D 2022

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

DEEP AESTHETIC ASSESSMENT OF BREAST CANCER SURGERY OUTCOMES

WILSON JOSÉ DOS SANTOS SILVA
DOCTORAL THESIS PRESENTED TO
FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO
IN ELECTRICAL AND COMPUTER ENGINEERING

U.PORTO

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

INESCTEC

Deep Aesthetic Assessment of Breast Cancer Surgery Outcomes

Wilson José dos Santos Silva

Doctoral Programme in Electrical and Computer Engineering

Supervisor: Jaime dos Santos Cardoso, PhD

Co-Supervisor: Maria João de Viseu Botelho Cardoso Aires de Campos, MD, PhD

December 21, 2022

Deep Aesthetic Assessment of Breast Cancer Surgery Outcomes

Wilson José dos Santos Silva

Doctoral Programme in Electrical and Computer Engineering

Approved in public examination by the Jury:

President:	Professor Luís Miguel Pinho de Almeida
Referee:	Professor José Rouco Maseda
Referee:	Professor Luis Filipe Barbosa de Almeida Alexandre
Referee:	Professor Jaime dos Santos Cardoso
Referee:	Professor João Manuel Ribeiro da Silva Tavares
Referee:	Professor João Manuel Patrício Pedrosa

Definitive version validated by the Supervisor:

Professor Jaime S. Cardoso

December 21, 2022

Resumo

Nos últimos anos, os tratamentos para o cancro de mama têm evoluído e melhorado consideravelmente, resultando num aumento substancial nas taxas de sobrevivência, com aproximadamente 80% das pacientes sobrevivendo pelo menos 10 anos. Considerando o impacto significativo que os tratamentos de cancro de mama podem ter na imagem corporal da paciente, afetando a sua autoestima e, por conseguinte, os seus relacionamentos sexuais e íntimos, é fundamental garantir que as mulheres recebem o tratamento mais adequado, optimizando tanto a sobrevivência como a estética resultante. Atualmente, a avaliação estética envolve, pelo menos parcialmente, uma avaliação subjetiva do resultado estético. Contudo, a utilização de métodos subjetivos coloca problemas ao nível da reprodutibilidade e imparcialidade da avaliação. Assim, é de extrema relevância desenvolver uma ferramenta objetiva para a avaliação estética dos resultados da cirurgia de cancro de mama, contribuindo para a definição de um *gold standard*. Tendo em conta que a fonte de informação do problema são fotografias/imagens e as bases de dados que existem são de pequenas dimensões, acrescentando à natureza subjetiva e sensível do problema, o desenvolvimento de uma ferramenta objetiva requer elevados esforços de investigação, nomeadamente, no que se refere a estratégias de regularização e interpretabilidade. Neste trabalho, fazemos contribuições fundamentais na área da interpretabilidade e contribuições aplicadas referentes à avaliação estética de tratamentos de cancro de mama. No entanto, ambos os tipos de contribuição científica tiveram como principal inspiração a avaliação estética dos tratamentos do cancro de mama.

A primeira parte desta tese foca-se nas contribuições fundamentais em interpretabilidade (*Explainable AI*). Estas contribuições foram motivadas a partir de discussões com médicos especialistas na área de cancro de mama e doenças torácicas (cirurgiões e radiologistas). O objetivo das mesmas é ter impacto tanto na avaliação estética dos tratamentos de cancro de mama como em outros contextos médicos (principalmente em radiologia). Várias contribuições podem ser aqui enumeradas, incluindo: a geração de explicações diversas e complementares (tanto com *Deep Neural Networks* como com *Ensemble Models*), o desenvolvimento de sistemas de *retrieval* mais adequados e interpretáveis para ajudar no diagnóstico, o desenvolvimento de modelos de *Deep Learning* que preservem a privacidade dos pacientes em explicações baseadas em casos, permitindo o seu uso generalizado, e finalmente a proposta de uma *framework* de avaliação quantitativa de explicações.

Em relação à segunda parte desta tese, esta é essencialmente dedicada às contribuições aplicadas à avaliação estética dos tratamentos de cancro da mama, incluindo também uma contribuição fundamental na avaliação de classificadores em contextos ordinais e altamente desbalanceados. Dado que a avaliação estética é um problema ordinal com desbalanceamento das classes (a maioria dos casos pertence às classes intermédias), é importante usar métricas adequadas para medir corretamente o desempenho dos classificadores desenvolvidos. A primeira contribuição destacada nesta parte é, portanto, a proposição de uma nova métrica desenvolvida com essas especificidades em mente: a natureza ordinal e desbalanceada do problema. As outras contribuições estão diretamente relacionadas à aplicação, consistindo no desenvolvimento de novos e melhores métodos de

detecção de *keypoints* para auxiliar na avaliação estética e, por fim, um novo modelo para realizar automaticamente a avaliação estética dos tratamentos do cancro de mama. Para além da classificação, i.e., avaliação estética, este modelo proposto também pode ser usado para encontrar os casos anteriores mais semelhantes do ponto de vista semântico, abrindo caminho para um novo sistema que ajudaria a gerir as expectativas das mulheres quanto aos possíveis resultados estéticos do tratamento.

Abstract

Treatments for breast cancer have continued to evolve and improve in recent years, resulting in a substantial increase in survival rates, with approximately an 80% survival at 10 years. Given the impact that breast cancer treatments can have on a patient's body image, consequently affecting her self-confidence and sexual and intimate relationships, it is paramount to ensure that women receive the treatment that optimizes not only survival but also aesthetic outcomes. Currently, the aesthetic assessment involves, at least partially, a subjective evaluation of the aesthetic result. However, the use of subjective methods poses problems at the level of reproducibility and impartiality of the evaluation. Thus, it is extremely relevant to develop an objective tool for the aesthetic assessment of breast cancer surgery outcomes, contributing to the definition of a gold standard. Being an image-based problem (through photographic evaluation) where only small databases exist, and given its subjective and sensitive nature, the development of such an objective tool requires research efforts in terms of regularization strategies and explainability. In this work, we propose fundamental contributions in the area of explainable artificial intelligence and applied contributions regarding the aesthetic evaluation of breast cancer treatments. Both fundamental and applied contributions were attained, having as the main motivation the aesthetic evaluation of breast cancer treatments.

The first part of this thesis covers fundamental contributions to Explainable Artificial Intelligence. These contributions came from discussions with medical experts in the field of breast cancer and thoracic diseases (surgeons, and radiologists), aiming to produce an impact both in the aesthetic evaluation of breast cancer treatments and in other medical application scenarios (mainly radiology). Several contributions can be highlighted, including: producing diverse and complementary explanations (with Deep Neural Networks and Ensemble Models), developing more correct and interpretable medical retrieval systems, developing privacy-preserving machine learning models to anonymize case-based explanations and allow their use when dealing with sensitive data, and also proposing a framework to quantitatively evaluate the quality of the explanations produced.

The second part of this thesis is mostly devoted to applied contributions to the aesthetic evaluation of breast cancer treatments, also including a more fundamental contribution to the assessment of ordinal classifiers. Given that the aesthetic assessment is an ordinal problem with imbalance (most cases fall into the two middle classes), it is important to have proper metrics to correctly measure the performance of the existing classifiers. The first contribution highlighted in this part is therefore the proposition of a novel metric developed with these specificities in mind: the ordinal and imbalance nature of the problem. The other two contributions are directly related to our application, consisting of the development of new and better keypoint detection methods to aid the aesthetic evaluation, and, finally, a new model to automatically perform the aesthetic assessment of breast cancer treatments. Besides classification, this model can also be used to find the most semantically similar past cases, paving the way to a new system that would help manage women's expectations regarding the possible treatment aesthetic outcomes.

Acknowledgements

I would like to start to thank those who guided me during this journey, providing supervision and immense inspiration: Prof. Jaime Cardoso and Prof. Maria João Cardoso. Prof. Maria João motivated me to work on this research topic since the beginning by demonstrating the clinical and social significance of the aesthetic evaluation of breast cancer treatments. Moreover, all discussions with her were extremely interesting and enlightening. Prof. Jaime was always a role model on how to be a machine learning scientist, demonstrating extreme scientific and moral rigor. The best supervisor one can aim to have. In the same line, I am also very grateful to Prof. Mauricio Reyes and Dr. Alexander Pöllinger, who supervised me during my internship in Switzerland, motivating me even more to try to have an impact in Medical Image Computing and Explainable AI.

Besides my supervisors, this work would not be possible without the support of my two affiliated institutions (Faculty of Engineering of the University of Porto, and INESC TEC) and the funding provided by the Portuguese Foundation for Science and Technology (through PhD grant number SFRH/BD/139468/2018).

My thanks also go to my great colleagues and friends from the VCMI group, with whom I learnt immensely, either through more formal collaborations (enumerated in this document), or just during coffee breaks and snack discussions. From those, I would like to address a special thanks to Kelwin, who was a *de facto* supervisor at the beginning of my PhD, and who introduced me to the amazing world of Explainable AI. You are able to learn more about deep learning from a ten minute discussion with him than by reading a deep learning manual. Furthermore, I want to give a special thanks to Diogo Pernes (our mathematical genius), João Pinto (the most prolific scientific writer), Eduardo (with whom you are able to have the most profound discussions), Tiago (my first Master's student, who is always productive, even when sleep deprived), Helena (the most dedicated student you will ever find), and Isabel (our NLP expert, even though she disagrees). A huge thanks also to Sara and Ana, very close friends with whom I organized the best summer school ever (VISUM2022). Besides all the aforementioned Portuguese (and Venezuelan) colleagues, two Swiss guys also had a special contribute to my scientific development during this path, and also deserve to be mentioned here. Alain, and Fabian, *merci vielmal*!

However, many more persons contributed indirectly to this work by providing emotional support and motivation during this commonly difficult period of doing a PhD, even made harder by a pandemic and successive lockdowns. To my friends from Bairrada and all the ones I made during my stays in Porto, Lisbon, Karlsruhe, and Bern, a huge thanks! A special ackowledgment to Patrícia who was always an enthusiastic supporter of my endeavours, and who was at my side during most of this challenge. Even though we chose different paths in life, you will always have a special place in my heart. Finally, I want to thank my family for the inexhaustible support, and would like to dedicate this work to them: my late maternal grandparents (Américo and Urânia), my paternal grandparents (Mário e Cidália), my parents (Idalécio e Célia) and my sister Melissa! Without you none of this would be possible.

"if you are going to try,
go all the way.
there is no other feeling like
that.
you will be alone with the
gods
and the nights will flame with
fire."

Charles Bukowski

Contents

I Prologue 1

1 Introduction 3
- 1.1 Context and Motivation . . . 3
- 1.2 Objectives and Document Structure . . . 3
- 1.3 Dissemination . . . 5
- 1.4 Collaborations . . . 8

II Explainable Artificial Intelligence 11

2 Towards Complementary Explanations for Machine Learning Models 13
- 2.1 Context and Motivation . . . 13
- 2.2 Related Work . . . 14
- 2.3 Quantitative Evaluation of Explanations . . . 16
- 2.4 Methodology . . . 17
  - 2.4.1 Complementary Explanations using Deep Neural Networks . . . 17
  - 2.4.2 Complementary Explanations using an Ensemble Model . . . 19
- 2.5 Results and Discussion . . . 22
- 2.6 Summary and Conclusions . . . 26

3 Interpretability-guided Medical Image Retrieval 29
- 3.1 Context and Motivation . . . 29
- 3.2 Related Work . . . 31
- 3.3 Methodology . . . 33
- 3.4 Results and Discussion . . . 36
- 3.5 Summary and Conclusions . . . 42

4 Privacy-preserving Case-based Explanations for Machine Learning Models 45
- 4.1 Context and Motivation . . . 45
- 4.2 Background . . . 47
- 4.3 Related Work . . . 49
- 4.4 Methodology . . . 50
- 4.5 Results and Discussion . . . 55
  - 4.5.1 Ablation Study . . . 60
- 4.6 Summary and Conclusions . . . 61

III Aesthetic Evaluation of Breast Cancer Treatments 63

5 Aesthetic Evaluation of Breast Cancer Treatment Outcomes 65
5.1 Context . . . 65
5.2 Related Work . . . 67
5.3 From handcrafted to deep-learning-based methodologies . . . 69
5.4 Discussion and Future Work . . . 70

6 Keypoint Detection for the Aesthetic Assessment of Breast Cancer Treatments 73
6.1 Context and Motivation . . . 73
6.2 Related Work . . . 74
6.3 Methodology . . . 75
6.3.1 Deep Keypoint Detection Algorithm . . . 75
6.3.2 Hybrid Keypoint Detection Algorithm . . . 78
6.3.3 Keypoint detection through deep segmentation . . . 79
6.4 Results and Discussion . . . 80
6.5 Summary and Conclusions . . . 83

7 Deep Aesthetic Assessment and Retrieval of Breast Cancer Treatment Outcomes 85
7.1 Context and Motivation . . . 85
7.2 Methodology . . . 86
7.3 Experimental Setup . . . 88
7.3.1 Data . . . 88
7.3.2 Evaluation . . . 88
7.4 Results and Discussion . . . 89
7.5 Summary and Conclusions . . . 91

8 Assessment of Ordinal Classifiers and its application to Aesthetic Evaluation 93
8.1 Context and Motivation . . . 93
8.2 Methodology . . . 97
8.2.1 Conceptual formulation . . . 97
8.2.2 Application from estimates in a confusion matrix . . . 99
8.2.3 Handling unobserved classes . . . 99
8.3 Experimental Setup . . . 101
8.4 Ordinal Assessment of Aesthetic Evaluation Classifiers . . . 109
8.5 Summary and Conclusions . . . 110

IV Conclusions 111

9 Conclusion 113
9.1 Fundamental Contributions . . . 114
9.2 Applied Contributions . . . 115
9.3 Final Remarks and Future Work . . . 115
9.4 Funding . . . 116

References 117

List of Figures

2.1 Illustration of explanation quality for decision rules and KNN (where the black dot is the new observation and the blue dot is the nearest-neighbor). 17
2.2 Proposed DNN architecture. 18
2.3 Feature impact analysis. 18
2.4 Ensemble Model. 19
2.5 Example of a Decision Tree applied to a particular feature X. 21
2.6 Scorecard implemented with a neural network. Input neurons represent the bins, and / the linear activation function. 22
2.7 Ensemble Model Proposed. 23
2.8 Visualization of the explanations. In the BCCT dataset we are considering the binary classification problem: {Poor, Fair} *vs.* {Good, Excellent}. Regarding the PH^2, the classification problem comes down to {Common, Atypical} *vs.* {Melanoma}. 25
2.9 Explanations obtained in the FICO Explainable ML Challenge. 26

3.1 Pleural Effusion test image and the least and most similar images of the catalogue according to our board-certified radiologist (in terms of disease and disease severity). The overall most similar image would be (b). However, such matching is not of radiological interest. 31
3.2 Overview of the proposed Interpretability-guided approach. Blocks in light gray (■) mean neural networks are being trained (i.e., weights are being updated), whereas blocks in dark gray (■) represent trained neural networks (i.e., weights are fixed). In the saliency maps, brighter colors mean higher relevance. Blue circles indicate ranking positions. CNN represents the deep model used as baseline. IG-CNN represents the CNN model architecture being trained with saliency maps. 34
3.3 Box-and-whisker plots regarding the nDCG results for the pleural effusion Top-10 (a) and Top-3 (b) retrieved images. SSIM is the statistically-based baseline, CNN is the CNN-based baseline, IG is the proposed interpretability-guided approach, ATT is the attention method, R2 is the ranking provided by the second board-certified radiologist, and R3 is the ranking provided by the third board-certified radiologist. 37
3.4 Example of test case and the Top-4 retrieved images given by each of the radiologists (R1 = ground-truth, R2, and R3) and each of the machine learning methods. In this split, both the CNN, IG, and ATT obtained nDCGs (Top-10) > 0.9. The green box means pleural effusion case and the red box means no pleural effusion (according to the dataset label). 38

3.5 Box-and-whisker plots regarding the nDCG results for (potential) pneumonia Top-10 (a) and Top-3 (b) retrieved images. SSIM is the statistically-based baseline, CNN is the CNN-based baseline, IG is the proposed interpretability-guided approach, ATT is the attention method, R4 is the ranking provided by the fourth board-certified radiologist, and R5 is the ranking provided by the fifth board-certified radiologist. 40
3.6 Example of test case and the Top-4 retrieved images given by each of the radiologists (R1 = ground-truth, R4, and R5) and each of the machine learning methods. In this split, both the CNN, IG, and ATT obtained nDCGs (Top-10) > 0.8. The green box means potential pneumonia case, red box means no potential pneumonia, and orange box means a disagreement between R1 and label, with R1 considering the case as potential pneumonia. 41
3.7 Box-and-whisker plots regarding the nDCG results for Top-10 retrieved images. CNN is the CNN baseline model, CNN(IG) is the CNN model having as inputs the Deep Taylor saliency maps, and IG is the proposed interpretability-based approach (i.e., it was trained (fine-tuned) with saliency maps, and has as inputs also saliency maps). 42

4.1 Diagram exemplifying the explanatory retrieval process. Consumers illustrated in red represent individuals who do not possess authorized access to the raw data (identity information) in the database. Consumers illustrated in green can access the raw data. 46
4.2 Overview of the PPRL-VGAN model's architecture. 51
4.3 Overview of our privacy-preserving model's architecture. We included in green a summary of the changes that occurred to the original PPRL-VGAN model, to apply it to the domain of case-based interpretability. 53
4.4 Architecture of the privacy-preserving model with generation of counterfactual explanations. 55
4.5 Results of the privacy-preserving model with multi-class identity recognition. The first and second columns represent the original images and their privatized versions, respectively. The third and fourth columns contain Deep Taylor saliency maps obtained from the original and privatized images, respectively. 56
4.6 Comparison between results from the network when explanatory evidence is considered (first row) and not (second row). 57
4.7 Results of the privacy-preserving model with Siamese identity recognition. The first and second columns represent the original images and their privatized versions, respectively. The third and fourth columns contain Deep Taylor saliency maps obtained from the original and privatized images, respectively. 58
4.8 Results of the privacy-preserving model with Siamese recognition, not considering overall privacy. The first and second columns represent the original images and their privatized versions, respectively. The third and fourth columns contain Deep Taylor saliency maps obtained from the original and privatized images, respectively. 59
4.9 Results of counterfactual generation using privacy-preserving model with Siamese identity recognition. The first and second rows represent privatized factual and counterfactual explanations, respectively. The final row contains a map with the differences between the factual and counterfactual explanations. 59
4.10 Results obtained by replacing the generator with other architectures. The first image is the original one, and the following images are privatized images using the original VAE, the ResNet VAE, and UNET, respectively. 60

5.1 The balance between optimal oncological and aesthetic outcomes. 66
5.2 The variability of patient self-evaluation. 67
5.3 Screenshots of the first two computer programs for the aesthetic assessment of LR. 68

6.1 Image gradient, from [35]. 75
6.2 Shortest paths from the bottom to the middle. 75
6.3 Shortest paths from the middle to the bottom. 75
6.4 Strong paths between middle and bottom. 75
6.5 Selected shortest paths. 75
6.6 Strong paths between top and bottom. 75
6.7 The endpoints are the highest points of the shortest path. 75
6.8 Automatic breast contour detection with the shortest path algorithm. 75
6.9 Nipple candidates detection [26]. 76
6.10 Harris corner detection. [26]. 76
6.11 Proposed iterative DNN architecture for keypoint detection. FCN stands for fully-convolutional network. Conv Backbone stands for convolutional backbone. . . . 77
6.12 Example of Ground Truth 77
6.13 Hybrid Keypoint Detection Algorithm. 78
6.14 A Novel Segmentation-Based Keypoint Detection Algorithm. 79
6.15 VIENNA dataset examples 80
6.16 Network outputs for a test instance example 81
6.17 Chronological scheme (from top to bottom) of the proposed Segmentation-Based Keypoint Detection Algorithm. Each column represents a single image. The first row is the ground-truth image, the second row is the ground-truth mask, the third row is the U-Net++ predicted mask, the fourth row is the set of all the detected contour keypoints and the fifth row is the set of breast keypoints after the post-processing step. 82
6.18 Example predictions when using the deep keypoint detection algorithm. Prediction is in blue and ground-truth is in red. 83
6.19 Example predictions when using the hybrid keypoint detection algorithm. Prediction is in blue and ground-truth is in red. 83
6.20 Example predictions when using the segmentation-based keypoint detection algorithm. Prediction is in blue and ground-truth is in red. 83

7.1 Overview of the proposed approach. Blocks in light gray mean deep neural networks are being trained (i.e., weights are being updated), whereas blocks in dark gray represent trained deep neural networks (i.e., weights are fixed). The block in white means that no weights are learnt. The “L2 distance” is computed based on the features from the previous to the last layer of the network, i.e., exactly before the classification decision. 87
7.2 Example of images used in this work. 88
7.3 Example of query and retrieved images for an Excellent aesthetic outcome. Binary label means class belongs to set {Excellent, Good}. Original label is the ordinal label previous to binarization (Excellent, Good, Fair, or Poor). LRP saliency map is also shown for the test image. 90
7.4 Binary label means class belongs to set {Excellent, Good}. Original label is the ordinal label previous to binarization (Excellent, Good, Fair, or Poor). LRP saliency map is also shown for the test image. 90

7.5 Example of query and retrieved images for a Fair aesthetic outcome. Binary label means class belongs to set {Fair, Poor}. Original label is the ordinal label previous to binarization (Excellent, Good, Fair, or Poor). LRP saliency map is also shown for the test image. 91
7.6 Example of query and retrieved images for a Poor aesthetic outcome. Binary label means class belongs to set {Fair, Poor}. Original label is the ordinal label previous to binarization (Excellent, Good, Fair, or Poor). LRP saliency map is also shown for the test image. 92

8.1 Illustration of A_{UOC} and the values of UOC^1_β obtained from an example confusion matrix. 100
8.2 Values of UOC, for several β values, and A_{UOC}, obtained with a confusion matrix with a single sample, according to its distance to the diagonal $|r-c|$. 101
8.3 Evolution of metric values with the total number of classes K, when applied to tridiagonal matrices. 102
8.4 Confusion matrices, regular and normalized, for the classifiers kNN, SVM, and Random Forest, used on the wine quality dataset. 105
8.5 Confusion matrices, regular and normalized, for the classifiers kNN, SVM, and Random Forest, used on the ESL dataset. 107
8.6 Confusion matrices for the SVM classifier. In (a) by training as a typical multi-class problem. In (b) using the Frank and Hall approach. 109

