1403.1687*, (pp. 1–10).

Yen, J.-C., Chang, F.-J., & Chang, S. (1995). A new criterion for automatic multilevel thresholding. *IEEE Transactions on Image Processing*, *4*(3), 370–378.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, (pp. 1–12).

Yu, L., Chen, H., Dou, Q., Qin, J., & Heng, P.-A. (2017). Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging*, *36*(4), 994–1004.

Yu, Z., Jiang, X., Zhou, F., Qin, J., Ni, D., Chen, S., Lei, B., & Wang, T. (2018). Melanoma recognition in dermoscopy images via aggregated deep convolutional features. *IEEE Transactions on Biomedical Engineering*, *66*(4), 1006–1016.

Yuan, Y., Chao, M., & Lo, Y.-C. (2017). Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Transactions on Medical Imaging*, *36*(9), 1876–1886.

Yuan, Y., & Lo, Y. (2019). Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 519–526.

Zalaudek, I., Argenziano, G., Soyer, H., Corona, R., Sera, F., Blum, A., Braun, R., Cabo, H., Ferrara, G., Kopf, A., et al. (2006). Three-point checklist of dermoscopy: an open internet study. *British Journal of Dermatology*, *154*(3), 431–437.

Zhao, Q., & Zhang, L. (2005). ECG feature extraction and classification using wavelet transform and support vector machines. In *International Conference on Neural Networks and Brain*, vol. 2, (pp. 1089–1092). Beijing, China: IEEE.

Zhong, Y. (2009). Intrinsic shape signatures: A shape descriptor for 3D object recognition. In *IEEE International Conference on Computer Vision. Workshops*, (pp. 689–696). Kyoto, Japan: IEEE.

Zhou, C., & Nayar, S. K. (2011). Computational cameras: Convergence of optics and processing. *IEEE Transactions on Image Processing*, *20*(12), 3322–3340.

Zhou, H., Chen, M., Zou, L., Gass, R., Ferris, L., Drogowski, L., & Rehg, J. M. (2008). Spatially constrained segmentation of dermoscopy images. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, (pp. 800–803). Paris, France: IEEE.

Zhou, Y., Smith, M., Smith, L., Farooq, A., & Warr, R. (2011). Enhanced 3D curvature pattern and melanoma diagnosis. *Computerized Medical Imaging and Graphics*, *35*(2), 155–165.

Zhou, Y., Smith, M., Smith, L., & Warr, R. (2010). Using 3D differential forms to characterize a pigmented lesion in vivo. *Skin Research and Technology*, *16*(1), 77–84.