List of Tables

2.1 Example of a scorecard. 20
2.2 Example of a scorecard with discretization based on Decision Tree of Fig. 2.5. The scores are only illustrative. 21
2.3 Quality of the predictions in terms of area under the ROC and Precision-Recall curves. Quality of the explanations in terms of correctness (Corr), completeness (Compl), and compactness (Compt). 24

4.1 Results of the privacy-preserving model with multi-class identity recognition. We expect low values in the privacy-related metrics and high values in glaucoma recognition accuracy. The best results for each metric are highlighted in bold. . . . 57
4.2 Results of the privacy-preserving model with Siamese identity recognition. The best results for each metric are highlighted in bold. 60
4.3 Results of replacing generator with ResNet VAE and UNET. For convenience, the first lines repeat the results shown in Table 4.1. 60
4.4 Comparison between the results of our privacy-preserving models and results of state-of-the-art models obtained from [114] in regards to privacy and preservation of explanatory evidence. 61

5.1 Examples of the dimensionless asymmetry measures [23]. (X_1, Y_1) and (X_2, Y_2) are the coordinates of both nipples (using the sternal notch as the centre of coordinates); NI_1 and NI_2 are the nipple to infra-mammary fold distances. 69

6.1 Average error distance for endpoints, breast contours and nipples measured in pixels 81
6.2 Average error distance for endpoints, breast contours and nipples, measured in pixels and average execution time of the models' inferences (on the test set of each cross-validation fold, which has approximately 43 to 45 images). Best results are highlighted in bold. **Note:** STD stands for standard deviation and Max stands for maximum error. 82

7.1 Results for the test set. Linear and RBF represent the SVM kernel, while 4 and 7 represent the number of symmetry features given as input to the SVM. 89

8.1 Results for the simulated confusion matrices, with $\beta_1 = 0.25$ and $\beta_2 = 0.75$ . . . 103
8.2 Results for the classifiers in real datasets, with $\beta_1 = 0.25$ and $\beta_2 = 0.75$ 106
8.3 Results for the test set. Multi-class and Frank Hall represent the two approaches for training the SVM model. Linear kernel was used since it generated the best results. 110

Abbreviations

AI	Artificial Intelligence
AMAE	Average Mean Absolute Error
AP	Anteroposterior (chest view)
AUC	Area Under the Curve
BC	Breast Cancer
BCCT	Breast Cancer Conservative Treatment
BCE	Breast Compliance Evaluation
BCSS	Breast Cancer Specific Survival
BRA	Breast Retraction Assessment
CAM	Channel attention module
CBIR	Content-based Image Retrieval
CNN	Convolutional Neural Networks
DCG	Discounted Cumulative Gain
DB	Database
DM	Digital Mammography
DNN	Deep Neural Networks
DREAM	Dialogue on Reverse Engineering Assessment and Methods
DT	Decision Tree
ED	Euclidean Distance
EORTC	European Organisation for Research and Treatment of Cancer
EUSOMA	European Society of Mastology
FP	False Positives
FTS	Features
GAN	Generative Adversarial Network
Grad-CAM	Gradient-Weighted Class Activation Mapping
GP	Gradient Penalty
IG	Interpretability-guided
IMF	Infra-Mammary Fold
KNN	K-Nearest Neighbors
KL	Kullback-Leibler
LBC	Lower Breast Contour
LR	Locoregional (treatment)
LRP	Layer-wise Relevance Propagation
MAE	Mean Absolute Error
MER	Misclassification Error Rate
MLDAM	Multi-level dual-attention mechanism
MLP	Multi-layer Perceptron
MMAE	Maximum Mean Absolute Error

MSE	Mean Squared Error
nDCG	normalized Discounted Cumulative Gain
OC	Ordinal Classification (Index)
OS	Overall Survival
PA	Posteroanterior (chest view)
PACS	Picture Archiving and Communication System
PAM	Position attention module
PR	Precision-Recall
PROMs	Patient Reported Outcome Measures
QoL	Quality of Life
RBF	Radial Basis Function
RF	Random Forest
ROC AUC	Area Under the Receiver Operating Characteristic curve
SSIM	Structural Similarity Index
STD	Standard Deviation
SVM	Support Vector Machine
TP	True Positives
UNR	Upward Nipple Retraction
UOC	Uniform Ordinal Classification (Index)
VAE	Variational Autoenconder
WGAN	Wasserstein Generative Adversarial Network

Part I

Prologue

Chapter 1

Introduction

1.1 Context and Motivation

Breast cancer is an increasingly treatable disease, with 10-year survival now exceeding 80% [54]. This high survival rate led to an increased interest in the quality of life after treatment, in many aspects, one of those being the aesthetic outcome.

Furthermore, with the development of new surgical possibilities and radiation therapies, it is even more critical to have the means to compare aesthetic results. To refine the current and new techniques, identify factors with a significant impact on the aesthetic outcome [45] and compare breast units, it is vital to have an objective and reproducible method to assess the aesthetic results. Nonetheless, to this date, there is no gold standard recognized method to evaluate a treatment's aesthetic outcome.

Thereby, it is extremely relevant to develop new methods in order to propose a system with desirable properties and performance to become a gold standard objective method. This document tackles some of the challenges related to the development of this method, addressing fundamental and applied research topics.

1.2 Objectives and Document Structure

The primary purpose of this thesis is to study and develop machine learning and computer vision techniques to make possible the development of an objective tool for the aesthetic assessment of breast cancer locoregional therapy outcomes, contributing to the definition of a gold standard. This gold standard would help manage the expectation of patients, improve treatment techniques and fairly compare different breast units. For this to be a reality, fundamental and applied research has to be conducted. Fundamental research will focus mainly on the assessment of ordinal classifiers and explainable artificial intelligence. Since the aesthetic evaluation is by nature an ordinal classification problem with highly imbalanced classes, the search for suitable performance metrics to correctly evaluate classifiers in this particular and demanding scenario is extremely important. Moreover, for the clinicians to trust the model, explanations for the aesthetic evaluation have to be

provided. Besides increasing trust, explainability may also help refine current techniques, manage patient expectations of the aesthetic result and educate new surgeons and nurses on what is relevant for the aesthetic evaluation of breast cancer treatments. Even though these addressed fundamental topics are motivated by the aesthetic evaluation of breast cancer treatments, they may have an impact in many other clinical and non-clinical (e.g., biometrics) applications, as explainability is a common requirement for all high-stake decisions. Applied research will be related to the detection of keypoints/fiducial points clinicians know are relevant to assessing the aesthetic outcome and to the development of an end-to-end deep learning model capable of performing the aesthetic assessment and of finding the most semantically similar past-cases (important when managing patient expectations). Thus, the following constitute the main scientific and technical objectives of the project:

- Proposal of new metrics for performance assessment of ordinal classification with imbalanced classes;
- Investigate deep learning techniques for fiducial point detection (from which aesthetic assessment can be performed);
- Development of interpretable models and explanatory techniques;
- Improvement of the current objective aesthetic classification of Breast Cancer Conservative Treatment (BCCT) using deep learning methodologies;
- Promotion of the new techniques and testing of the developed tool in national and international healthcare institutions, including the Champalimaud Foundation Breast Unit;

This document will present the main contributions related to the aforementioned goals, being divided into two parts. The first part of this thesis is devoted to Explainable Artificial Intelligence, and is formed by the following chapters:

- Chapter 2: in this chapter, we present a framework to quantitatively assess the quality of explanations, and two classification models that are able to generate clinical decisions and from which we are capable of extracting diverse and complementary explanations in the form of rule-based and case-based explanations. We also propose an approach to extract a single explanation from an ensemble.
- Chapter 3: radiologists often turn to public or internal image databases to search for similar disease-matching images to aid the diagnosis when in doubt about a suspected condition. Thus, we propose a new medical image retrieval approach driven by interpretability saliency maps to automatically find the most similar disease-matching images.
- Chapter 4: we analyse and propose privacy-preserving models to anonymize case-based explanations, simultaneously optimizing a multi-task objective, privacy, realism and explanatory evidence.

In the second part of this thesis,

- Chapter 5: we do an analysis of the motivation, state-of-the-art, current and future opportunities in the field of the aesthetic assessment of breast cancer treatments.
- Chapter 6: current objective assessment methods for the aesthetic evaluation of breast cancer treatments depend on high-level features extracted based on certain keypoints annotated in the images. In this chapter, we propose several algorithms to improve the automatic annotation of these keypoints.
- Chapter 7: we propose the first deep learning model for evaluating the aesthetic outcome of breast cancer treatments. This model not only outperforms the previous state-of-the-art for the binary version of the aesthetic assessment problem, but can also be used to find the most semantically similar past cases, which can help guide patient expectations.
- Chapter 8: in this chapter, we propose a new metric specifically developed to assess the performance of classifiers deployed in a scenario where the classes have a natural order (ordinal classification) and where the distribution of the classes is highly imbalanced.

1.3 Dissemination

The contributions of the doctoral research on explainable artificial intelligence, and aesthetic evaluation of breast cancer treatments described in this thesis have been disseminated as part of fifteen scientific publications. Moreover, three additional publications resultant from the input of explainable artificial intelligence to the field of biometrics (which will not be detailed in this thesis) were also produced. These are (clustered by type and in reverse chronological order):

- Articles in journals:

    7. W. Silva, T. Gonçalves, K. Härmä, E. Schröder, V. C. Obmann, M. C. Barroso, A. Poellinger, M. Reyes, and J. S. Cardoso, "Computer-aided Diagnosis through Medical Image Retrieval in Radiology," *Scientific Reports*, 2022. doi:10.1038/s41598-022-25027-2
    6. P. Neto, T. Gonçalves, J. R. Pinto, W. Silva, A. F. Sequeira, A. Ross, J. S. Cardoso, "Explainable Biometrics in the Age of Deep Learning," *ACM Computing Surveys*, 2022 [submitted, awaiting decision]. Publication outside the scope of the thesis but related to my Explainable AI contributions.
    5. H. Montenegro, W. Silva, A. Gaudio, M. Fredrikson, A. Smailagic, and J. S. Cardoso, "Privacy-preserving Case-based Explanations: Enabling visual interpretability by protecting privacy," *IEEE Access*, 10, 28333-28347, 2022. doi:10.1109/ACCESS.2022.3157589

4. H. Montenegro, W. Silva, and J. S. Cardoso, "Privacy-Preserving Generative Adversarial Network for Case-Based Explainability in Medical Image Analysis," *IEEE Access*, 9, 148037-148047, 2021. doi:10.1109/ACCESS.2021.3124844
3. A. F. Sequeira, T. Gonçalves, W. Silva, J. R. Pinto, and J. S. Cardoso, "An exploratory study of interpretability for face presentation attack detection," *IET Biometrics*, 10(4), 441-455, 2021. doi:10.1049/bme2.12045. Publication outside the scope of the thesis but related to my Explainable AI contributions.
2. T. Gonçalves, W. Silva, M. J. Cardoso, and J. S. Cardoso, "A novel approach to keypoint detection for the aesthetic evaluation of breast cancer surgery outcomes," *Health and Technology*, 10 (4), 891-903, 2020. doi:10.1007/s12553-020-00423-8
1. J. S. Cardoso, W. Silva, and M. J. Cardoso, "Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment," *The Breast*, 49, 123-130, 2020. doi:10.1016/j.breast.2019.11.006

- Articles in international conference proceedings:

11. H. Montenegro, W. Silva, and J. S. Cardoso, "Disentangled Representation Learning for Privacy-Preserving Case-Based Explanations," in *MAD 2022: Workshop on Medical Applications with Disentanglements at MICCAI 2022*, Set. 2022. Even though the publication is in line with the research topics presented in this document, my contributions to the work were not enough for considering its inclusion in the thesis.
10. W. Silva, M. Carvalho, C. Mavioso, M. J. Cardoso, and J. S. Cardoso, "Deep Aesthetic Assessment and Retrieval of Breast Cancer Treatment Outcomes," in *10th Iberian Conference on Pattern Recognition and Image Analysis (IbPria 2022)*, May 2022 [best student paper award]. doi.org/10.1007/978-3-031-04881-4_9
9. D. Mata, W. Silva, and J. S. Cardoso, "Increased Robustness in Chest X-ray Classification through Clinical Report-driven regularization," in *10th Iberian Conference on Pattern Recognition and Image Analysis (IbPria 2022)*, May 2022. Even though the publication is in line with the research topics presented in this document, it is still preliminary work, lacking consistency to be included in the thesis.
8. H. Montenegro, W. Silva, and J. S. Cardoso, "Towards Privacy-preserving Explanations in Medical Image Analysis," in *1st Workshop on Interpretable Machine Learning in Healthcare at ICML (IMLH 2021)*, Jul. 2021. Available online at: https://arxiv.org/abs/2107.09652
7. W. Silva, A. Poellinger, J. S. Cardoso, and M. Reyes, "Interpretability-Guided Content-Based Medical Image Retrieval," in *23rd International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2020)*, Oct. 2020. doi:10.1007/978-3-030-59710-8_30 [MICCAI Student Travel Award]

6. A. F. Sequeira, W. Silva, J. R. Pinto, T. Gonçalves, and J. S. Cardoso, "Interpretable Biometrics: Should We Rethink How Presentation Attack Detection is Evaluated?," in *8th International Workshop on Biometrics and Forensics (IWBF 2020)*, Apr. 2020. doi:10.1109/IWBF49977.2020.9107949. Publication outside the scope of the thesis but related to my Explainable AI contributions.
5. T. Gonçalves, W. Silva, and J. S. Cardoso, "Deep Aesthetic Assessment of Breast Cancer Surgery Outcomes," in *XV Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON 2019)*, Sep. 2019. doi:10.1007/978-3-030-31635-8_236
4. W. Silva, K. Fernandes, and J. S. Cardoso, "How to produce complementary explanations using an Ensemble Model," in *International Joint Conference on Neural Networks (IJCNN 2019)*, Jul. 2019. doi:10.1109/IJCNN.2019.8852409
3. W. Silva, E. Castro, M. J. Cardoso, F. Fitzal and J. S. Cardoso, "Deep Keypoint Detection for the Aesthetic Evaluation of Breast Cancer Surgery Outcomes," in *16th IEEE International Symposium on Biomedical Imaging (ISBI 2019)*, Apr. 2019. doi:10.1109/ISBI.2019.8759331
2. W. Silva, K. Fernandes, M. J. Cardoso, and J. S. Cardoso, "Towards complementary explanations using deep neural networks," in *iMIMIC 2018: Workshop on Interpretability of Machine Intelligence in Medical Image Computing at MICCAI 2018*, Set. 2018. doi:10.1007/978-3-030-02628-8_15
1. W. Silva, J. R. Pinto, and J. S. Cardoso, "A Uniform Performance Index for Ordinal Classification with Imbalanced Classes," in *International Joint Conference on Neural Networks (IJCNN 2018)*, Jul. 2018. doi:10.1109/IJCNN.2018.8489327

The research conducted during these doctoral studies has also been partially presented to the scientific community at the 2021 CMU Portugal Doctoral Symposium and on seven abstracts presented at four editions of the *Portuguese Conference on Pattern Recognition (RECPAD)*, which are enumerated below (in reverse chronological order):

7. H. Montenegro, W. Silva and J. S. Cardoso, "Anonymising Case-based Explanations for Medical Image Analysis," in *27th Portuguese Conference on Pattern Recognition (RECPAD 2021)*, Nov. 2021.
6. W. Silva, and J. S. Cardoso, "Complementary and case-based explanations for clinical decision support," in *27th Portuguese Conference on Pattern Recognition (RECPAD 2021*, Nov. 2021. [Best Poster Award]
5. T. Gonçalves, W. Silva, and Jaime S. Cardoso, "A Deep Image Segmentation Approach to Breast Keypoint Detection," in *26th Portuguese Conference on Pattern Recognition (RECPAD 2020)*, Oct. 2020.

4. W. Silva, J. R. Pinto, T. Gonçalves, A. F. Sequeira, and Jaime S. Cardoso, "Explainable Artificial Intelligence for Face Presentation Attack Detection," in *26th Portuguese Conference on Pattern Recognition (RECPAD 2020)*, Oct. 2020.
3. T. Gonçalves, W. Silva, and J. S. Cardoso, "Towards a Deep Keypoint Detection Algorithm for the Aesthetic Assessment of Breast Cancer Surgery Outcomes," in *25th Portuguese Conference on Pattern Recognition (RECPAD 2019)*, Oct. 2019.
2. W. Silva, K. Fernandes, M. J. Cardoso, and J. S. Cardoso, "Interpretable Ensemble Model for the Aesthetic Evaluation of Breast Cancer Treatments," in *25th Portuguese Conference on Pattern Recognition (RECPAD 2019)*, Oct. 2019.
1. W. Silva, K. Fernandes, M. J. Cardoso, and J. S. Cardoso, "Understanding Deep Neural Networks decisions in Medical Imaging," in *24th Portuguese Conference on Pattern Recognition (RECPAD 2018)*, Oct. 2018.

1.4 Collaborations

The doctoral work presented in this thesis included close collaborations with researchers and clinicians from other institutions, namely:

- Prof. Maria João Cardoso and Dr. Carlos Mavioso from the Champalimaud Foundation Breast Unit, who provided image data regarding the aesthetic evaluation of breast cancer treatments, keypoint and classification annotations, and feedback for all the results, articles, and applications we developed. Moreover, Prof. Maria João Cardoso co-supervised this thesis, providing inspiration and clinical supervision.
- Prof. Florian Fitzal from the Medical University of Vienna, who provided the VIENNA dataset used in the keypoint detection articles.
- Prof. Mauricio Reyes from the University of Bern, who supervised me during my internship at the University of Bern, in Bern, Switzerland. Prof. Mauricio Reyes provided scientific supervision related to explainable artificial intelligence and medical image retrieval.
- Prof. Alexander Poellinger from Inselspital (Bern University Hospital), who (informally) supervised me during my stay in Bern. Prof. Alexander Poellinger provided clinical supervision related to explainable artificial intelligence and medical image retrieval, also extensively supporting my work with radiology knowledge.
- Dr. Kirsi Härmä, Dr. Erich Schröder, Dr. Verena Carola Obmann, and Dr. Maria Cecilia Barroso from Inselspital (Bern University Hospital), who provided retrieval annotations for Chest X-ray data (medical image retrieval article).

The author of this thesis also extensively collaborated with colleagues from the VCMI group at INESC TEC, namely, Kelwin Fernandes, João Ribeiro Pinto, Tiago Gonçalves, and Helena

Gonçalves, with whom several of the publications presented in this thesis were done. Furthermore, he also collaborated with other colleagues in the development of additional publications (Pedro Neto, and Ana F. Sequeira) and organization of scientific events (Ana Rebelo, Sara Oliveira, Isabel Rio-Torto, Alain Jungo, and Fabian Balsiger). During the time of the PhD, he was involved in the organization of the following scientific events:

- Steering Committee Member for VISUM Summer School 2022
- Organization Team Member for MICCAI Hackathon (2021, 2022)
- Program Committee Member for xAI4Biometrics at WACV (2021, 2022)
- Program Chair for iMIMIC at MICCAI (2020, 2021, 2022)
- Project Committee Member for VISUM Summer School (2018, 2019, 2020, 2021)

The research work developed throughout the PhD was related to, directly motivated by, or motivated the proposal of several scientific projects, namely:

- NanoStima-RL5: Advanced Methodologies for Computer-Aided Detection and Diagnosis (active member) [National Project]
- CLARE: Computer-aided cervical cancer screening (active member) [FCT Research Project]
- TAMI: Transparent Artificial Medical Intelligence (participated in the writing of the proposal and was an active member) [CMU Portugal Research Project]
- CINDERELLA: Clinical Validation of an AI-based approach to improve the shared decision-making process and outcomes in Breast Cancer Patients proposed for Locoregional treatment (project partially motivated by this thesis' research) [European Project]
- CAGING: Causality-driven Generative Models for Privacy-preserving Case-based Explanations (played an important role in the writing of the proposal, Co-PI of the project) [FCT Exploratory Research Project]

Additionally, the author has collaborated, as co-supervisor, on the following master dissertations related to his doctoral studies (in reverse chronological order):

4. Maria Carvalho, "Towards Biometrically-morphed Medical Case-based Explanations", Master in Bioengineering, Universidade do Porto, 2022 - as co-supervisor, alongside Professor Maria J. Cardoso (co-supervisor) and Professor Jaime S. Cardoso (supervisor);
3. Diogo Mata, "Biomedical multimodal explanations – increasing diversity and complementarity in Explainable Artificial Intelligence", Master in Biomedical Engineering, Universidade do Porto, 2022 - as co-supervisor, alongside Professor Jaime S. Cardoso (supervisor); CTM Best Master's Thesis Award (2022).

2. Helena Montenegro, “A privacy-preserving framework for case-based interpretability in machine learning”, Master in Informatics and Computing Engineering, Universidade do Porto, 2021 - as co-supervisor, alongside Professor Jaime S. Cardoso (supervisor); CTM Best Master’s Thesis Award (2021), APRP Best Master’s Thesis Award (2021), 3rd place in the Fraunhofer Portugal Challenge (2021).

1. Tiago Gonçalves, “Deep Aesthetic Assessment of Breast Cancer Surgery Outcomes”, Master in Bioengineering, Universidade do Porto, 2019 - as co-supervisor, alongside Professor Jaime S. Cardoso (supervisor).

Part II

Explainable Artificial Intelligence

Chapter 2

Towards Complementary Explanations for Machine Learning Models

Foreword on Author Contributions

The results of this work have been disseminated in the form of two papers in international conference:

- [150] W. Silva, K. Fernandes, M. J. Cardoso, and J. S. Cardoso, "Towards Complementary Explanations Using Deep Neural Networks," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, Springer, Cham, 2018. 133-140.
  https://doi.org/10.1007/978-3-030-02628-8_15
- [153] W. Silva, K. Fernandes, and J. S. Cardoso, "How to produce complementary explanations using an Ensemble Model," in *International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2019.
  https://doi.org/10.1109/IJCNN.2019.8852409

2.1 Context and Motivation

In the most recent years, many machine learning models are replacing or helping humans in decision-making scenarios. The recent success of deep neural networks (DNN) in the most diverse applications led to widespread use of this technique, being of potential utility in solving the aesthetic evaluation problem. Nonetheless, their high accuracy is not accompanied by high interpretability. On the contrary, they remain mostly as black-box models. In this way and despite the success of DNN, in areas such as medicine and finance, which have legal and safety constraints, their use is somehow restricted. Therefore, and in order to take advantage of the DNN potential, it is critical to develop robust strategies to explain the behavior of the model. Moreover, these strategies have to be aligned with our knowledge about human learning. It is known that human beings have different ways of thinking and learning [127]. There are people for whom a visual explanation is more easily apprehended and, on the contrary, there are people who prefer a verbal explanation. In order to satisfy all consumers of explanations, an interpretable model should be able to provide different styles of explanations and with different levels of granularity. Furthermore, it should present as many explanations as the decision maker needs to be confident about his/her decisions. It is also important to mention that some observations require more complex

explanations than others, which reinforces the idea of different depth in the explanations. As in the case of the aesthetic evaluation of breast cancer treatments, this diversity and complementarity could be of extreme importance as the consumers of the explanations have diverse backgrounds (surgeons, nurses, and patients), where a verbal explanation alone can be of value to a surgeon but not to the patient, while a visual explanation alone does not possess enough detail to differentiate similar cases, and analyse small improvements.

2.2 Related Work

The literature is rich in the use of different machine learning methods and in different approaches to obtain interpretability, or in a broader sense, to produce explanations for the decisions that models make. Nevertheless, one can think of interpretability as a three-stage process, closely related to the development cycle of a data science solution. In accordance with this idea, Kim and Doshi-Velez [55] grouped the different strategies in pre-, in-, and post-model.

Pre-Model

The first stage, the pre-model, focuses on trying to understand the data itself and takes place before the construction of the machine learning model. Here, visualization and exploratory data analysis play a significant role.

Visualization is a quite common technique in the business intelligence community, and it basically consists of visualizing the behaviour/distribution of the data according to the different features available [173]. Exploratory data analysis, a concept firstly introduced by Tukey [171], also focuses on the understanding of the data, but it is more general than visualization, also including quantitative techniques apart from the graphical ones. These two strategies can be fundamental for building trust in the subsequent machine learning model. Furthermore, an understanding of the behaviour of the data in accordance with the features available might help in the construction of new features (hand-crafted), which is especially important when one is working with simpler models. By incorporating interpretability-based criteria directly into the model design, one can also extract and select better and more interpretable features [90]. When dealing with highly complex data distributions, the use of prototypes, i.e., examples that characterize well the data or a particular class of elements, is beneficial. However, in the context of having a distribution, in which some data points are not well characterized by a given prototype, prototypes are not enough and the MMD-critic framework proposed by Kim *et al.* [91], which also selects the criticisms, is fundamental. In summary, pre-model interpretability is not enough when isolated, but it plays an important role when integrated into the general context of interpretability. Only by first understanding the data distribution that we are dealing with, can we rely on the subsequent decisions and explanations that a given machine learning model can provide.

In-Model

In-model approaches, on the other hand, focus on integrating interpretability inside the model. In order to build an interpretable model or to make a machine learning model more interpretable, there are several different strategies available.

One of the first strategies that comes to mind when thinking about making an interpretable model, because it is closely related to, and inspired by, human nature, is to build a model based on rules, which are able to characterize the different classes in question. A widely known example of such a model is a decision tree [18]. However, other models like decision lists [134], and rule sets [178] are also valid options. Also related are the per-feature based models, like the generalized additive models [82] in general and their discretized version, scorecards [63; 138], which are extensively used in industry, and in particular in finance. Nonetheless, the interpretability of these models is limited by the semantic meaning of the original features, the complexity of the rules, and by the size of the model, or depth in the case of decision trees.

Another strategy, closely related to the way human beings think, is to build models based in cases instead of in rules. Exploring again the idea of prototypes, decisions and explanations can be obtained through cluster divisions, with each cluster being characterized by a prototype [89]. However, the quality of an explanation generated using this approach is limited from the beginning by the representativeness of the prototypes. Moreover, the formation of the clusters typically depends on the distance metric considered, which may not be the most appropriate for the context under study. More recent works focus instead on prototypical parts to make the decision, increasing interpretability and robustness [10; 40].

Now, instead of thinking in natural/obvious ways to base models, we can search for procedures to increase the interpretability of a certain model, which is not interpretable in its usual implementation. For example, complex and non-interpretable models, like deep neural networks, can be made more interpretable using some regularization techniques that simplify the model. One of such techniques, which aims to achieve sparsity [97], is the well known $L1$ regularization [74]. This regularization technique consists of the sum of the absolute values of the model individual parameters, w_i, and is mathematically described in Equation (2.1).

$$\Omega(\theta) = ||\omega||_1 = \sum_i |\omega_i| \tag{2.1}$$

In the context of linear regression, the addition of this term results in the famous LASSO model [166]. With this addition, a subset of weights becomes zero, which means that some features are discarded, a property that obviously increases the interpretability of the method.

Another property with great interest in the interpretability domain is a monotonic relationship to some or, ideally, all of the input features [78]. In the context of neural networks, it can be obtained by constraining the weights to be positive, or negative, depending on the increasing or decreasing nature of the function to be learned [148]. However, it is important to note that the introduction of these regularization techniques ($L1$ and monotonicity), which help improve

interpretability, comes at the cost of model complexity and therefore can have a significant negative impact on model performance.

More recent efforts focused on guiding the neural networks into learning relevant concepts (e.g., high-level clinical concepts) [42; 92; 155], and on integrating causal knowledge into the network [67; 177].

Post-Model

The last stage, post-model, has the aim of understanding the model decisions but already after a model has been built. In here, a possible strategy can be the perturbation of the input provided to the model and the analysis of the consequent impact on the model output. This strategy is known as sensitivity analysis. When working with images, a possible perturbation is the occlusion of some parts of the image [61]. Related with this approach are the gradient-based methods. Instead of occluding regions of the image, these methods use gradient information to identify the areas of the image that mostly contribute to the final decision (e.g., class that the images belong to) [8; 142; 147; 158; 159]. Nonetheless, there is no guarantee that changes made to the input represent a realistic scenario, and spatial explanations typically do not have a rich semantic meaning. One can also try to understand what the model is doing by looking at the feature representations that the model has learnt in solving a particular task. In the case of neural networks, this is done by observing the latent or hidden units of the network. However, an understanding of what is being represented in the learned semantic space is usually not easy. Therefore, some techniques were developed, which mainly consist of inverting the representations back to the input pixel space [56; 163; 185] or connecting the representations to semantic concepts [11; 92; 150; 153]. Other widely used techniques to produce saliency maps consist of optimization approaches [133], or decomposition [7; 111].

A different strategy is to mimic a more complex model with a simpler one. Being simpler, a model is consequently more interpretable. In the context of neural networks, an example of this would be to try to imitate the behaviour of a very deep model with a more shallow one [5]. Two issues with this approach are the fact that a simpler model may not exist, and that it is difficult to verify if the mimic model is really representative of what the more complex model is doing.

2.3 Quantitative Evaluation of Explanations

Despite the vast amount of effort that has been invested around interpretable models, the concept itself is still vaguely defined and lacks of a unified formal framework to assess it.

The efficacy of an explanation depends on its ability to convince the target audience. Thus, it is surrounded by external intangible factors such as the background of the audience and its willingness to accept the explanation as a truth. While it is hard to fully assess the quality of an explanation, some proxy functions can be used to summarize the quality of a prediction under certain assumptions. Defining an explanation as a simple model that can be applied to a local context of the data, we suggest that a good explanation should maximize the following properties:

- **Completeness** An explanation should be complete, i.e., it should be general enough for it to be applied to more than one observation. Defining the covered set as the set of training cases covered by the explanation, completeness is the ratio between the sizes of the covered set and the training set (presented in percentage terms), e.g., the blue rows in Fig. 2.1 where the decision rule precondition holds and the observations within the same distance of the neighbor explanation (Fig. 2.1).

- **Correctness** An explanation should be correct, in the sense that if we consider the explanation itself as a model, it should be able to correctly identify the class to which the current observation belongs. Quantitatively, it means the percentage of the covered set that is correctly classified, or in other words, it is the accuracy of the explanation as a model, e.g., the label agreement between the blue rows and between the points inside the n-sphere.

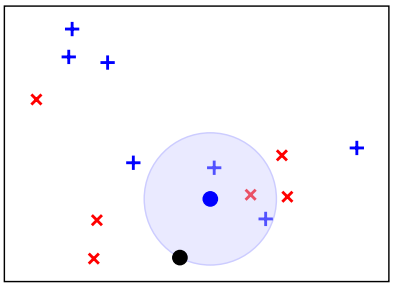
- **Compactness** An explanation should be compact, or in other words, it should be succinct. If an explanation is very long, it is explaining nothing. Quantitatively, compactness can be measured as the size in bytes of the explanation, e.g., the number of conditions in the decision rule and the feature dimensionality of a neighbor-based explanation.

We named this framework "the three C's of interpretability" and its application is presented in Fig. 2.1, where we illustrate the explanation quality assessment of a rule-based explanation and a case-based explanation.

A	B	Y
0	4	1
2	3	1
4	2	0
3	6	1
3	3	1

If $A \geq 1 \wedge B \leq 5 \wedge B \geq 2$
then $y = 1$

Compl: 3/5
Corr: 2/3
Compt: 3



Compl: 4/14
Corr: 3/4
Compt: 2

Figure 2.1: Illustration of explanation quality for decision rules and KNN (where the black dot is the new observation and the blue dot is the nearest-neighbor).

2.4 Methodology

2.4.1 Complementary Explanations using Deep Neural Networks

In addition to their high accuracy in various classification problems, DNN have the ability to jointly integrate different strategies of interpretability, such as, the previously mentioned, case-based, monotonicity and sensitivity analysis. Thus, it is a model that presents itself at the forefront to satisfy the explanation consumers in their search for valuable and diverse explanations.

We will focus on binary classification settings with a known subset of monotonic features. Without loss of generality, we will assume that monotonic features increase with the probability

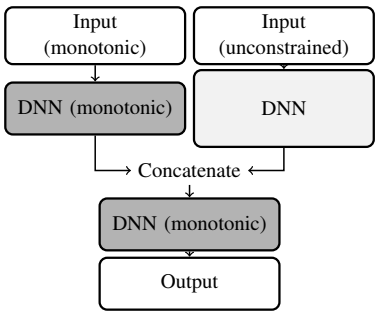


Figure 2.2: Proposed DNN architecture.

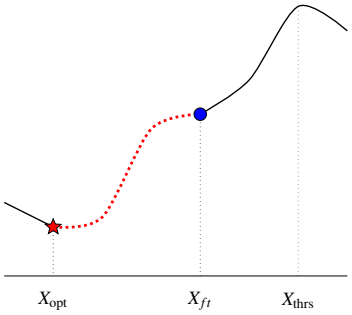


Figure 2.3: Feature impact analysis.

of observing the positive class. The proposed architecture consists of two independent streams of densely connected layers that process the monotonic and non-monotonic features respectively. We impose constraints on the weights of the monotonic stream to be non-negative to facilitate interpretability. Then, both streams are merged and processed by a sequence of densely connected layers with non-negative constraints. Thus, we are promoting that the non-monotonic stream maps its feature space into a latent monotonic space. It is expected that the non-monotonic features will require additional expressiveness to transform a non-monotonic space into a monotonic one. In this sense, we validate topologies where the non-monotonic stream has at least as many, and possibly more, layers than the monotonic stream. Figure 2.2 illustrates the proposed architecture.

Explanation by local contribution (rule-based explanation)

To measure the contribution, C_{ft}, of a feature ft on the prediction y, we can find the assignment X_{opt} that approximates X to an adversarial example (see Eq. (2.2)):

$$(\bar{y} - f(X))^2 \tag{2.2}$$

where $\bar{y} = 1 - y$ is the opponent class, $y \in \{0, 1\}$, and $f(X)$ is the estimated probability. We can use backpropagation with respect to ft to find the value X_{opt} (see Fig. 2.3) that minimizes Eq. (2.2). It is relevant to note that for monotonic features, such value is known a priori. Since some features may have a generalized higher contribution than others, resulting in repetitive explanations, we balanced the contribution on the target variable with the range of the feature domain traversed from the initial value to the local minimum X_{opt}. Namely:

$$C_{ft} = |f(X) - f(X')| \cdot \frac{X_{ft} - X_{opt}}{X_{\max} - X_{\min}} \tag{2.3}$$

where X' is the input vector after assigning X_{opt} to the feature ft. Thus, the contribution can be measured by approximating X to the adversarial space. On the other hand, the inductive rule constructed for ft covers the space between X_{ft} and the value X_{thrs} where the probability of the predicted class is maximum.

Explanation by similar examples (case-based explanation)

DNN are able to learn intermediate semantic representations adapted to the predictive task. Thus, we can use the nearest neighbors in the semantic space as an explanation for the decision. While the latent space is not fully interpretable, we can evaluate which features (and at which degree) impact the distance between two observations using sensitivity analysis. In this sense, two types of explanations can be extracted:

- **Same class ("factual"):** the nearest neighbor from the same class, and the explanation for why they are from the same class.
- **Opposite class ("counterfactual"):** the nearest neighbor from the opposite class in the latent space and what features make them different.

2.4.2 Complementary Explanations using an Ensemble Model

The combination of different models inside a global one allows the use of different interpretability strategies at the same time, which improves the quality of the explanations. Moreover, having different models results in a diversity of types of explanations, and usually also in better classification performance.

Diversity is important because people and application domains vary a lot, and what is suited for a certain person or for a particular application domain may not be suited for others.

Decisions regarding the interpretability strategies to be used, and the way we propose to generate explanations, were based on the previously presented **"three C's of interpretability"** framework.

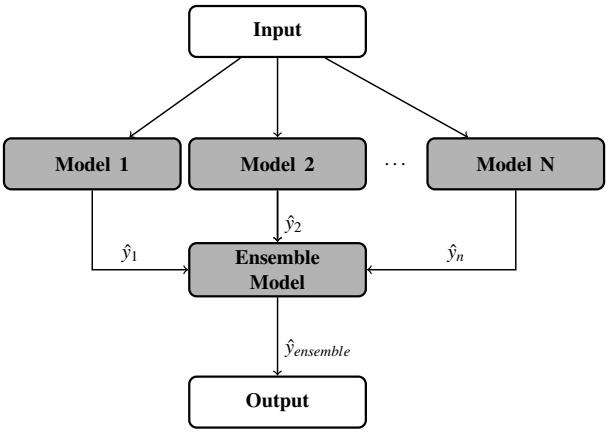


Figure 2.4: Ensemble Model.

Having this framework into consideration, we are able to propose a method to generate explanations for the predictions of ensemble models. For an ensemble model with N models (Fig. 2.4), the prediction made by the ensemble is given by the majority vote within its elements. Thus, we can look to the subset M of models that agree with the global decision and select the explanations that these models provide to a pool of candidate explanations for the ensemble. The final

global explanation of the ensemble for a given test point x is the one from the pool of M candidate explanations ($E_n(x)$) that has the highest correctness ($corr(E_n(x))$ - Eq. (2.4).

$$E_{global}(x) = \underset{E_n(x)}{\text{argmax}}(corr(E_n(x)), n \in \{1,...,M\} \wedge M \leq N \tag{2.4}$$

In this work, we also instantiate an ensemble model to provide complementary explanations, which is constituted by the following models:

Deep Neural Networks

The first part of the ensemble is the network we presented before, which is illustrated in Fig. 2.2. In this ensemble model, we will only consider the case-based explanations provided by the deep neural network, being the rule-based explanations generated by the other additional models (Scorecards, and Random Forest).

Scorecards

A scorecard is an intrinsically interpretable model widely used in financial applications, particularly regarding credit scoring [63]. In Table 2.1, we exemplify how this model works. Considering a new observation, depending on the values of its features, the observation will get a particular number of points per feature, resulting in a total score that will determine the class that it belongs.

Table 2.1: Example of a scorecard.

	Bins	**Points**
Feature X		
	Up to **x1**	10
	x1 to **x2**	25
	x2 to **x3**	38
	x3 and up	43
Feature Y		
	Up to **y1**	50
	y1 to **y2**	65
	y2 and up	70
Total Score		**88**

One of the critical aspects when building a scorecard is defining the way the discretization of the features is made. Typical strategies include performing the discretization using equal-width or equal-frequency algorithms. Equal-width consists on dividing the range of feature values in a pre-defined number of bins with the same width. On the other hand, equal-frequency consists on dividing the range of feature values also in a pre-defined number of bins but with each bin having the same number of training observations. Both algorithms belong to the unsupervised category, i.e., the division of the bins does not take into account the class of the observations. In this work, we perform the binning of the scorecard using a decision tree, and its thresholds

as the cut-off points of the bins. Figure 2.5 provides an illustration of a decision tree applied to a particular feature X. The discretization of feature X is done accordingly to the thresholds previously computed, resulting in the illustrative example presented in Table 2.2.

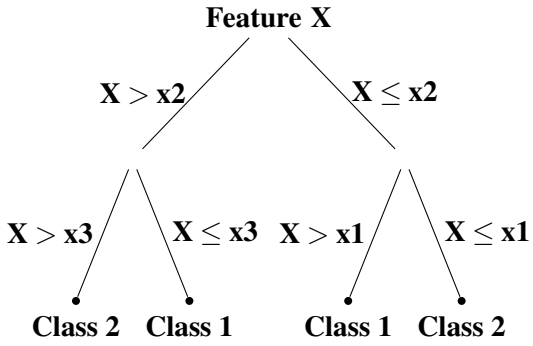


Figure 2.5: Example of a Decision Tree applied to a particular feature X.

Table 2.2: Example of a scorecard with discretization based on Decision Tree of Fig. 2.5. The scores are only illustrative.

	Bins	**Points**
Feature X		
	Up to **x1**	50
	x1 to **x2**	25
	x2 to **x3**	30
	x3 and up	60

Our implementation of the scorecard was based on neural networks, with the weights of the neurons being the weights of the scorecard (Fig. 2.6). Regarding in-model interpretability, we have considered three regularization techniques. The first one was to use differential-coding in the bins [149], which promotes a smooth variation in the points attributed to consecutive bins. The second one was to use $L1$ regularization over the differential-coding of the bins to ensure a sparse number of scores. Finally, the third technique was to impose non-negative constraints on the weights of the neural network/scorecard, which in conjunction with the differential-coding ensures monotonicity. Moreover, we also increased the interpretability of the model after the same has been built (post-model interpretability). For this, we merge neighboring bins that are monotonic on the decision for a given local sample in an attempt to improve the completeness of the explanations.

Random Forest

The last model we considered as part of our ensemble is itself an ensemble, a Random Forest [17]. A Random Forest is an ensemble of decision trees, and it was created because decision trees alone tend to overfit to the training data. When using a multitude of trees and averaging the predictions

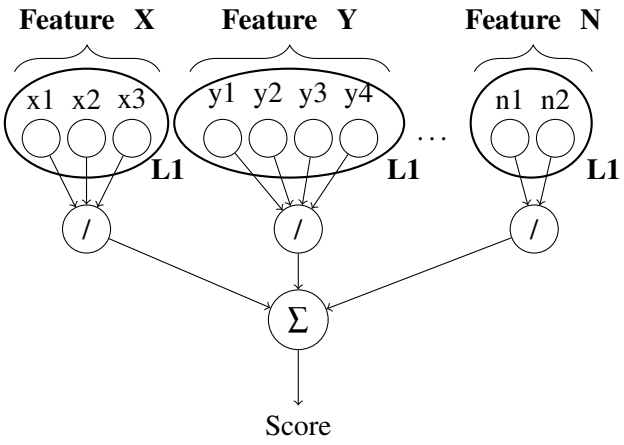


Figure 2.6: Scorecard implemented with a neural network. Input neurons represent the bins, and / the linear activation function.

from all trees of the ensemble, we are able to reduce overfitting, and therefore, to have a more robust model. This robustness comes at the expense of more complexity, and, consequently, less interpretability. Although Random Forests are not considered interpretable, the individual trees within the ensemble are (at least given a small limited depth). Thus, we can find an explanation for the ensemble decision using the approach we proposed, selecting the explanation from the tree with the path from its source to the tree leaf that leads to a more correct explanation.

Considering post-model interpretability, we prune the tree branches that do not lead to further class refinement, and so we are able to produce more complete explanations.

Ensemble Model

The model we instantiate is then constituted by the previously presented (sub-)models: Deep Neural Network, Scorecard, and Random Forest. The prediction made by the ensemble, $\hat{y}_{ensemble}$, is given by the majority vote computed based on the predictions $\hat{y}_{dnn}$, $\hat{y}_{s}$, and $\hat{y}_{rf}$, which are the predictions made by the Deep Neural Network, Scorecard, and Random Forest, respectively (Figure 2.7). The final explanation is chosen accordingly to the proposed method.

2.5 Results and Discussion

We validate the performance of the proposed deep neural network and ensemble model on three datasets, one from the financial domain and the others from the biomedical/medical domain. We will present the results for the deep neural network as baseline and part of the ensemble. For an exclusive analysis of the deep neural network performance, please check Silva *et al.* [150].

Unlike the medical datasets, the original features of the financial dataset are "raw" features, defined without the help of an expert. Thus, it is important to consider pre-model interpretability strategies, with a subsequent step of feature engineering that will define new extended features. Our results are computed using the extended feature version.

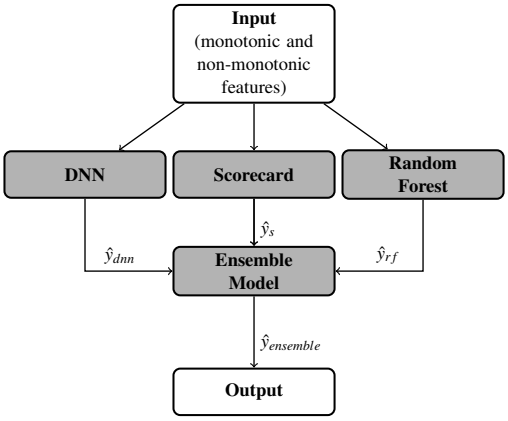


Figure 2.7: Ensemble Model Proposed.

The financial dataset being used is the 2018 FICO Explainable Machine Learning Challenge's Credit dataset [62]. This dataset is an anonymized dataset of Home Equity Line of Credit (HELOC) applications. The problem related to the dataset is a binary classification with the target variable being called of RiskPerformance. RiskPerformance can be:

- **Good:** which means that the applicant made his/her payments without ever being more than 90 days overdue.
- **Bad:** which means that the applicant was 90 days past due or worse at least once over a period of 24 months from when the credit account was open.

Regarding the medical datasets, we considered one relative to dermoscopy image classification [107] and another to aesthetic evaluation of breast cancer treatments [23].

The first medical dataset, dermoscopic image classification [107], has 14 high-level features acquired from 200 patients. Features describe the presence of certain colors on the nevus and abnormal patterns. The goal of the problem related to this dataset is to classify each observation in three different classes: Common, Atypical, and Melanoma. For this work, as we only consider binary classification, we have binarized the problem into two different ones: Common vs. Atypical and Melanoma, and Common and Atypical vs. Melanoma. The second medical dataset, aesthetic evaluation of breast cancer treatments [23], has 23 high-level features acquired from 143 patients. Features describe breast asymmetry in terms of shape, and local and global differences in color. Local differences in color aim to detect scars in the breasts. The aesthetic evaluation of breast cancer treatments consists of an ordinal classification problem composed of four different classes: Poor, Fair, Good, and Excellent. Here, we considered the three binary classification tasks:

- **Excellent** vs. **Good, Fair, and Poor**
- **Excellent, and Good** vs. **Fair, and Poor**
- **Excellent, Good, and Fair** vs. **Poor**

Table 2.3: Quality of the predictions in terms of area under the ROC and Precision-Recall curves. Quality of the explanations in terms of correctness (Corr), completeness (Compl), and compactness (Compt).

PH^2: Dermoscopy Images [107]							
Binarization	**Model**	**Predictions**		**Explanations**			
		ROC	**PR**	**Type**	**Corr**	**Compl**	**Compt**
Common *vs.* Atypical, Melanoma	Random Forest	99.53	99.70	Rule	97.30	10.49	33.32
	Scorecard	99.17	99.60	Rule	82.97	24.23	23.96
	DNN	**99.74**	**99.83**	Similar	97.11	39.00	19.32
				Opponent	74.59	70.61	37.69
	Ensemble	99.64	99.76	Best	92.27	16.69	30.96
Common, Atypical *vs.* Melanoma	RF	96.33	87.41	Rule	95.01	13.59	32.88
	SC	95.86	87.85	Rule	82.94	38.10	24.00
	DNN	96.02	89.30	Similar	91.49	8.15	33.27
				Opponent	84.02	62.12	46.24
	Ensemble	**96.64**	**89.43**	Best	94.76	18.65	35.25
BCCT: Breast Aesthetics [23]							
Binarization	**Model**	**Predictions**		**Explanations**			
		ROC	**PR**	**Type**	**Corr**	**Compl**	**Compt**
Excellent *vs.* Good, Fair, Poor	Random Forest	90.06	75.28	Rule	97.68	7.97	33.14
	Scorecard	92.95	78.25	Rule	96.81	26.19	24.68
	DNN	91.03	73.00	Similar	87.25	1.46	79.79
				Opponent	92.82	67.86	157.81
	Ensemble	**93.72**	**79.58**	Best	94.91	14.98	74.36
Excellent, Good *vs.* Fair, Poor	Random Forest	86.00	82.27	Rule	94.80	4.87	46.32
	Scorecard	86.72	**83.31**	Rule	81.57	22.60	24.87
	DNN	86.78	82.82	Similar	72.52	17.34	80.36
				Opponent	81.16	31.28	138.00
	Ensemble	**87.28**	82.47	Best	86.61	21.43	87.69
Excellent, Good, Fair *vs.* Poor	Random Forest	83.49	97.32	Rule	97.50	13.95	27.12
	Scorecard	85.03	97.35	Rule	98.00	10.60	33.34
	DNN	80.61	96.55	Similar	85.69	95.20	124.94
				Opponent	92.04	46.87	149.68
	Ensemble	**85.61**	**97.72**	Best	96.27	12.67	63.73
FICO Explainable ML Challenge [62]							
Binarization	**Model**	**Predictions**		**Explanations**			
		ROC	**PR**	**Type**	**Corr**	**Compl**	**Compt**
Negative *vs.* Positive	Random Forest	**77.61**	75.46	Rule	84.18	5.77	57.87
	Scorecard	76.35	74.55	Rule	73.57	17.96	30.97
	DNN	76.71	74.88	Similar	87.22	0.38	114.64
				Opponent	49.70	99.01	199.24
	Ensemble	77.41	**75.58**	Best	82.11	30.76	101.74

We compared the performance of the proposed Ensemble Model against its constituents alone: a Deep Neural Network (previously proposed), a Scorecard (which is a highly interpretable model), and a Random Forest. We used 10-fold cross-validation to choose the best hyper-

parameter configuration and to generate explanations. Scorecard bins per feature were limited to 20 for FICO dataset, and 10 for the medical datasets. The depth of trees within the Random Forest was limited to 5, being a good trade-off between predictive performance and model interpretability. We show in Table 2.3 the performance of the four models regarding the quality of the predictions and their respective explanations. To measure the quality of the predictions, we considered the area under the ROC curve, and Precision-Recall (PR). Our proposed Ensemble model leads to higher predictive performance in the majority of the scenarios considered, and when it does not, it is in line with the best performing model. Regarding the quality of the explanations, the evaluation was based on the "three C's of interpretability", considering the correctness, completeness, and compactness of the explanations generated by each model. The Ensemble proposed, despite increasing the diversity of the explanations, maintains very high correctness in the explanations that it generates.

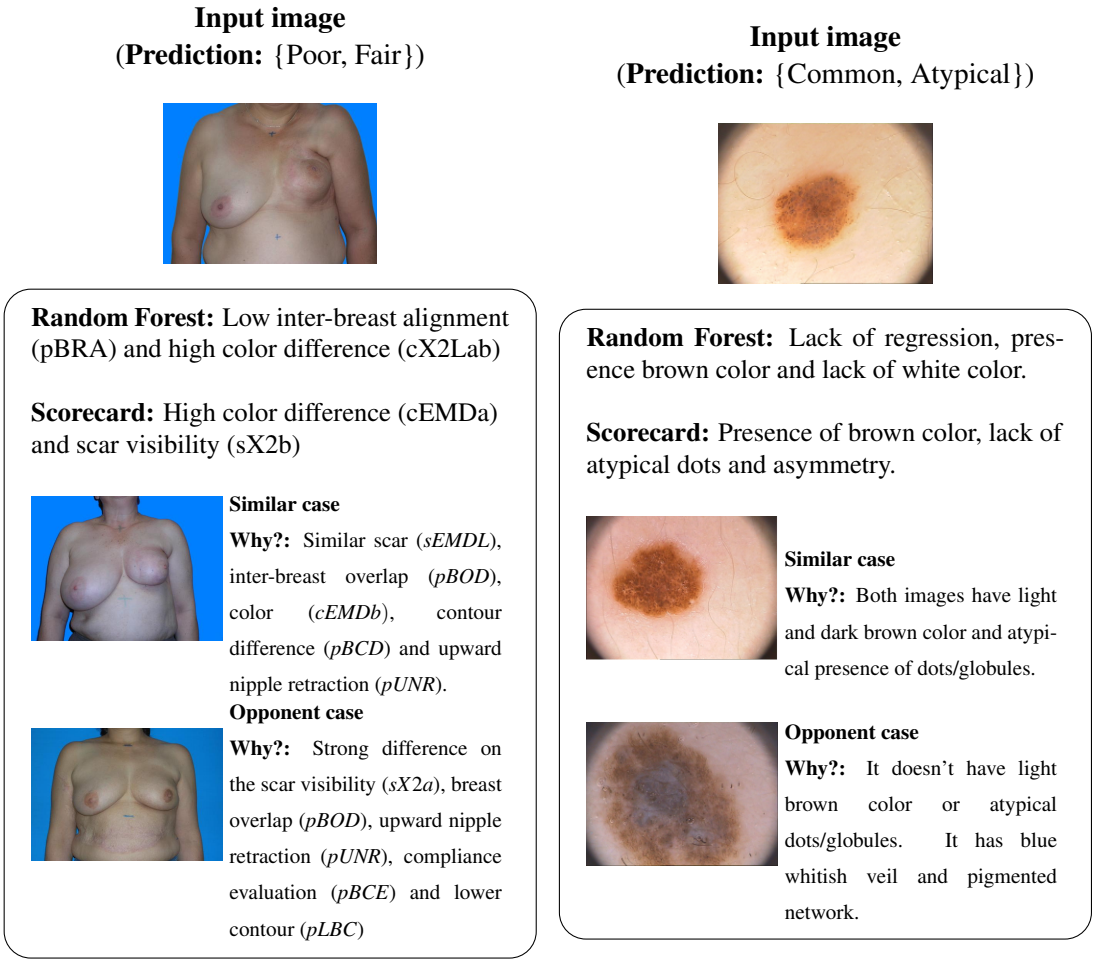


Figure 2.8: Visualization of the explanations. In the BCCT dataset we are considering the binary classification problem: {Poor, Fair} *vs.* {Good, Excellent}. Regarding the PH^2, the classification problem comes down to {Common, Atypical} *vs.* {Melanoma}.

Figures 2.8 and 2.9 illustrate the explanations obtained by the proposed ensemble on the three datasets. As can be seen in the example, different models and explanation strategies tend to support

the decision using different subsets of features. Namely, they offer complementary evidence of the predicted class.

Prediction: Rejected

- **Scorecard:** The consolidated version of risk markers is below 81.50, the maximum of the relative values of the average number of months in file and the percentage of trades never delinquent is above 0.67, the average number of months in file is below 97.50, the sum of the relative values of the average number of months in file and the percentage of trades never delinquent is above 0.57, and the average relative position of the client's features is above 0.50.
- **Random Forest:** Although the sum of the relative values of the percentage of trades never delinquent and the net fraction revolving burden is below 0.89, the following facts support the decision: condition of the number of months since most recent delinquency not met and the average relative position of the client's features is above 0.51.
- **Deep Neural Network:**
  - **Similar:** This client is rejected because No usable/valid trades or inquiries observed in the number of months Since Most Recent inquiries (excluding the last 7 days), the average number of months in file with value 41.0, the number of satisfactory trades with value 2.0, the number of months Since Most Recent inquiries (excluding the last 7 days) with value 0.0, and the number of Months Since Most Recent Delinquency with value 15.0 are similar to client in row 5662. Client 5662 could not payoff.
  - **Opponent:** The minimum of the relative values of the number of months Since Most Recent inquiries (excluding the last 7 days) and the Net Fraction Revolving Burden with value 0.0, the average number of months in file with value 41.0, the percentage of Trades Never Delinquent with value 100.0, the number of Months Since Most Recent Delinquency with value 15.0, and the Net Fraction Revolving Burden with value 0.0 are different to client in row 4340, with values 0.5, 219.0, 86.0, 1.0, and 31.0 respectively.. Client 4340 paid.

Figure 2.9: Explanations obtained in the FICO Explainable ML Challenge.

2.6 Summary and Conclusions

The use of machine learning models in areas like medicine and finance is highly restricted due to interpretability concerns. Both clinicians and patients, and clients and regulators want to understand the decisions provided by the models. Given the diversity of target users, the variability in application domains and in examples within the same application, the existence of a method able to generate correct explanations along with diversity and complementarity is fundamental.

In this chapter, we presented a framework to quantitatively evaluate the quality of an explanation. Based on that same framework, we proposed an approach to select the global explanation of an ensemble. Moreover, we also developed a Deep Neural Network and an Ensemble Model capable of generating complementary explanations, fulfilling the need for correctness and diversity.

The proposed models are evaluated in three datasets, one financial (FICO Explainable Machine Learning Challenge) and two biomedical (Dermoscopic Image Classification, and Aesthetic

Evaluation of Breast Cancer Treatments), with one of these being related to the main topic of this thesis. Regarding the quantitative results, the proposed ensemble leads to higher predictive performance when compared with the Deep Neural Network, Scorecard, and Random Forest alone. Moreover, the explanations that it generates have very high values of correctness, without losing too much completeness, and maintaining reasonable compactness. In its turn, qualitative results show that the goal of obtaining diversity in the explanations generated is fulfilled. Moreover, the explanations are in accordance with the analysis of experts (clinicians in the case of medical applications, and bankers in the case of credit).

Chapter 3

Interpretability-guided Medical Image Retrieval

Foreword on Author Contributions

The results of this work have been disseminated in the form of one paper in an international conference and one submitted article to an international journal:

- [154] W. Silva, A. Poellinger, J. S. Cardoso, and M. Reyes, "Interpretability-Guided Content-Based Medical Image Retrieval," in *23rd International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2020)*, Oct. 2020. doi:10.1007/978-3-030-59710-8_30
- [156] W. Silva, T. Gonçalves, K. Härmä, E. Schröder, V. C. Obmann, M. C. Barroso, A. Poellinger, M. Reyes, and J. S. Cardoso, "Computer-aided Diagnosis through Medical Image Retrieval in Radiology," *Scientific Reports*, 12, 20732 (2022). doi:10.1038/s41598-022-25027-2

3.1 Context and Motivation

In the previous chapter, we presented our first works in interpretability/explainable AI. We also introduced the different approaches available to make models more interpretable and to generate explanations for their decisions. While all approaches have their own relevance, in terms of radiology, case-based explanations arise as the most useful and relevant.

As previously pointed out, case-based explanations, or explanations-by-example, are images retrieved from a database (past cases) similar to the observation being analyzed in terms of relevant task-related features. Case-based explainability, in the form of image retrieval, is commonly used in scenarios such as medical image diagnosis to obtain examples of similar disease-matching images that can be compared to a case under analysis and provide additional insights to explain and support a diagnosis. The retrieval process begins with a user entering an image into the retrieval system. Then, the retrieval system ranks the examples in its database according to a semantic similarity measure and presents to the user the most similar examples.

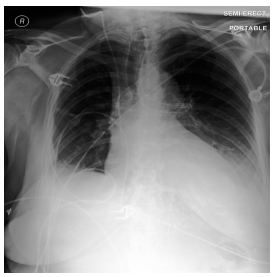
The increasing use of advanced cross-sectional imaging and the evolution of the information technology infrastructure to meet the demands of higher imaging volumes (i.e., improved

computational power, storage capacity, and workflow efficiency in the picture archiving and communication system (PACS) environment), contributed to a substantial increase in the number of images generated per examination [106]. Consequently, this has increased the workload of radiologists, which must now interpret more examination images in less time, thus creating the possibility for increased detection errors as a result of increased fatigue and stress, lowering the quality of the healthcare delivered by the radiologists to the patients [98; 174]. Moreover, as the ratio of diagnostic demand to the number of radiologists increases, the diminished effective available time per diagnostic becomes a critical issue [154]. According to the current paradigm, in case of doubt about a suspected condition, radiologists often turn to public or internal image databases where similar disease-matching images of the diseases the radiologist has narrowed down can be searched and compared against (e.g., *Radiopaedia*). After reviewing all possible differential diagnoses (those considered originally and those that came up during the search), the radiologist weighs these diagnoses and usually gives 2 to 4 of them as possible diagnoses.

In this process, the radiologist ranks the images and creates an ordered set of images in his/her head. This task is time-consuming and often ineffective since it requires several iterations until a proper matching image supporting the final diagnosis is found. Moreover, these databases are limited in the variability of cases presented to the users, which is exacerbated in conditions of low prevalence. Hence, it is extremely relevant to develop disease-targeted content-based image retrieval (CBIR) systems that automatically present disease-matching similar images to the one being analysed. A CBIR system usually focuses on two different tasks: feature representation, which consists of finding a low-dimensional representation of the image that is suitable for characterising it well enough; and, feature indexing and search, which focus on the efficiency of the retrieval process [101]. Our work focuses on the first step, i.e., on finding the most appropriate feature representation for the task at hand.

Finding the most appropriate feature representation is an arduous task since the clinical analysis is typically constricted to a small region of the image, discarding most of the available information. As such, finding the overall most similar image (i.e., including all pixels in the image) is not the objective. Instead, we are interested in finding the most similar image in terms of disease and disease severity. As illustrated in Fig. 3.1, those can be quite far apart, as Fig. 3.1 (b) is, overall, more similar to Fig. 3.1 (a) than Fig. 3.1 (c), while in terms of disease and disease severity, it is the opposite, with Fig. 3.1 (b) being the least similar image and Fig. 3.1 (c) the most similar (from a catalogue of 10 images).

Given that the disease features are located in a small region of the entire image, the medical CBIR system should also be paying attention to that specific region, ignoring the remaining information. However, most CBIR systems perform their analysis taking the entire image into account, particularly the more traditional methods. Deep learning approaches have a better focus on the disease-related characteristics as they learn the appropriate feature representations to solve the classification task of interest. Thus, they represent an improvement in terms of focus when compared to the more traditional approaches. Nonetheless, we hypothesise that this can be further improved by increasing, even more, the focus of the network in the regions that matter to the deci-



(a) Test Image

(b) Least Similar Image

(c) Most Similar Image

Figure 3.1: Pleural Effusion test image and the least and most similar images of the catalogue according to our board-certified radiologist (in terms of disease and disease severity). The overall most similar image would be (b). However, such matching is not of radiological interest.

sion, and exploring two different techniques: one driven by interpretability, and another based on attention mechanisms.

3.2 Related Work

Medical Image Retrieval

The importance of having a good medical image retrieval system to help clinicians make a diagnosis was clearly pointed out in the previous section. Here, we will focus on presenting the most relevant CBIR works available in the literature. The main difficulties in the development of CBIR systems are related to the development of algorithms that generate useful semantic representations of medical images in order to effectively retrieve the most similar examples [52] and the integration of these algorithms in end-user applications [75; 187]. We will focus on the first difficulty. In that regard, several works were presented in the literature to find the most suitable representation to perform the retrieval: Tizhoosh [167] explored the use of bar code annotations as an auxiliary method for feature-based image retrieval; Srinivas *et al.* [164] implemented a clustering method that uses dictionary learning to group large medical databases and relies on different similarity measures (e.g., Euclidean) to perform image retrieval; Hofmanninger and Langs [85] proposed the re-mapping of visual features extracted from medical imaging data based on weak labels to obtain descriptions of local image content capturing clinically relevant information; Seetharaman and Sathiamoorthy [141] presented a unified learning framework for heterogeneous medical image retrieval based on a full range auto-regressive model with a Bayesian approach to extract meaningful image features; Ma *et al.* [104] created a method that consists of a weighted graph whose nodes represent the images and edges measure their pairwise similarities; Nowaková *et al.* [120] presented a novel method for fuzzy medical image retrieval using vector quantisation with fuzzy signatures in conjunction with fuzzy S-trees; Qayyum *et al.* [131], Ayyachamy *et al.* [3] and Owais *et al.* [126] trained CNNs on multimodal and multi-class data sets, and used the learned features and the classification results to retrieve medical images; Cai *et al.* [20] used a Siamese Network in the learning process, with the CNN of each branch being used to extract features, followed by

the application of a binary hash-mapping to reduce the dimensions of the feature vectors; Minarno *et al.* [109] used a CNN-based auto-encoder method in the feature extraction process to improve the results of the retrieval process; Mbilinyi *et al.* [105] used a deep metric learning approach and the triplet loss to learn a model that receives an image and a text description highlighting specific diagnoses the retrieved images should have. In summary, feature representation is performed in one of the following ways: statistical measures, hand-crafted features, learned features, or a combination of the previously mentioned strategies. However, to the best of our knowledge, none of the previously proposed approaches explicitly focuses the training process on the disease-related characteristics without requiring additional labels. In this work, we aim to utilize AI interpretability methods to guide the retrieval process, with a focus on the disease and without necessitating any additional related label information.

Explainable Artificial Intelligence

The literature in Explainable Artificial Intelligence is presented in the previous chapter, where we presented all approaches as being either pre-model, in-model or post-model. In this work, we will focus on post-model interpretability strategies, as we are interested in finding the most relevant regions for the medical decisions (explicit attention) without limiting in any way the learning process or requiring any additional label. This can be done by identifying the areas of the image that mostly contribute to the final decision. To find these relevant regions, we used *Deep Taylor* [111], which is a relevance propagation approach (similar to Layer-wise Relevance Propagation (LRP) [7]), that uses deep Taylor decomposition to efficiently assess the importance of single pixels in image classification problems. The choice of this interpretability method in specific was mainly driven by its recognized quality, but also because it was the method that produced the saliency maps more in-line with what our board-certified radiologist considered as relevant medical information.

Attention Mechanisms

A different alternative to the use of post-hoc interpretability methods to focus the network on the disease-related characteristics is the use of implicit attention mechanisms. This application of attention mechanisms in deep learning algorithms was inspired by the field of psychology, according to which humans tend to selectively concentrate on a part of the information [176]. For instance, the human visual system tends to selectively focus on specific parts of an image while ignoring others [183]. The use of attention was initially proposed by Bahdanau *et al.* [9], for the task of neural machine translation. In this work, the authors use an encoder-decoder architecture presenting two challenges: 1) the decoder needs to compress all the input information into a single fixed-length vector and pass it to the decoder; 2) ensuring model alignment between input and output sequences was not possible. Hence, it was necessary to develop an attention mechanism that could support the decoder in focusing on the relevant parts of the inputs [38]. Naturally, during the training phase, an extra task is added: the learning of the attention weights. Nevertheless,

this approach showed improved results against the state-of-the-art and paved the way for the creation of novel attention-based methodologies. Attention models can be classified into different categories according to their input sequences, output sequences, candidate states (hidden states of the encoder) and query states (hidden states of the decoder) [38]. Of relevance for this work are self-attention and multi-level attention, with self-attention being when the query and candidate states belong to the same input sequence, and multi-level attention when we apply the attention mechanism on multiple levels of abstraction of the input sequence. Additional details on the attention mechanisms used in this work will be presented later when discussing their application in content-based image retrieval.

3.3 Methodology

Structural Similarity Index (SSIM)

The first method to be considered for evaluation in the retrieval task is the classic statistically-based structural similarity index (SSIM) [179]. As in [154], the SSIM was computed directly between test and catalogue images, using its default values. Since higher SSIM values represent higher similarity, the top retrieved image is the one with the highest similarity index.

Convolutional Neural Network (CNN)

The second method to be considered is already a deep learning based method, where the relevant features are automatically identified [85; 101; 146; 181]. In our preliminary work [154], we use the DenseNet-121 [86] as our CNN architecture. However, in these experiments, we do not initialize its weights with the ImageNet pre-training. Instead, we use a pleural effusion CNN model pre-trained in the CheXpert dataset [154] for the pleural effusion condition, and the new pleural effusion model for the pneumonia condition, as pre-training using data more similar to the final domain is more effective than using ImageNet pre-training [116; 180]. Similarity between images is computed based on the Euclidean distance in the feature space of the previous to the last layer of the model. Since shorter distances represent higher similarity, the top retrieved image is the one with the shortest distance to the test image. The distance between two images is formalized in Eq. 3.1, where I_t represents the test image t, I_c represents the catalogue image c, θ_{CNN} represents the CNN model parameters, and F represents the function that translates the original image into a latent representation constituted by the features in the previous to last layer of the network (i.e., in a vector of dimension 1024).

$$d_{CNN}(I_t, I_c) = ||F(\theta_{\text{CNN}}, I_t) - F(\theta_{\text{CNN}}, I_c)||_2 \tag{3.1}$$

Interpretability-guided Network (IG)

The third method being considered uses the exact same architecture as the CNN model, but has as input the saliency maps, instead of the original images (Fig. 7.1) in order to focus the network on

the disease-related characteristics [154]. Those saliency maps are computed using the Deep Taylor interpretability method [7], and are based on the previously presented CNN model. Following the same strategy as before, the network was initialized with the IG model pre-trained in CheXpert [154] for the pleural effusion condition, and with the new IG pleural effusion model for the pneumonia condition. The similarity is computed based on the previous to last layer of the model. The distance between two images is formalized in Eq. 3.2, where I_t represents the test image t, I_c the catalogue image c, θ_{CNN} the CNN model parameters, θ_{IG} the IG parameters, S the function that generates the saliency maps, and F the function that translates the original image into a latent representation constituted by the features in the previous to last layer of the network.

$$d_{IG}(I_t, I_c) = ||F(\theta_{\text{IG}}, S(\theta_{\text{CNN}}, I_t)) - F(\theta_{\text{IG}}, S(\theta_{\text{CNN}}, I_c))||_2 \tag{3.2}$$

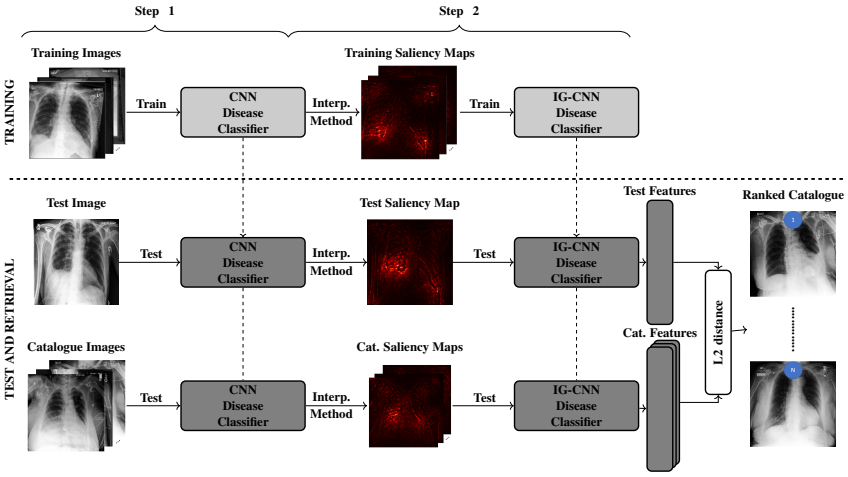


Figure 3.2: Overview of the proposed Interpretability-guided approach. Blocks in light gray (■) mean neural networks are being trained (i.e., weights are being updated), whereas blocks in dark gray (■) represent trained neural networks (i.e., weights are fixed). In the saliency maps, brighter colors mean higher relevance. Blue circles indicate ranking positions. CNN represents the deep model used as baseline. IG-CNN represents the CNN model architecture being trained with saliency maps.

Attention Network (ATT)

The fourth method is driven by attention mechanisms. Recently, a CNN with a multi-level dual-attention mechanism (MLDAM) has been proposed for macular optical coherence tomography classification [110]. The main novelty of this work in the context of medical image classification is the joint application of a *self-attention* and a *multi-level attention* mechanisms that allow the network to learn relevant features in coarser as well as finer sub-spaces. In their article [110], the authors state that this technique enables the network to utilise the information of coarser features preventing loss of any useful information, thus enabling the network to yield more focused features and better convergence. Regarding the impact of the application of attention mechanisms in the interpretability of deep learning algorithms, Chen and Ross [39] proposed the joint use of a

position attention module (PAM) and a channel attention module (CAM) to refine the pixel values at spatial and channel levels. These refined features are then fused through an element-wise sum. The authors performed an analysis of the saliency maps produced by the gradient-weighted class activation mapping (Grad-CAM) [142] and concluded that the use of attention modules had enabled the network to shift the focus on to the annular iris region.

In this work, we aimed to assure diversity in the levels and scales of the features extracted from the DenseNet-121 [86]. Following the notation in Mishra *et al.* [110], let I_A, I_B and I_C be the multi-level features extracted from the backbone. We extracted features from different dense-blocks resulting in a I_A with shape $[512, 28, 28]$, a I_B with shape $[1024, 14, 14]$ and a I_C with shape $[1024, 7, 7]$.

In line with the previous deep methods, the similarity is computed by measuring the Euclidean distance in the previous to last layer. The distance between two images is formalised in Eq. 3.3, where I_t represents the test image t, I_c the catalogue image c, θ_{ATT} the Attention model parameters, and A represents the function that translates the original image into a latent representation constituted by the features in the previous to last layer of the network.

$$d_{ATT}(I_t, I_c) = ||A(\theta_{\text{ATT}}, I_t) - A(\theta_{\text{ATT}}, I_c)||_2 \tag{3.3}$$

Deep Learning Networks Training

All deep learning methods (i.e., CNN, IG, and ATT) were trained to solve binary classification tasks (e.g., pleural effusion vs. non-pleural effusion). Thus, we use the binary cross-entropy as our loss function (Eq. 3.4, where y is the binary indicator, ln the natural logarithm, p the predicted probability, and θ the model parameters).

$$\mathscr{L}(\theta) = -(y \ln(p(\theta)) + (1-y) \ln(1-p(\theta))) \tag{3.4}$$

For the pleural effusion condition, the deep learning models were trained for 10 epochs, with a batch size of 32, and using the *Adadelta* optimiser [184]. Since the data for the pleural effusion condition is highly imbalanced, the misclassifications were weighted with the inverse of the frequency of the respective class to promote a similar focus of the network in both classes [165].

Regarding the pneumonia condition, the deep learning models were trained for 15 epochs, with a batch size of 32, and using the *Adam* optimiser [93] with a learning rate $l_r = 1 \times 10^{-4}$. The *Adam* optimiser was chosen over the *Adadelta* due to converging issues during the training of the CNN model, and was kept for the training of the other deep models (IG and ATT) for consistency.

For both conditions, small rotations and translations were used as data augmentation. Hyper-parameter values were empirically optimised for the CNN models and replicated for all the others. Final models were selected based on the F1 score (Eq. 8.12) in the validation set.

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3.5}$$

We note that the training process was agnostic to the ranking task at hand. No information of ranking was provided at any point, neither in the loss function nor in the selection of the best performing model in validation.

The methods were implemented using Keras [44] with TensorFlow backend in a workstation equipped with an NVIDIA Tesla V100 (32 GB) GPU. For the generation of the saliency maps, we used the *iNNvestigate* toolbox [1] implementation of the Deep Taylor method [7].

Evaluation and Comparison

The quality of the retrieval is evaluated by computing the normalised Discounted Cumulative Gain (nDCG) - Eq. 3.6, which is the normalised version of the Discounted Cumulative Gain (DCG) - Eq. 3.7, being a common metric in learning to rank tasks [60]. The normalisation is done over the maximum possible value of the DCG metric (in our work, the maximum possible value is obtained when the ranking of the method is exactly the same as our ground-truth). The subscript p represents the number of retrieved images we are considering for the evaluation (e.g., when we perform the evaluation over the entire set of retrieved images, $p = 10$). In Eq. 3.7, rel_i represents the relevance value assigned to the ranking position i, with the least similar image having a relevance of 1 and the most similar image having a relevance of 5.5 (i.e., the relevance of two contiguous positions differs by 0.5). Thus, the first positions of the catalogue ranking have more importance than the last ones, with the importance being gradually reduced as we go from the first to last ranked image.

$$\mathrm{nDCG}_p = \frac{DCG_p}{IDCG_p} \tag{3.6}$$

$$\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{3.7}$$

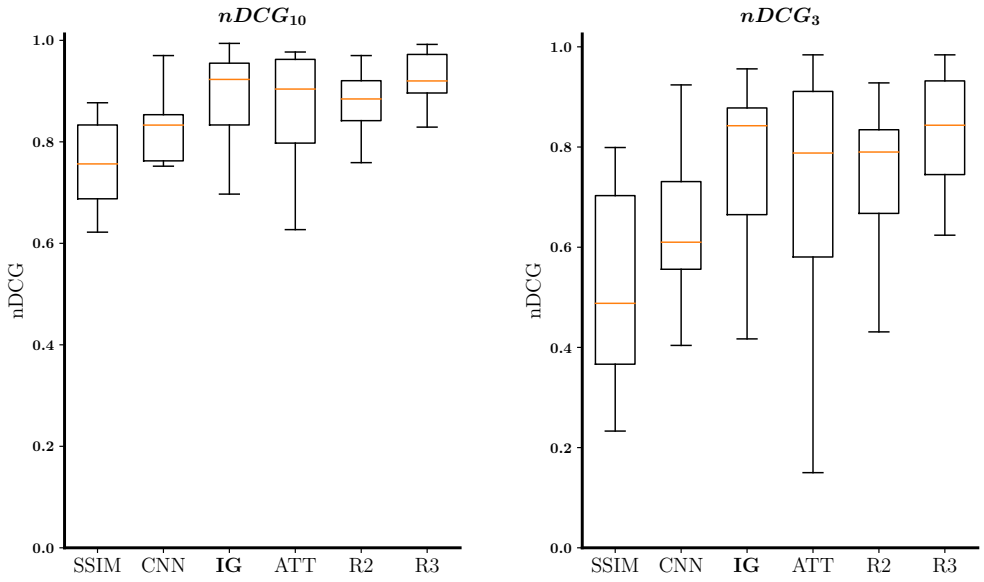
In order to contextualize the retrieval results of our machine learning methods, we also asked our partner board-certified radiologists to provide their similarity rankings for the pleural effusion and pneumonia conditions. Thus, we are able to check the inter-rater variability in ranking tasks, also helping us to have a more complete evaluation of our methods' quality.

3.4 Results and Discussion

Pleural Effusion

Our first experiments were conducted for the pleural effusion condition. All images used here were frontal X-ray images acquired in an AP view fashion. Thus, experiments were performed with 61203 training images, 534 validation images, and 1072 test images. The training images were used to find the optimal set of parameters, the validation images to select the final classification model, and the test images for the assessment. To evaluate the ranking quality, ten different query images and ten catalogues of ten images each were randomly created, splitting the test data into

query and catalogue images by using ten different random seeds (keeping the proportion of the classes). Afterwards, our main board-certified radiologist provided us with a ranking of those ten images in relation to the respective query image, serving as our ground-truth ranking. Moreover, we also asked two other board-certified radiologists to provide their rankings in order to compare inter-rater variability with our models' performance.



(a) nDCG results for Top-10

(b) nDCG results for Top-3

Figure 3.3: Box-and-whisker plots regarding the nDCG results for the pleural effusion Top-10 (a) and Top-3 (b) retrieved images. SSIM is the statistically-based baseline, CNN is the CNN-based baseline, IG is the proposed interpretability-guided approach, ATT is the attention method, R2 is the ranking provided by the second board-certified radiologist, and R3 is the ranking provided by the third board-certified radiologist.

In Fig. 3.3 (a), we present the nDCG results obtained with the statistical and machine learning models (i.e., SSIM, CNN, IG, and ATT) and also the results obtained by considering the rankings provided by two other radiologists (R2, and R3) for the Top-10 retrieved images. By observing the box-and-whisker plot, we conclude that the proposed interpretability-guided approach (IG) and the attention-based method (ATT) are the ones that lead to the best nDCG results for the Top-10 retrieved images, with the interpretability-guided approach outperforming the attention-driven method. Those results are in line with those from the other radiologists, demonstrating the high-quality of both methods. Furthermore, the CNN approach leads to better results than the SSIM method, as was expected. The same can be observed in Fig. 3.3 (b), where the nDCG results for the Top-3 retrieved images are presented (in clinical practice having the three most similar images is typically enough to help the radiologist make the diagnosis). In this scenario, nDCG values are worse than in the previous experiment due to only considering the Top-3 retrieved images, highly

penalizing a "failure" in one or more of these images. This also contributed to an increase in the variability of the results obtained, particularly in the case of the ATT method. Nonetheless, IG and ATT approaches remained the best methods and are still in line with the performance of the two radiologists.

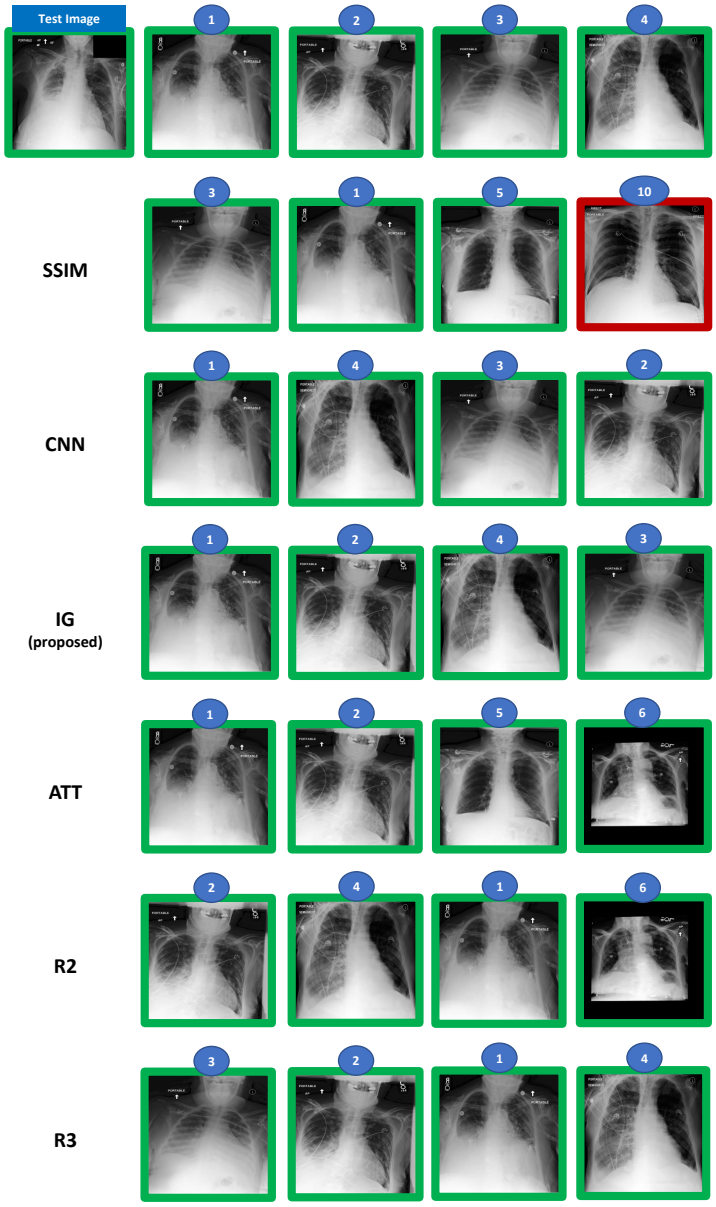


Figure 3.4: Example of test case and the Top-4 retrieved images given by each of the radiologists (R1 = ground-truth, R2, and R3) and each of the machine learning methods. In this split, both the CNN, IG, and ATT obtained nDCGs (Top-10) > 0.9. The green box means pleural effusion case and the red box means no pleural effusion (according to the dataset label).

In Fig. 3.4, we show the Top-4 retrieved results obtained by each of the methods, and provided

by the radiologists in comparison with the ground-truth defined by our main radiologist for one split, corresponding to a specific test case and catalogue. In this split, all machine learning methods (i.e., CNN, IG, and ATT) attained extremely high nDCG results. Both CNN and IG retrieved the same Top-4 images, with the only difference being the ranking of these four images, with IG's ranking being closer to the one provided by our main radiologist than CNN's ranking. Even though the ATT's Top-4 retrieved images differ from the ones selected by our main radiologist, one of those images was also selected by one of the other radiologists (i.e., R2). SSIM was the worst method, selecting the least similar image (a non-pleural effusion case) for the Top-4 retrieved images.

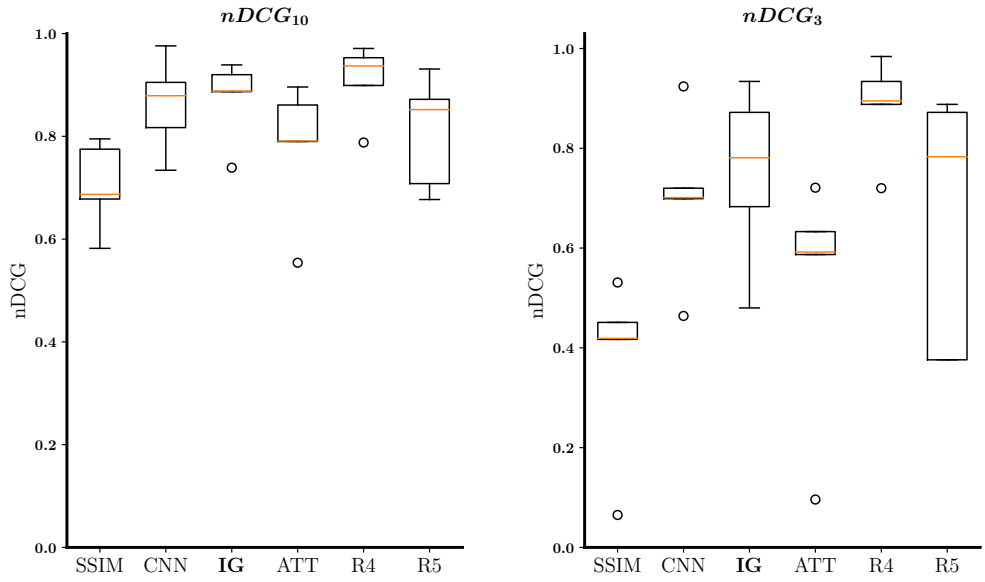
(Potential) Pneumonia

The following experiments were conducted for the pneumonia condition. All images used here were frontal X-ray images acquired in a PA view fashion. Thus, for the experiments, we considered 18226 training images, 133 validation images, and 258 test images. For the ranking evaluation, five different query images and five catalogues of ten images each were created, splitting the data into query and catalogue images by using five different seeds (keeping the proportion of the classes). Afterwards, the catalogue's ten images were ranked in terms of their potential as pneumonia cases to the respective query image. Even though our dataset annotations for training and validation were pneumonia annotations, our main radiologist considers a Chest X-ray as only indicative of potential pneumonia, and not of a definitive diagnosis. Thus, catalogue images were ranked having in mind their potential as pneumonia cases.

In Fig. 3.5 (a), we present the nDCG results obtained with the statistical and machine learning models and also the results obtained by considering the rankings provided by two other radiologists (R4, and R5). By observing the box-and-whisker plot, we infer that the proposed interpretability-guided approach (IG) is the method with the best retrieval ranking performance. On the contrary, for this condition, the attention-based method (ATT) had a poor ranking performance, obtaining nDCG results worse than the ones obtained with our deep learning baseline method (CNN), and only surpassing the performance of the statistical baseline (SSIM). The relative performance of the four methods was the same when we measured the nDCG performance for the Top-3 retrieved images (as shown in Fig. 3.5 (b)). IG's results also fall within the inter-rater variability of the radiologists, which demonstrates the quality of the method.

When we compare the quantitative results obtained for the pneumonia condition with the ones obtained for the pleural effusion, we observe that they were considerably worse in general. That may be due to pneumonia being a more difficult to diagnose condition, and also to different interpretations of what a pneumonia Chest X-ray is (in several catalogue images, there was a disagreement between the MIMIC-CXR label and the diagnosis provided by our main board-certified radiologist).

In Fig. 3.6, we present an example query case and the respective Top-4 retrieved images obtained by each of the methods and provided by the radiologists in comparison with the ground-truth



(a) nDCG results for Top-10

(b) nDCG results for Top-3

Figure 3.5: Box-and-whisker plots regarding the nDCG results for (potential) pneumonia Top-10 (a) and Top-3 (b) retrieved images. SSIM is the statistically-based baseline, CNN is the CNN-based baseline, IG is the proposed interpretability-guided approach, ATT is the attention method, R4 is the ranking provided by the fourth board-certified radiologist, and R5 is the ranking provided by the fifth board-certified radiologist.

defined by our main radiologist. In this split, all deep learning models had a reasonably good ranking performance, with the interpretability-guided approach (IG), and the attention-based method (ATT) retrieving in the first position the most similar image in terms of pneumonia to the test image. As can be observed here, some images in this catalogue had different diagnoses given by MIMIC-CXR and by our main radiologist, namely the fourth and sixth ranking positions (images with orange boxes). Moreover, the direction of the disagreement was the same, with our main radiologist considering the cases as of potential pneumonia, and the MIMIC-CXR label being non-pneumonia. Nonetheless, even with this label disagreement, the performance obtained with our interpretability-guided approach (IG) was reasonably good, exceeding nDCG results of 0.88 for the Top-10 retrieved images in all but one split.

Ablation study

We also studied the relevance of training with the saliency maps, and not only using them to compute the features in the semantic space. In Fig. 3.7, we present the Top-10 nDCG results for both pleural effusion and potential pneumonia, considering the CNN baseline model and these two versions, i.e., using only the saliency maps as inputs to the CNN model - CNN(IG) - and our proposed method where we use the saliency maps both in the training and retrieval processes -

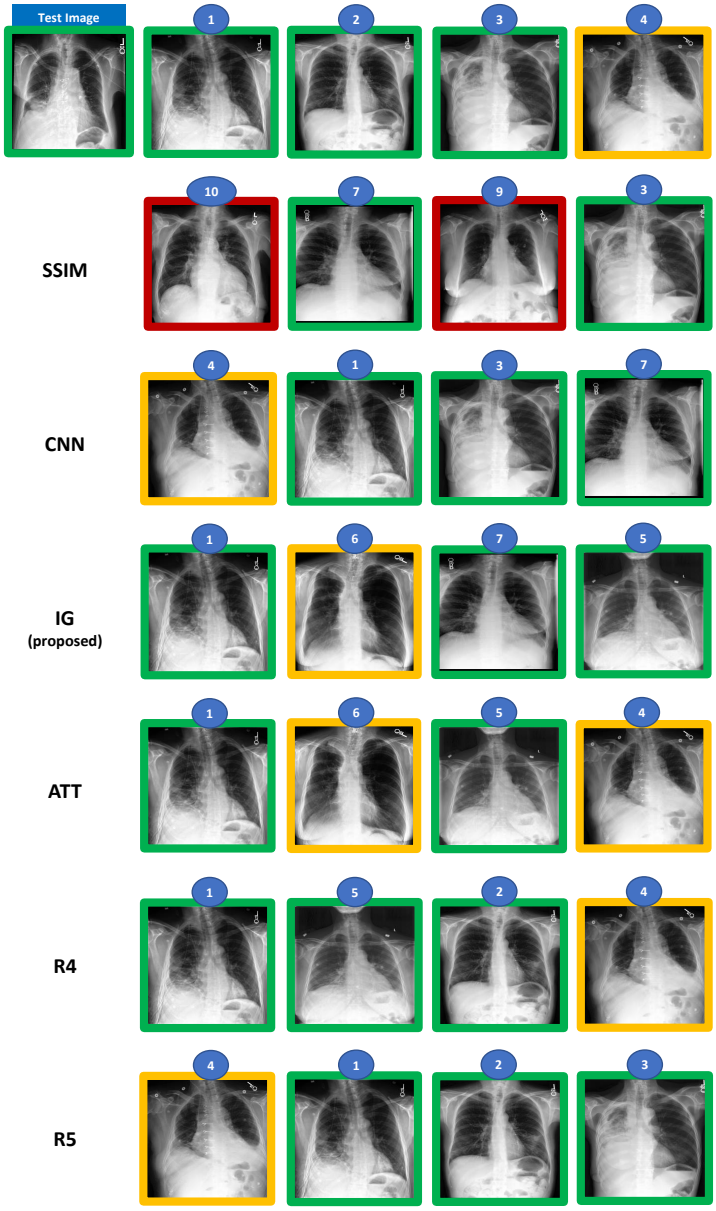
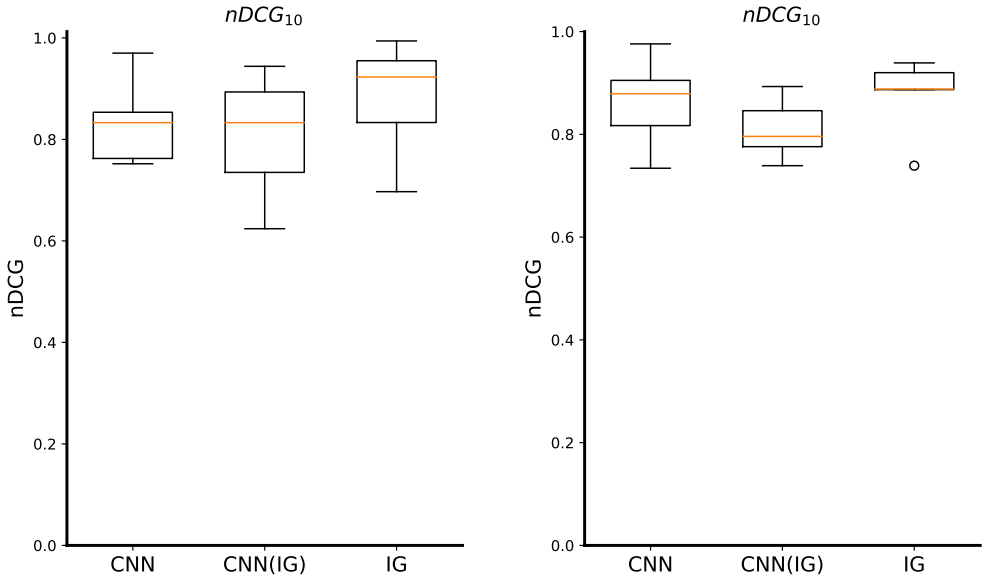


Figure 3.6: Example of test case and the Top-4 retrieved images given by each of the radiologists (R1 = ground-truth, R4, and R5) and each of the machine learning methods. In this split, both the CNN, IG, and ATT obtained nDCGs (Top-10) > 0.8. The green box means potential pneumonia case, red box means no potential pneumonia, and orange box means a disagreement between R1 and label, with R1 considering the case as potential pneumonia.

IG. By observing Fig. 3.7, we conclude that the use of saliency maps in the training process helps to attain better results, which can be explained by an increase in the focus of the network on the relevant disease regions, and by learning this new saliency map distribution.



(a) Pleural Effusion (b) (Potential) Pneumonia

Figure 3.7: Box-and-whisker plots regarding the nDCG results for Top-10 retrieved images. CNN is the CNN baseline model, CNN(IG) is the CNN model having as inputs the Deep Taylor saliency maps, and IG is the proposed interpretability-based approach (i.e., it was trained (fine-tuned) with saliency maps, and has as inputs also saliency maps).

3.5 Summary and Conclusions

We have investigated the use of different content-based image retrieval methods in a Chest X-ray retrieval task, intending to study their potential to support a medical diagnosis. For radiologists, more important than having a decision support system providing prediction labels is to have a system that is able to present them with similar clinical cases, as it is usually the way they proceed when encountering a difficult diagnosis scenario. Moreover, radiologists feel more comfortable working with images than with textual descriptions, motivating the use of case-based reasoning or explainability.

For the medical image retrieval to be successful, the comparison between images has to take into account the particular nature of medical images, i.e., that the information of interest is commonly found in a specific region of the image, while the remaining information is irrelevant. Indeed, the structural similarity index method showed the poorest performance for both conditions, demonstrating that a general image comparison does not represent disease similarity. Driven by that notion, we proposed an interpretability-guided approach and investigated the use of attention mechanisms for the retrieval task. The proposed interpretability-guided medical image retrieval approach outperformed all the other studied methods for both considered conditions. Our approach has an explicit attention mechanism that is also more intelligible than the implicit attention

mechanism of the attention-driven method, leading to a more interpretable solution. Moreover, it obtained a performance in line with other human experts (board-certified radiologists) for both conditions. In turn, the attention-based medical image retrieval method had an excellent performance for the pleural effusion condition (in line with both our proposed method and the other radiologists) but failed for the pneumonia condition.

It is important to emphasize that all methods were only trained to solve binary classification tasks, not using any ranking information. This means that the annotation effort required is significantly lower than if ranking information was also needed. Even though we did not use ranking information in the training process, our proposed approach correctly captured the ranking information, obtaining excellent nDCG results for both conditions. The test and catalogue images were not only ranked but also labelled by our main board-certified radiologist. Particularly for the pneumonia condition, we observed some disagreements in the diagnosis, which may be indicative of the usage of different definitions, and may hinder the method's performance. Nonetheless, even for the pneumonia condition, our proposed method obtained an excellent ranking performance.

This work aims to be the first step towards a deeper focus on medical image retrieval as a decision support system, helping radiologists make better and quicker decisions. However, further studies and investigations are required in order to translate these algorithms into clinics. Considering the evaluation aspect, it is crucial to have more extensive studies, more datasets, other clinical problems, and more radiologists involved in the annotation and evaluation process. Regarding the technical side, there are several open problems or investigation opportunities, namely, the use of multimodal data, the introduction of causal knowledge [36], privacy-preserving image retrieval [113], and also exploring federated learning settings [188]. By the use of multimodal data, we mean the integration of the clinical reports in the learning process, also with the possibility of accompanying the top retrieved images with a generated clinical report to provide complementary information, which can be particularly interesting when the end-user is a general practitioner instead of a radiologist. In this work, we observed that by using post-hoc interpretability saliency maps, we were able to focus model attention on more clinically relevant regions. However, those methods are only able to capture correlations. Thus, they may also focus on confounding information [68]. In order to prevent this from happening, the integration of a causal structure is essential. For clinical applications where personal characteristics are exposed, primarily if these systems are used for educational purposes, where the images are shown to unauthorized personnel, it is extremely important to anonymize the retrieved cases before showing them. Even though Montenegro *et al.* [113; 114] already explored the use of privacy-preserving methods to anonymize medical images, further research is required in order to improve realism and preservation of clinical information. Finally, it is also relevant to explore federated learning settings for this particular purpose, as the training process would benefit considerably from using different datasets acquired with different scanners and representing different population characteristics.

Chapter 4

Privacy-preserving Case-based Explanations

Foreword on Author Contributions

The results of this work have been disseminated in the form of one paper in an international conference, two articles in international journals and a Master's thesis:

- [114] H. Montenegro, W. Silva, and J. S. Cardoso, "Towards Privacy-preserving Explanations in Medical Image Analysis," in *1st Workshop on Interpretable Machine Learning in Healthcare at ICML (IMLH 2021)*, Jul. 2021. Available online at: https://arxiv.org/abs/2107.09652
- [113] H. Montenegro, W. Silva, and J. S. Cardoso, "Privacy-Preserving Generative Adversarial Network for Case-Based Explainability in Medical Image Analysis," *IEEE Access*, 9, 148037-148047, 2021. doi:10.1109/ACCESS.2021.3124844
- [115] H. Montenegro, W. Silva, A. Gaudio, M. Fredrikson, A. Smailagic, and J. S. Cardoso, "Privacy-preserving Case-based Explanations: Enabling visual interpretability by protecting privacy," *IEEE Access*, 10, 28333 - 28347, 2022. doi:10.1109/ACCESS.2022.3157589
- [112] H. Montenegro, "A privacy-preserving framework for case-based interpretability in machine learning," Master's thesis, Universidade do Porto, Portugal, 2021.

The Master's thesis [112] was supervised by Jaime S. Cardoso and co-supervised by Wilson Silva. This line of work is the natural sequence of the topics discussed in the previous chapter. My contributions consisted of the original research idea and conceptualization of the work.

4.1 Context and Motivation

The relevance of case-based explanations was highlighted in the previous chapter. However, sometimes case-based methods cannot be applied in domains with sensitive visual data, such as in the medical field, as they may compromise the privacy of individuals. For example, case-based explanations are difficult or impossible to use in the medical domain when data is shared with unauthorized personnel (e.g., medical students, interns, patients and family members). Privacy is less of an issue in saliency map methods, where by design, the only sensitive information they reveal is the input image under consideration.

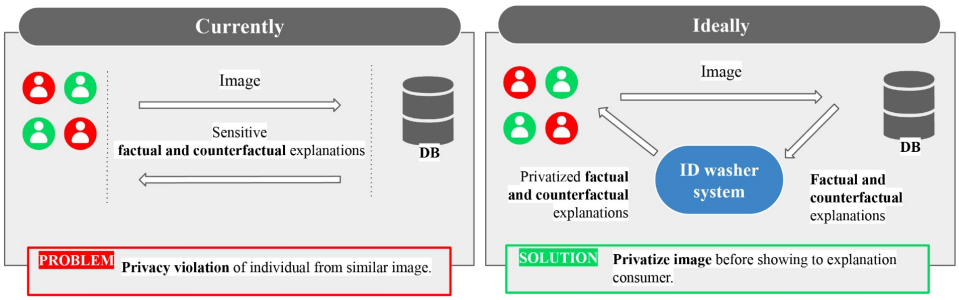


Figure 4.1: Diagram exemplifying the explanatory retrieval process. Consumers illustrated in red represent individuals who do not possess authorized access to the raw data (identity information) in the database. Consumers illustrated in green can access the raw data.

While privacy may not be an issue for retrieval with some public datasets, retrieval of private data poses challenges. Retrieving an image with sensitive identity information may violate the privacy of the individual present in the image. To address this issue, the case-based explanation must go through a privatization process to wash the identity from the image before presented to the consumer, as illustrated in Figure 4.1. The greatest challenge in creating the washer model is to ensure that no identity is leaked in the privatized version of the explanation and that explanatory evidence and realism are preserved.

In this chapter, the authors will demonstrate the weaknesses in the application of current privacy-preserving methods to medical data. Most of the current strategies fail to preserve relevant semantic features that serve as explanatory evidence in the context of case-based explanations. Furthermore, some privacy-preserving methods also fail to ensure privacy for all the subjects in the training data. This fact inhibits the use of these methods in the privatization of medical case-based explanations and highlights the need for new privacy-preserving approaches.

We will also detail our proposition that for a privacy-preserving explanation to work, it must fulfil the following prerequisites: privacy, explanatory evidence, and intelligibility. Regarding privacy, we argue that explanations must protect the privacy of all data subjects through a privatization mechanism that is independent of the training data. When it comes to explanatory evidence, the privatized explanations must contain the explanatory features that allow humans to understand their similarity to the case being explained. Finally, the explanations must be intelligible, as they are intended for humans.

Furthermore, we develop a privacy-preserving generative model that privatizes case-based explanations taking into account the above requirements, surpassing the state-of-the-art privacy-preserving methods. We guarantee privacy for the entire training data using a multi-class identity recognition network to promote a uniform identity distribution in the privatized images. The model also ensures the preservation of explanatory evidence by reconstructing relevant explanatory features obtained using interpretability saliency maps. Since the main domain motivating our research is medicine, we also adapt this model to situations where the data lacks images per identity (something pervasive in the medical field). For such scenarios, we use a Siamese identity recognition network [19] to aid privatization. Finally, we extend our model by generating counter-

factual explanations based on privatized factual explanations. In order to have a robust evaluation, we validate our model using the dataset Warsaw-BioBase-Disease-Iris v2.1 [168; 169], which is both medical and biometric and thus has well-defined identities.

4.2 Background

In this section, we will firstly provide the necessary background knowledge supporting the understanding of both the literature work and the developed methods.

Deep Generative Models

Generative Models learn the probability distribution of a training dataset and can use it to generate new data samples. In the context of privacy-preserving methods, these models generate anonymized images. Some case-based interpretability methods incorporate generative models to generate explanations. The most relevant generative models for this section are Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).

GANs [73] comprise two adversarial networks, a generator and a discriminator, that compete with each other in a minimax game. The generator generates new data samples intended to fool the discriminator. The discriminator distinguishes between real and generated instances. Ideally, the generated images from a trained GAN are close to a desirable latent distribution. The objective in Equation 4.1 describes the minimax game. The generator G minimizes the objective, while the discriminator D maximizes it, x represents real samples, and z represents the input to the GAN, often random noise. GANs are recognized as difficult to train, and the generator often undergoes mode collapse, a phenomenon describing an overall lack of diversity in output.

$$\min_G \max_D E_{x\sim p(x)}\left[\log D(x)\right] + E_{z\sim p(z)}\left[\log(1-D(G(z)))\right] \tag{4.1}$$

Privacy-preserving methods use conditional GANs (cGAN), a variation on the GAN restricted by predefined conditions. In specific, privacy-preserving GANs are conditioned by the input image, which contains certain features that must be preserved during privatization.

One variant of a GAN that is used in one privacy-preserving work is the WGAN-GP [77], which addresses mode collapse and stabilizes training with Wasserstein loss and a Gradient Penalty. The discriminator loss is shown in Equation 4.2, where $\hat{x}$ represents random samples obtained by a weighted average between real and generated samples, and λ refers to a non-negative penalty coefficient. The generator loss function is shown in Equation 4.3.

$$\mathscr{L}_D = E_z[D(G(z))] - E_x[D(x)] + \lambda\ E_{\hat{x}}[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2] \tag{4.2}$$

$$\mathscr{L}_G = -E_z[D(G(z))] \tag{4.3}$$

VAEs [94] learn an approximation of the data distribution using two networks: an encoder and a decoder. The encoder maps samples $x \sim p_{\text{data}}(x)$ in the original data space into a latent space with a simpler data distribution (usually a Gaussian distribution). The decoder maps samples $z \sim p(z)$ from the latent space into the original data space. This network allows the generation of new images by sampling from $p(z)$ and converting the samples to the original data space through the decoder. The loss function used to train a VAE, represented in Equation 4.4, contains two terms: a reconstruction term and a regularization term. The reconstruction loss term approximates a reconstructed image obtained through the VAE to its original version, and can be represented by loss functions like cross-entropy or mean squared error. The regularization loss term uses Kullback-Leibler (KL) Divergence to reduce the distance between the encoder's distribution $q_\theta(z \mid x)$ and the original data distribution $p(z \mid x)$.

$$\mathscr{L} = -E_{x,z}[\log p(x \mid z) + D_{KL}(q_\theta(z \mid x)||p(z \mid x))] \tag{4.4}$$

Besides the two Deep Generative Models introduced, there are other models in the literature with the potential to generate high-quality images, as is the case of Autoregressive Models and Normalizing Flows [132]. Research regarding these two models has been growing in recent years due to their capacity to model the data distribution explicitly as a tractable distribution without the need to perform approximations. These models may also be relevant in the future development of privacy-preserving models. Nonetheless, they are not used in any of the works analyzed in the related work sections.

Siamese Classification Networks

A Siamese Classification Network performs binary classification to recognize whether two images belong to the same class or to different classes. It was initially introduced by Bromley *et al.* [19] for signature verification. This network calculates a semantically-related distance between two data samples. The architecture comprises two identical networks with shared weights, responsible for extracting features from the data. After feature extraction, the network computes the distance between the samples' embeddings. A relevant loss function that can be used to train these networks is the contrastive loss [80]. This loss is represented in Equation 4.14, where D represents the Euclidean distance between the samples' embeddings, m represents a margin to bound this distance, and Y is 1 when the images do not belong to the same class and 0 otherwise. The contrastive loss ensures that embeddings from images belonging to different classes are more distant than embeddings from images sharing the same class.

$$\mathscr{L} = \frac{1}{2} \times (1-Y) \times D + \frac{1}{2} \times Y \times \{\max(0, m-D)\}^2 \tag{4.5}$$

This type of network is used in one of the works we are going to present in this chapter [113] but also on some other privacy-preserving networks [121; 182] as an identity recognition network to calculate the identity-related distance between an image and its privatized version.

4.3 Related Work

Privacy-preserving methods have been applied in medical imaging with the purpose of increasing the availability of medical data to train artificial intelligence algorithms [87]. Anonymization and pseudonymization techniques remove or alter metadata associated with the medical images (e.g., the patients' names). However, the images themselves expose identity, which can be used to identify the patients through re-identification techniques [140]. Encryption [186] results in unintelligible images that cannot be shown to humans as case-based explanations. Other privacy-preserving techniques avoid disclosing sensitive information about the data during a model's training. For instance, Federated Learning [95] consists of training the models in the data owners' servers to avoid sharing private medical data [137; 145]. Differential privacy [57] has also been applied to hide the contributions of individual patients during a model's training [188]. Nevertheless, these techniques cannot be applied to privatize case-based explanations, which are meant to be exposed to humans, as they act on the model and not on the data itself.

No privacy-preserving method for medical imaging considers altering the image to remove a patient's identifiable features while preserving disease-related information and the image's intelligibility. However, there are privacy-preserving methods in the literature capable of generating intelligible privatized images, that have been applied in domains other than the medical field. We discuss the methods in regards to their application to case-based explanations. Furthermore, we consider that identity-related features in the images may be entangled with explanatory features that must be preserved. We distinguish these methods in traditional and Deep Learning methods.

Traditional privacy-preserving methods are applied over the whole input, as they cannot identify sensitive image regions. These methods require an additional pre-processing step to locate the image regions that need to be privatized. The most well-known traditional method consists of applying filters such as blur to an image [66]. The most significant issue in this type of method is that relevant explanatory features are lost at the same rate as identity features. As such, privatized images with acceptable degrees of privacy do not preserve explanatory evidence [114]. Another famous class of privacy-preserving techniques is the K-Same-based family [76; 118], which was developed for face de-identification. In these methods, the privatized images are an average of various training images, guaranteeing K-Anonymity, where the highest probability of a person being recognized in the image is $\frac{1}{K}$. This technique imposes limitations on privacy, as the privatization process directly uses images from other subjects in the database, and in explanatory

evidence preservation. An alternative to those methods is face-swapping [15], which consists of replacing the faces in an image with models from a public database. Although this method guarantees privacy, if identity-related features and explanatory features are entangled, the replacement of the image regions that contain identity-related features will result in the loss of the associated explanatory features.

In Deep Learning, privacy-preserving models usually comprise a generative network responsible for generating privatized images and an identity recognition network that guides the privatization process. Some models directly obtain identity vectors from the images by disentangling identity-related features from the remaining features, as is the case with the CLEANIR model [43] and the R^2VAE model [72]. These identity vectors can then be altered to hide the original identity of the images. Other privacy-preserving strategies focus on creating privatized images that do not share the same identity as the original images by using a Siamese identity recognition network [19] to guide the generation of privacy-preserving images [121; 182]. These networks ensure image utility by maximizing the structural similarity between the original and privatized images.

The biggest problem in the previous deep learning methods is that none guarantees the preservation of relevant semantic features needed for a particular classification task. Privacy-preserving methods that preserve task-related features use a task-related classifier to ensure the feature preservation process. PPRL-VGAN [41] was developed for privacy-preserving facial expression recognition. It privatizes images through identity replacement. Although this model successfully hides the identity from the original image, it exposes the identities of other subjects in the data. As such, this model still violates privacy. Furthermore, this model only preserves the task-related class of the original image and not its explanatory features.

In general, none of the privacy-preserving models explores the explicit preservation of the original images' explanatory evidence. Furthermore, some of the models still possess privacy issues as they directly use training data in the privatization process.

4.4 Methodology

PPRL-VGAN

Initially, we explored the state-of-the-art deep learning model PPRL-VGAN [41], which preserves privacy through identity replacement. This method was chosen for being the only approach available in the literature that considered the preservation of semantic features needed for a particular task. In specific, this network was originally developed for privacy-preserving facial expression recognition.

PPRL-VGAN is a Generative Adversarial Network (GAN) comprising a conditional Variational Autoencoder (VAE) as the generator and a multi-task classifier as the discriminator, as shown in Figure 4.2. The generator G generates an image recognized as the given replacement identity c, preserving the task-related features of the original image. The discriminator D aids the generative task through a fake/real classifier D^1 to promote realism in the generated images, an

identity recognition network D^2 to aid the identity replacement process, and a task-related classification network D^3 to preserve relevant semantic features. Given an image I in the original data space with probability distribution p_d, the discriminator's loss function is shown in Equation 4.6, where the variables y^{id} and y^e correspond to the target labels for identity recognition and for the semantic task, respectively, and λ_x^D are parameters used to calibrate the importance of each task x during training.

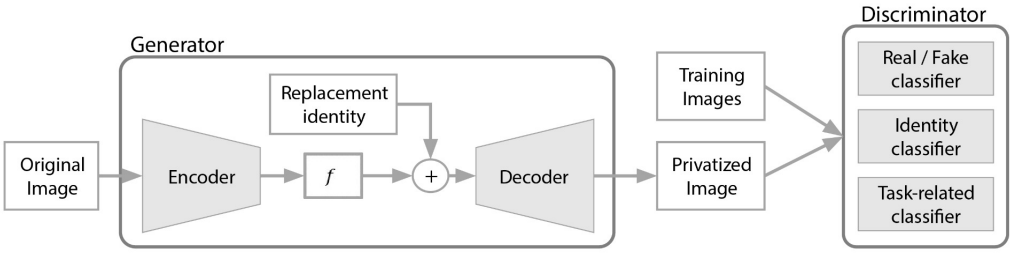


Figure 4.2: Overview of the PPRL-VGAN model's architecture.

$$\mathscr{L}_D = E_{(I,y,c)\sim p_d(I,y,c)}[\lambda_1^D \{\log D^1(I) + \log(1 - D^1(G(I,c)))\} + \lambda_2^D \log D^2_{y^{id}}(I) + \lambda_3^D \log D^3_{y^e}(I)] \tag{4.6}$$

The loss function used to train the generator is shown in Equation 4.7. This function includes a loss term for each of the discriminator's tasks and one term to regularize the VAE's latent space. Given an image's latent representation $f(I)$, the regularization loss term uses Kullback-Leibler Divergence (KL) to approximate the prior distribution on the latent space $p(f(I))$ and the conditional distribution $q(f(I) \mid I)$ parameterized by the encoder.

$$\mathscr{L}_G = E_{(I,y,c)\sim p_d(I,y,c)}[\lambda_1^G \log(1 - D^1(G(I,c))) + \lambda_2^G \log(1 - D^2_{y'(c)}(G(I,c))) + \lambda_3^G \log(1 - D^3_{y^e}(G(I,c)))] + \lambda_4^G KL(q(f(I) \mid I) \mid\mid p(f(I))) \tag{4.7}$$

From PPRL-VGAN to Privacy-preserving Case-based Explainability

There are various weaknesses with PPRL-VGAN that prevent its use for the privacy-preservation of case-based explanations. The most critical ones are the privacy violation inherent to using identity replacement as the privatization mechanism and the non-preservation of explanatory evidence as it exists in the original image [114]. Regarding applying this model to medical data, the model also has difficulty in disentangling identity-related factors in cases where most subjects only have images from one disease-related class [114]. Moreover, a multi-class identity recognition network is challenging to train when the data only has a small number of images per identity, as frequently happens in the medical context.

Privacy-Preserving Network with Multi-class Identity Recognition

Using the PPRL-VGAN model as a base, we defined a novel privacy-preserving network for the privatization of case-based explanations.

To ensure that privacy is preserved for every subject in the training data, we removed the replacement identity given to the decoder. Instead of creating an image that looks like the replacement identity, we try to keep the identity recognition close to random guessing (i.e., close to a uniform distribution).

By promoting a uniform distribution across identities, the generative task became more complex, leading to poor image quality and mode collapse problems. We pre-trained the identity recognition model and the task-related classifier on the dataset used to train the privacy-preserving model, to facilitate the generative task and improve image quality. In PPRL-VGAN, the mode intentionally collapsed to the identity given as replacement and to the task-related class from the original image. However, in our case, the mode collapse was unintentional and affected the explanatory value of the images, as they all looked identical. To fix this problem and improve image quality, we replaced the generative framework with a WGAN-GP network [77], using Wasserstein loss with gradient penalty to stabilize the discriminator.

We explicitly preserve explanatory evidence by using interpretability saliency maps to reconstruct relevant task-related features in the privatized images. In specific, we use Deep Taylor [111] to create masks containing the relevant image features. We input these masks into the generative network and concatenate them with the original images inside the VAE's encoder, after feature extraction and before calculating the parameters of a Gaussian distribution. In the loss function, we use the squared L2 loss to reconstruct relevant features. We also ensure that the privatized images are assigned the same classification score as the original images to aid the preservation of explanatory features.

We summarized the changes introduced to the PPRL-VGAN model in Figure 4.3. With these changes, we obtained a privacy-preserving model with three modules: a generative module, a privacy module, and an explanatory module.

Generative Module: The generative module is responsible for the generation of intelligible images, given an image I from the original data space's probability distribution p_d. It is composed of a GAN with a VAE as the generator G. The discriminator, D, is trained using Wasserstein loss and gradient penalty, as shown in Equation 4.8, where $\hat{x}$ corresponds to random samples and λ is the weight associated with the gradient penalty term. In the generator, there are two terms: a realness term to promote the generation of realistic images (Equation 4.9), and a regularization term in the VAE. The regularization term, shown in Equation 4.10, consists of approximating the prior distribution on the latent space $p(f(I))$, where $f(I)$ corresponds to the image I's latent representation, and the conditional distribution $q(f(I) \mid I)$ parameterized by the encoder.

$$\mathscr{L}_D = E_{I\sim p_d(I)}[D(G(I))] - E_{I\sim p_d(I)}[D(I)] + E_{\hat{x}\sim p_{\hat{x}}}[\lambda(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2] \tag{4.8}$$

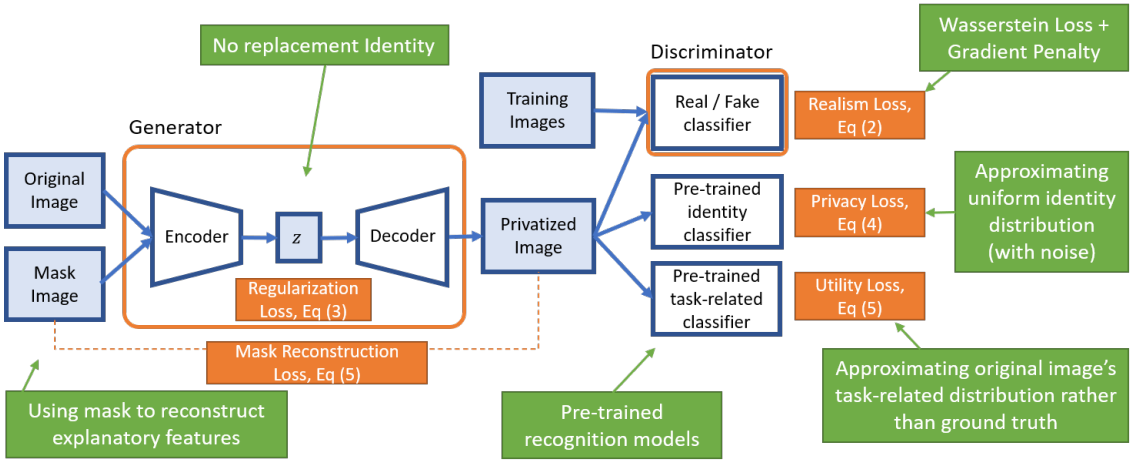


Figure 4.3: Overview of our privacy-preserving model's architecture. We included in green a summary of the changes that occurred to the original PPRL-VGAN model, to apply it to the domain of case-based interpretability.

$$\mathscr{L}_{realness} = E_{I\sim p_d(I)}[-D(G(I))] \tag{4.9}$$

$$\mathscr{L}_{reg} = E_{I\sim p_d(I)}[KL(q(f(I) \mid I)||p(f(I)))] \tag{4.10}$$

Privacy Module: The privacy module is responsible for anonymizing the images, guaranteeing privacy for the subjects in the image and in the database. Using a pre-trained multi-class identity recognition network D_{id}, we promote a uniform identity distribution in the privatized images. As such, the generator contains a privacy term in the loss function, represented in Equation 4.11. In this equation, U represents a uniform distribution with noise.

$$\mathscr{L}_{privacy} = E_{I\sim p_d(I)}[-D_{id}(G(I))log(U)] \tag{4.11}$$

Explanatory Module: The explanatory module is responsible for guaranteeing the privatized images' explanatory value. We preserve the explanatory evidence through the reconstruction of explanatory features in the images, using Deep Taylor saliency maps, M, obtained by applying the task-related classifier D_{exp} on the original images. We also approximate the privatized image classification score to the one in the original images. The generator loss terms representative of this module are shown in Equation 4.12.

$$\mathscr{L}_{exp} = E_{(I,M)\sim p_d(I,M)}[\lambda_3 D_{exp}(I)log(D_{exp}(G(I))) + \lambda_4(I\times M - G(I)\times M)^2] \tag{4.12}$$

Finally, the entire generator's loss is depicted in Equation 4.13. λ_x are parameters to control the importance of each loss term x.

$$\mathscr{L}_G = \lambda_1 \mathscr{L}_{realness} + \lambda_2 \mathscr{L}_{privacy} + \mathscr{L}_{exp} + \lambda_5 \mathscr{L}_{reg} \tag{4.13}$$

Privacy-Preserving Network with Siamese Identity Recognition

As it stands, our privacy-preserving model cannot be used in domains where the number of images per subject is scarce, which is frequently the case in the medical context, since a multi-class identity recognition network is hard to train in these scenarios. To widen the range of applications of our model, we replace the multi-class identity recognition network with a Siamese network [19], pre-trained on the dataset used to train the privacy-preserving model.

The Siamese identity recognition network compares the original image with its privatized version and computes their identity-related distance, which can be used to classify whether the images belong to the same identity or not. We trained this network using a contrastive loss [80], represented in Equation 4.14. In this equation, m represents a margin to limit the distance between images, Y represents the label assigned to an image pair (1 when the images belong to the same identity, and 0 otherwise), and ED represents the Euclidean Distance between the image pair embeddings.

$$ContrastiveLoss = \frac{1}{2} \times Y \times ED^2 + \frac{1}{2} \times (1-Y) \times [\max(0, m - ED)]^2 \tag{4.14}$$

By using this network, we ensure that the privatized image is different from the original image in terms of identity. To guarantee that the generated images also do not look like the images of the other identities present in the dataset, we use the Siamese network to increase the identity-related distance between the privatized image and the images from each of the subjects present in the database. In practice, at each epoch during training, we randomly select one image from each of the identities and promote that this image is far from the privatized image.

The privacy term of the generator loss function, when using the Siamese network, is represented in Equation 4.15, where N is the number of identities that exist in the dataset.

$$\begin{aligned} \mathscr{L}_{privacy} = E_{(I,N)\sim p_d(I,N)}[\lambda_2 [\max(0, m - ED(I, G(I)))]^2 + \\ \lambda_6 \sum_{i=0}^{N} \frac{[\max(0, m - ED(G(I), I_N))]^2]}{N}] \end{aligned} \tag{4.15}$$

Generation of Counterfactual Explanations

We also apply our model to the generation of counterfactual explanations. We add a counterfactual generation module to the previously defined privacy-preserving network in the form of a counter-

factual decoder responsible for mapping an image's latent representation to its counterfactual. To generate counterfactual explanations, we aim to perform the smallest number of alterations to the privatized factual explanations to change their predicted class. As such, the counterfactuals' decoder is trained to minimize the pixel-wise distance between the factual and counterfactual explanations while changing the original image's task-related prediction. We use the saliency masks with the explanatory features to promote changes in the image regions relevant to the explanatory classification task while preserving the remaining image parts. This network's architecture is shown in Figure 4.4.

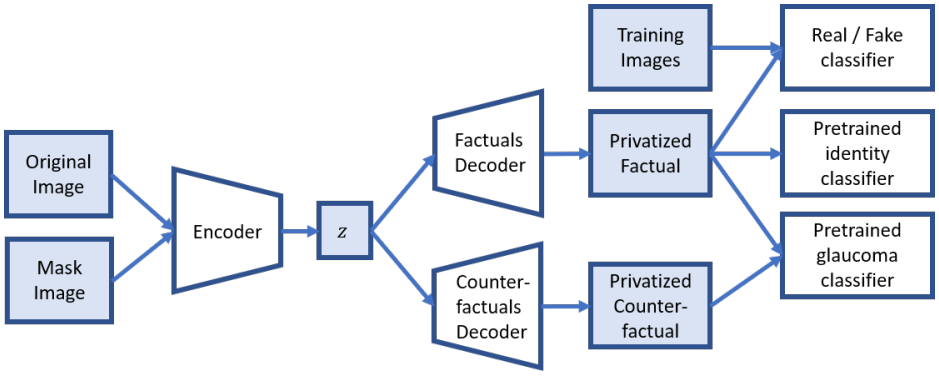


Figure 4.4: Architecture of the privacy-preserving model with generation of counterfactual explanations.

Regarding the training approach, we first train the factual decoder as in the previously presented networks, with the counterfactual decoder frozen. Then, we freeze the factual decoder and transfer its weights to the counterfactual decoder to train it. The generator's loss function used to train the counterfactual decoder is represented in Equation 4.16. In this equation, $F(I)$ and $C(I)$ denote the privatized factual and counterfactual explanations, respectively.

$$\begin{aligned}\mathscr{L}_C = E_{(I,M)\sim p_d(I,M)}[\lambda_7(F(I)\times(1-M)-C(I)\times(1-M))^2+ \\ \lambda_8 D_{exp}(I)log(1-D_{exp}(C(I)))]\end{aligned} \tag{4.16}$$

4.5 Results and Discussion

For the experiments, we used the medical and biometric dataset Warsaw-BioBase-Disease-Iris v2.1 [168; 169], composed of 2,996 iris images with various eye pathologies acquired from 115 different patients. We only used the 1,795 images taken from the device IrisGuard AD100, and we focused on one of the pathologies, glaucoma. The images were labeled according to the presence or absence of glaucoma. In the pre-processing stage, we cropped the images to remove labels in their lower corners, horizontally flipped the patients' right eye images, and centered the iris of the eye in the middle of the image. The images' resolution was set to 64×64 and they were split into 65% for training, 15% for validation, and 20% for testing. To obtain masks with relevant

glaucoma features located inside the iris, we generated iris segmentation masks and performed an AND operation between the Deep Taylor saliency maps and the iris segmentation masks.

Privacy-Preserving Model with Multi-class Identity Recognition

In the privacy-preserving model with multi-class identity recognition, we used as parameters $\lambda_1 = 0.4$, $\lambda_2 = 1$, $\lambda_3 = 2$, $\lambda_4 = 0.001$ and $\lambda_5 = 0.002$. We used $\lambda = 10$ in the discriminator's loss, as suggested in the original WGAN-GP paper [77]. We used the Adam optimizer with a learning rate of $2e^{-5}$. The model was trained for 1,184 epochs. The results are presented in Figure 4.5. Although the images possess some visible noise, they can be considered intelligible. We notice that the network has some difficulty creating a realistic eye structure surrounding the iris. In the visual results, the privatized image's Deep Taylor saliency maps closely resemble the ones from the original images, evidencing the correct preservation of explanatory evidence.

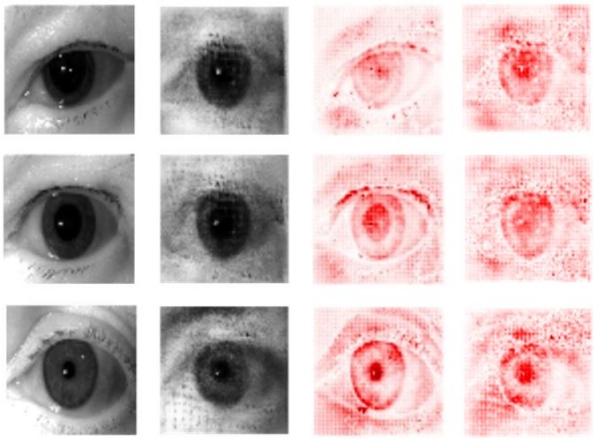


Figure 4.5: Results of the privacy-preserving model with multi-class identity recognition. The first and second columns represent the original images and their privatized versions, respectively. The third and fourth columns contain Deep Taylor saliency maps obtained from the original and privatized images, respectively.

We include in Table 4.1 the results achieved with this network. The identity recognition network's accuracy is evaluated at recognizing the subject from the original image. To evaluate privacy at the whole dataset's level, we analyze the maximum score that the identity recognition model assigns to an identity when making a prediction about a privatized image. We also evaluate the divergence between the privatized images' identity distribution and the uniform distribution, using KL Divergence. Finally, we assess the Glaucoma Recognition network's accuracy at detecting the original images' glaucoma score in the privatized images.

The low accuracy in identity recognition suggests that the privacy-preserving model succeeds at privatizing the images. The values for the maximum identity score and KL divergence suggest that the network has difficulty recognizing any identity, as these values are significantly lower than the baseline. Furthermore, the high values in glaucoma recognition accuracy advocate for the network's high capacity of preserving explanatory evidence.

Table 4.1: Results of the privacy-preserving model with multi-class identity recognition. We expect low values in the privacy-related metrics and high values in glaucoma recognition accuracy. The best results for each metric are highlighted in bold.

Dataset	**Identity Recognition Accuracy**	**Maximum Identity Score**	**Average KL Divergence**	**Glaucoma Recognition Accuracy**
Original testing set (baseline)	89.71%	88.22%	4.24	100.00%
Privatized set with explanatory evidence	**0.88%**	**33.15%**	**2.53**	**91.47%**
Privatized set without explanatory evidence	**0.88%**	34.49%	2.60	89.41%

During the network's development, the most significant challenge we came across was to manage the trade-off between privacy, intelligibility and explanatory evidence. In most cases, improving one of these dimensions would result in worsening the remaining ones. In our model, the most sacrificed dimension was intelligibility, as the generated images have poorer quality than the original ones. When we try to remove one of the other dimensions, the image quality improves. For instance, removing explanatory evidence results in the higher-quality results shown in Figure 4.6.

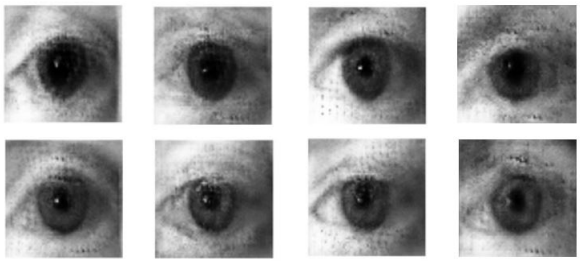


Figure 4.6: Comparison between results from the network when explanatory evidence is considered (first row) and not (second row).

Privacy-Preserving Model with Siamese Identity Recognition

Using the privacy-preserving model with Siamese identity recognition, with parameters $\lambda_1 = 0.4$, $\lambda_2 = 5$, $\lambda_3 = 2$, $\lambda_4 = 0.001$, $\lambda_5 = 0.002$ and $\lambda_6 = 10$, we obtained the results shown in Figure 4.7. The model was trained for 900 epochs.

This model provides higher-quality images than the previous multi-class identity recognition model. Nonetheless, the model also suffers from a trade-off between privacy, intelligibility, and explanatory evidence. For instance, when we remove the overall privacy term ($\lambda_6 = 0$), we obtain privatized explanatory features that resemble more closely the ones from the original images, as shown in Figure 4.8.

Table 4.2 exposes the results obtained with this model. To evaluate privacy, we use the previously developed multi-class identity recognition model as an evaluation network. Then, we use

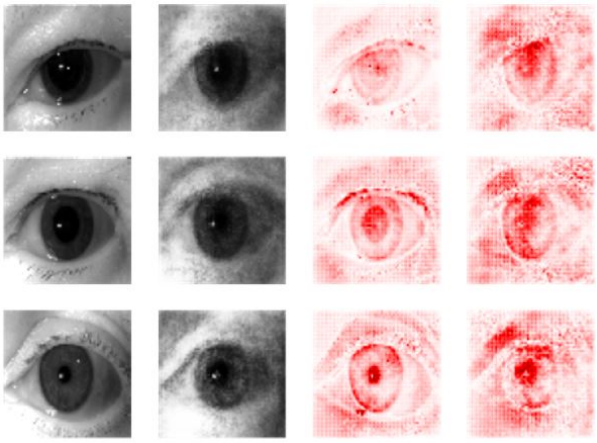


Figure 4.7: Results of the privacy-preserving model with Siamese identity recognition. The first and second columns represent the original images and their privatized versions, respectively. The third and fourth columns contain Deep Taylor saliency maps obtained from the original and privatized images, respectively.

the Siamese identity recognition model's accuracy at recognizing that the original and privatized images belong to different identities. To calculate this accuracy, we verify whether the distance between image pairs is higher than 0.777, corresponding to the average distance value obtained when using the Siamese network on image pairs from the original testing set. To evaluate the privacy in the whole dataset, we obtain the identity recognition accuracy when comparing the privatized images with an image from each identity available in the dataset. We also evaluate the number of pairs that are considered to be from the same identity (real pairs). In this table, we expect to achieve low values in multi-class identity recognition and average number of real pairs, and high values in the remaining metrics.

We obtained a higher privacy degree by considering overall privacy, as seen by the lower accuracy in multi-class recognition and higher accuracy in Siamese identity recognition. Furthermore, when we consider overall privacy, there are fewer images from the dataset's subjects that are considered to be from the same identity as the privatized images. The privatized set with overall privacy also achieved higher glaucoma recognition accuracy.

Counterfactual Generation

By adding a counterfactual generation module to the privacy-preserving model with multi-class identity recognition, with parameters $\lambda_7 = 0.001$ and $\lambda_8 = 1$, we were capable of inverting the glaucoma classification of the original image with 90.29% accuracy. With the model that uses Siamese identity recognition, we achieved 90.88% accuracy in inverting the images' glaucoma classification. Furthermore, in both models, the differences between the factual and the counterfactual explanations are located mainly in the iris region. An example of the obtained results using the Siamese identity recognition model is shown in Figure 4.9.

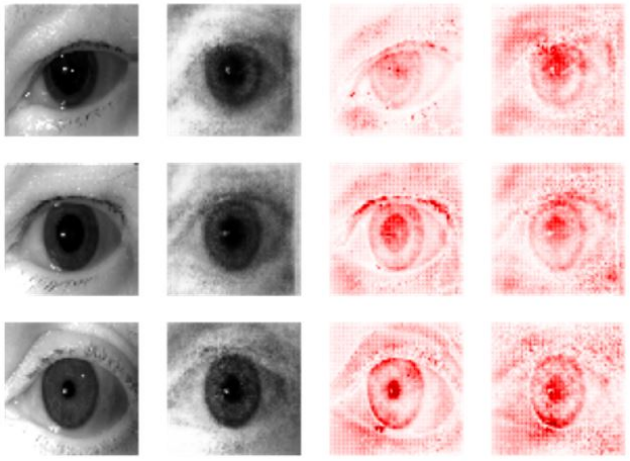


Figure 4.8: Results of the privacy-preserving model with Siamese recognition, not considering overall privacy. The first and second columns represent the original images and their privatized versions, respectively. The third and fourth columns contain Deep Taylor saliency maps obtained from the original and privatized images, respectively.

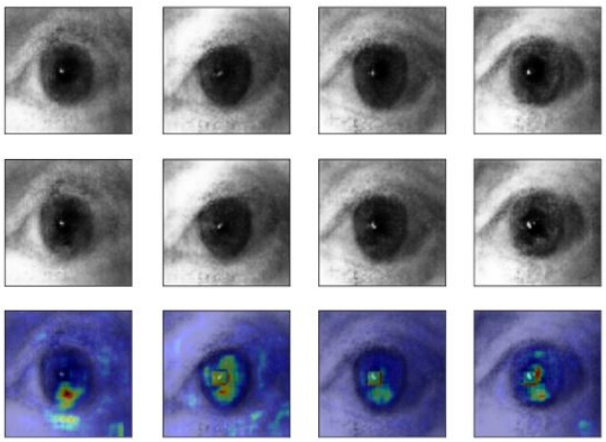


Figure 4.9: Results of counterfactual generation using privacy-preserving model with Siamese identity recognition. The first and second rows represent privatized factual and counterfactual explanations, respectively. The final row contains a map with the differences between the factual and counterfactual explanations.

In this experiment, we used the Deep Taylor glaucoma masks to promote changes located inside the iris and, thus, avoid alterations in zones that are irrelevant to glaucoma classification that may occur as an adversarial attack. However, even with these masks, the counterfactual decoder may be performing an adversarial attack on the glaucoma classification network, tricking it into misclassifying the samples and generating adversarial samples instead of counterfactual explanations.

Table 4.2: Results of the privacy-preserving model with Siamese identity recognition. The best results for each metric are highlighted in bold.

Dataset	Multi-class Identity Recognition Accuracy	Siamese Identity Recognition Accuracy	Siamese Recognition Accuracy (Whole Dataset)	Average Number of Real Pairs	Glaucoma Recognition Accuracy
Original testing set (baseline)	89.71%	83.80%	-	-	100.00%
Privatized set with no overall privacy ($\lambda_6 = 0$)	**0.88%**	89.41%	78.91%	22.99	88.53%
Privatized set with overall privacy ($\lambda_6 = 10$)	1.76%	**92.65%**	**91.99%**	**8.74**	**91.47%**

4.5.1 Ablation Study

To verify how the generator used in the privacy-preserving models fares in comparison with other state-of-the-art architectures, we replaced it with a ResNet VAE, which contains ResNet [83] as the encoder and decoder, and with a UNET architecture [135]. We performed this experiment with the multi-class identity recognition version of the privacy-preserving model. The results are shown in Table 4.3 and in Figure 4.10.

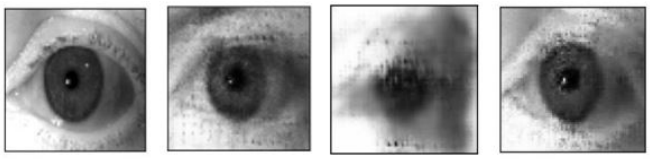


Figure 4.10: Results obtained by replacing the generator with other architectures. The first image is the original one, and the following images are privatized images using the original VAE, the ResNet VAE, and UNET, respectively.

Table 4.3: Results of replacing generator with ResNet VAE and UNET. For convenience, the first lines repeat the results shown in Table 4.1.

Dataset	Identity Recognition Accuracy	Maximum Identity Score	Average KL Divergence	Glaucoma Recognition Accuracy
Original testing set (baseline)	89.71%	88.22%	4.24	100.00%
Privatized set with original VAE	0.88%	33.15%	2.53	91.47%
Privatized set with ResNet VAE	**0.59%**	**9.0%**	**0.74**	84.41%
Privatized set with UNET	4.41%	26.86%	2.03	**91.76%**

Although the results obtained with the ResNet VAE present higher privacy, the images lack intelligibility and explanatory value, hindering their use as explanations. The UNET has a higher capacity to preserve features, as verified by the higher accuracy in identity recognition and glaucoma recognition. Furthermore, since the image generated by the UNET is extremely similar to the original one, this network might be performing an adversarial attack on the identity recognition network instead of adequate anonymization.

Given these results, we can conclude that the original generator with a standard convolutional VAE is the one that provides better and more balanced results, guaranteeing both privacy and the explanatory value of the images.

State-of-the-art Comparison

In this section, we compare our privacy-preserving models with the state-of-the-art methods blurring, K-Same-Select [76] and PPRL-VGAN[41]. These methods had previously been applied to the Warsaw-BioBase-Disease-Iris v2.1 dataset in [114]. The results in terms of identity recognition and glaucoma recognition are summarized in Table 4.4.

Table 4.4: Comparison between the results of our privacy-preserving models and results of state-of-the-art models obtained from [114] in regards to privacy and preservation of explanatory evidence.

Method	**Identity Recognition Accuracy**	**Glaucoma Recognition Accuracy**
Blurring	19.41%	75.59%
K-Same-Select	**0.88%**	77.06%
PPRL-VGAN	1.76%	86.56%
Ours (Multi-class)	**0.88%**	**91.47%**
Ours (Siamese)	1.76%	**91.47%**

Our privacy-preserving models have a higher capacity to preserve explanatory features than the methods from the literature while obtaining comparable results in identity recognition. Furthermore, our models promote privacy for every patient in the dataset, unlike K-Same-Select and PPRL-VGAN, which directly use identities from the dataset in the privatization process (through image averaging or identity replacement). As such, our privacy-preserving models are the most appropriate to be applied to the domain of medical case-based explanations.

4.6 Summary and Conclusions

In these works, we explored a state-of-the-art privacy-preserving model in the context of case-based explainability, and starting from this model we created a new privacy-preserving approach specifically developed to anonymize case-based explanations. The new approach tackles the most significant weaknesses of current privacy-preserving models, guaranteeing privacy, intelligibility, and preservation of explanatory evidence. At first, we used a multi-class identity recognition

model to guide image privatization. Then, we widened the range of application of our model by using a Siamese identity recognition network to guide the privatization, enabling the model to be used when medical data only has a small number of images per subject.

Our approach regarding the preservation of explanatory evidence consisted of using interpretability saliency maps to reconstruct relevant features. However, *post hoc* techniques are often criticized for not reflecting a model's real reasoning [100]. As such, using these methods to preserve explanatory features when anonymizing explanations obtained through intrinsic interpretability methods clashes with the intrinsic methods' goal of providing accurate representations of a model's reasoning. In such cases, if the intrinsic interpretability method defines a similarity measure to semantically compare two images, it should be possible to use this measure to approximate the privatized image to the original image in regard to explanatory features.

We have also applied the model to generate counterfactual explanations based on privatized factual explanations. The counterfactual explanations highlight the changes in an image that would lead to a reversal of the class prediction. We used interpretability saliency maps to promote changes in image regions related to the classification task. Nonetheless, the resulting explanations may be adversarial examples whose alterations are not related to the concepts associated with the classification task. Even though we only considered a binary classification task in our work, the approach is generalizable to the multi-class scenario. To apply the counterfactual generation model to multi-class classification problems, the counterfactual decoder could be trained to receive the latent representation of an image and the target class of the counterfactual, allowing to retrieve counterfactual explanations representative of each class.

Future work should consider integrating privacy in the image retrieval process to optimize the selection of explanatory cases and using causality to ensure that features preserved in the privacy-preserving explanations are causally related to the explanatory task.

In conclusion, this line of work contributes to enabling the use of case-based explanations in contexts where the data violates the privacy of individuals, as is the case in most medical imaging applications, including the aesthetic evaluation of breast cancer treatments.

Part III

Aesthetic Evaluation of Breast Cancer Treatments

Chapter 5

Aesthetic Evaluation of Breast Cancer Treatment Outcomes

Foreword on Author Contributions

Some parts of this chapter were originally published in or adapted from:

- [27] J. S. Cardoso, W. Silva, and M. J. Cardoso, "Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment," *The Breast*, 49, 123-130, 2020. doi:10.1016/j.breast.2019.11.006

5.1 Context

According to Globocan, 2018 witnessed about 2.1 million new breast cancers (BC), accounting for almost 1 in 4 cancer cases among women [16]. BC is the most frequently diagnosed cancer in the majority of the countries worldwide and is also the leading cause of cancer death in over 100 countries. However, BC is an increasingly treatable disease, and 10-year survival now exceeds 80% in most high-income countries. Given this high rate, survivorship issues have become a critical concern, especially the ones with an impact on long-lasting patient Quality of Life (QoL). The locoregional (LR) treatments for BC (surgery and radiation therapy) are undertaken by the majority of BC patients and usually have a significant impact on body image. In case of a poor aesthetic result, women will have to live with the potential disfiguring aesthetic consequences of their LR intervention. Both treatments, surgery and radiotherapy, can individually impact the aesthetic outcome and, when combined, there is usually an added effect that will eventually worsen the final aesthetic outcome. In general, it is estimated that 30% of all women submitted to LR treatment have a fair/poor aesthetic outcome with the consequent negative impact on psychosocial recovery and QoL [84].

Concerning LR treatments - besides the oncological criteria (re-interventions, recurrences), there are no standard available tools to evaluate the quality/impact regarding the aesthetic outcome. Nowadays, due to better screening and optimized treatments, locoregional recurrences and reinterventions have almost universally attained the optimal goals, but aesthetic results need also

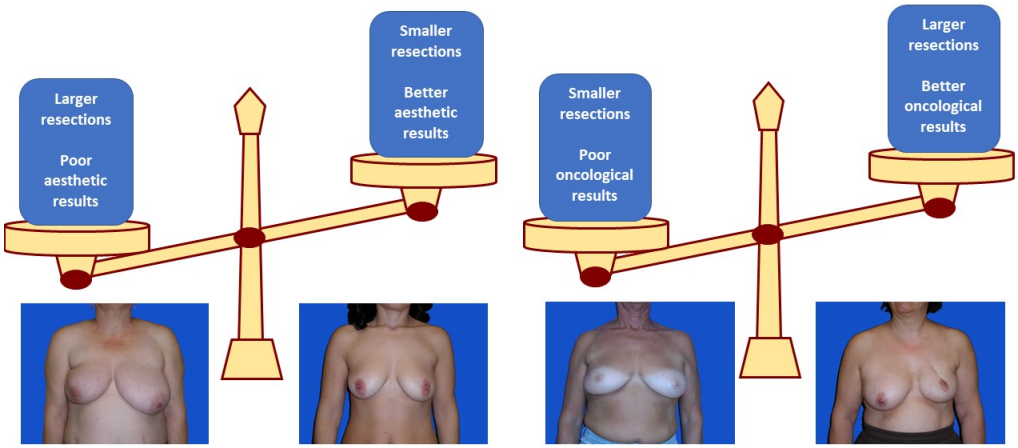


Figure 5.1: The balance between optimal oncological and aesthetic outcomes.

to be improved as a way to guarantee a better QoL and also as a standard measure to allow Breast Units to audit outcomes and improve LR approaches whenever needed.

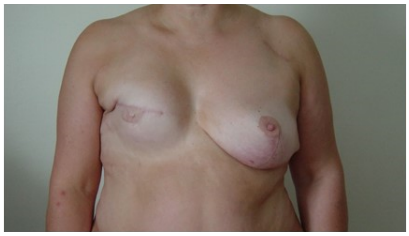
As an example, classic breast conservation surgery, according to EUSOMA, should represent 70-80% of surgeries for small breast cancers until 3 cm, and reinterventions in these cases should not exceed 10% [14; 136]. To be granted certification, all European accredited Breast Units must fulfill these mandatory criteria. However, it is easily understood that these values are more frequently attained if larger resections are undertaken, with a subsequent higher negative impact on aesthetic outcome, as asymmetry will be a detrimental feature in these cases. On the other hand, a more limited approach can be at risk of having a higher reintervention rate due to closer margins but will most probably result in better aesthetic outcomes. This fundamental balance between a cleaner resection and a better aesthetic result is a key issue and still very difficult to evaluate due to the lack of a reliable evaluation method (Figure 5.1).

Many methodologies have been proposed and studied for the purpose of aesthetic evaluation, which are mainly patient-based [46; 96], expert-based [175], and the so-called objective protocols [37; 81]. However, none of the methods is recognised as a gold standard.

This research topic has now reached a turning point that deserves to be addressed and discussed. Researchers have recently started to explore diverse deep learning methodologies, which bring significant improvements in robustness, but also raise new challenges regarding data availability. At this turning point, this section aims to showcase the evolution and current landscape of methods for aesthetic quantification of the LR treatment for BC. After presenting the most significant advances in aesthetic assessment in the literature, we present some of the ideas that were later used to develop the models to be presented in some of the next chapters (namely, Chapters 6, and 7). We also discuss the most relevant challenges and promising future opportunities regarding research.

5.2 Related Work

Intuitively, self-assessment through Patient Reported Outcome Measures (PROMs) should be the most valued form of evaluation of aesthetic outcome; unfortunately, in spite of being a valuable measure of patient satisfaction, and hence very important, it has very low reproducibility values when compared to other evaluation methods due to the lack of knowledge patients have about how they are expected to look by the end of treatment and also due to personal factors that can have an impact on this evaluation [32]. Although there is an undeniable truth residing in the fact that the most important outcome should be evaluated by the patient herself, to use this type of evaluation that is inevitably biased would never allow a true evaluation of results, making any analysis and eventual quality control virtually impossible (Figure 5.2).



(a) Self-evaluation: **Excellent**.

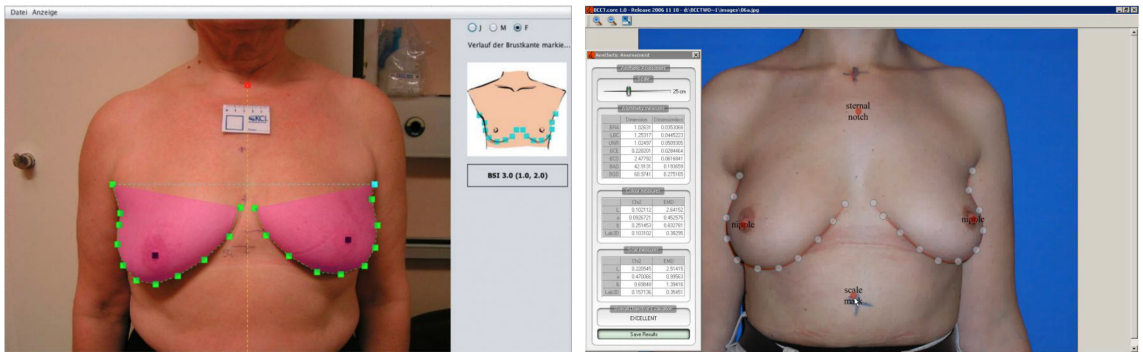
(b) Self-evaluation: **Poor**.

Figure 5.2: The variability of patient self-evaluation.

Trying to overcome the self-assessment problems, the introduction of aesthetic evaluation by experts, through patients' photographs, has also been frequently used [12; 129], especially when a new technique of LR treatment needs to be evaluated. Although a step forward, and frequently used as the gold-standard evaluation, in the absence of a better one, it is still very time-consuming, expensive and also presents low to medium reproducibility values [28].

The introduction of objective methods was started by Pezner *et al.* [128], in 1985, with the first objective measure to evaluate asymmetry, one of the most important aspects of aesthetics: Breast Retraction Assessment (BRA). This line of thought was followed by other authors, who contributed with new measurements to value mainly asymmetry: Limbergen *et al.* [102] proposed two new asymmetry measurements, the Lower Breast Contour (LBC) and the Upward Nipple Retraction (UNR); Tsouskas and Fentiman [170] described the Breast Compliance Evaluation (BCE), which is the difference between the distance from the nipple to the Infra-Mammary Fold (IMF). Although all these works defined measures that could be "objectively" computed from the photograph of the patient, the measures were manually extracted (and therefore not completely independent from the user). Moreover, the measures were related to the asymmetry impact on the aesthetic result, leaving all other factors outside the analysis.

The last decade of the twentieth century also witnessed the introduction of attempts to enrich the collected data with multiple views of the patient [124], special cameras - telecameras [139], 3D scanners [123] and Moiré topographic [119] in order to facilitate the aesthetic evaluation. However, the benefits of using more information were offset by the cost and the complexity of



(a) BAT (b) BCCT.core

Figure 5.3: Screenshots of the first two computer programs for the aesthetic assessment of LR.

the data acquisition process, leaving the single frontal image as the preferred acquisition protocol for the aesthetic assessment [30; 34]. Nevertheless, 3D imaging can provide a stronger starting point than conventional photography for physics-based models that may be beneficial for surgical planning and supporting physician-patient communication [58].

The first decade of the twenty-first century witnessed the softwarization of the solutions [29; 64]. In particular, the first two computer programs developed specifically for the LR aesthetic assessment were presented almost simultaneously, BCCT.core and BAT (Figure 5.3).

The development of BCCT.core also brought with it the adoption of machine learning methodologies to integrate disparate measures into a global assessment of the aesthetic result [23]. In such (which are now thought of as shallow) machine learning approaches, engineers do not need to be concerned with constructing precise and exact rules to combine the multiple measures. Instead, they focus on statistical models or simple neural networks as an underlying engine and then automatically learn or 'tune' the parameters of the engine using the past data to make them handle uncertainty and generalize well for yet to be seen patients. These approaches require historical data, and a rich set of photographs of patients, each individually evaluated by (a panel of) expert(s). These influential works adopted a Delphi Panel procedure to reach a consensus and guide the machine learning process [28].

Still, in the first decade of the 2000s, another line of efforts tackled the automation of the detection of fiducial points in the photograph to support the computation of the measures, alleviating the dependency on the user to achieve the overall assessment. The most relevant anatomical landmarks include the nipples, breast contours (with particular emphasis in the endpoints) and incisura jugularis [22; 99; 122; 161]. Although full automation was not achieved, the process was much less dependent on the user, often requiring only minor corrections. Once these marks were in place, the process flows transparently, with the computation of several measures and their combination in the overall assessment.

It is also worth emphasizing a change that simplified the acquisition process. Initial measures, like BRA and LBC, are quantities to which a physical dimension is assigned with a corresponding unit of measurement. To be properly recorded, they require a known scale to be present, enabling

the conversion from pixels in the digital photograph (or units in the analogue photograph) to the true physical dimension. Thus, dimensionless quantities, which are based on ratios, were introduced to dismiss the need for the scale and therefore simplify the evaluation process [23].

Original measure	Dimensionless measure
BRA	$\text{pBRA} = \frac{BRA}{0.5\left(\sqrt{X_1^2+Y_1^2}+\sqrt{X_2^2+Y_2^2}\right)}$
LBC	$\text{pLBC} = \frac{LBC}{0.5(Y_1+NI_1+Y_2+NI_2)}$
BCE	$\text{pBCE} = \frac{BCE}{0.5(NI_1+NI_2)}$
UNR	$\text{pUNR} = \frac{UNR}{0.5(Y_1+Y_2)}$

Table 5.1: Examples of the dimensionless asymmetry measures [23]. (X_1, Y_1) and (X_2, Y_2) are the coordinates of both nipples (using the sternal notch as the centre of coordinates); NI_1 and NI_2 are the nipple to infra-mammary fold distances.

An additional bonus related to these dimensionless quantities was that they were defined to dismiss the need to know which was the treated breast, further facilitating the full automation of the process, see Table 5.1.

5.3 From handcrafted to deep-learning-based methodologies

The recent artificial intelligence (AI) breakthroughs achieved with deep learning are also reaching this field. The machine learning traditional workflow is typically based on extracting predesigned features (also referred to as handcrafted or engineered features) from the patient photographs. For instance, BCCT.core algorithm relied on handcrafted measures like BRA to estimate the overall aesthetics. The algorithm did not learn that BRA was indeed useful for the aesthetics' evaluation, it was predesigned by an expert. This feature engineering is a bottleneck requiring significant human expertise. By limiting the model to the use of handcraft features, one may be missing the integration of information not captured by the handcrafted features but still relevant for the aesthetic outcome.

Without any preconception about how to construct features relevant to the aesthetic evaluation, deep learning breaks away the aforementioned difficulties by the use of a deep, layered model structure, often in the form of neural networks, and the associated end-to-end learning algorithms. With deep learning, the feature extraction and analysis parts are fully coupled. The algorithm learns, directly from the image, to compute features and use those features in the analysis of the aesthetic result. Feature construction and prediction are now unified in a single process.

In Chapter 6, we propose and present deep and hybrid models, aiming to replace the current state-of-the-art approach to automatically detect the keypoints relevant for the computation of the asymmetry measures mentioned before. In Chapter 7, we present a deep learning approach aiming to improve the overall aesthetic evaluation.

Even though deep learning may help in the automation required for both keypoint detection and aesthetic evaluation of breast cancer treatments, it may suffer from a lack of transparency

that could hinder its use in clinical practice. Nonetheless, we can explore the methods we developed and presented in Part II of this document. For instance, the methods presented in Chapter 2 [150; 153] were already evaluated in the context of the automatic aesthetic assessment, promising accountability for future certified Breast Units (and also with applications in teaching, etc.).

5.4 Discussion and Future Work

There is a definite conviction that objective methods will play a central role in assessing breast surgery procedures and in future Breast Units. The recent recommendations are pointing in that direction [31] and the traction that BCCT.core is gaining in the research community [33; 108] support this view. Although several challenges impede bringing aesthetic evaluation into daily clinical practice now, it is expected to be a critical component in future BC management workflows.

Some needs are likely to be satisfied in the near future as they are the corollary of current efforts. The full automation of the process with high accuracy will bring efficiency and user independence to the process. The current deep based approaches, supported in large sets of past data, are likely to attain this goal. Although the full integration in current Hospital Information Systems is not foreseen in the near future, the deployment of the tools of aesthetic assessment as web applications or cloud services may facilitate the widespread adoption. The softwarization in the 2000s was an important landmark, with software developed specifically for the task, but it provided only desktop applications. While a desktop application must be installed on the computer before it can run, web applications allow us to access it on demand by using a web browser, not requiring installed software. Furthermore, they can be accessed on any device with internet connection, providing maximum accessibility with minimal system requirements. Web development tools are nowadays much more mature than 10 years ago, making the engineering task a lesser effort. Although concerns with security and privacy in e-health cloud-based systems are justified, solutions exist to properly secure health data in the cloud [4; 160].

Despite all the progress in the field, there is still a lack of big data and extensive international studies. For example, Esteva *et al.* [59] have developed a deep neural network for skin lesions classification using a dataset of 129450 clinical images. Efforts of similar dimension can be found in other medical domains. It is fundamental to set up and populate a sizeable interoperable repository of photographs of breast cancer LR treated patients, enabling the development, testing, and validation of AI-based solutions to improve aesthetic evaluation and overall quality of life follow-up. Following current trends, the platform should promote access to anonymized image data sets to be made more openly reusable across the globe for developing AI solutions. The Dialogue on Reverse Engineering Assessment and Methods (DREAM) initiative recently hosted an open crowd-sourced Digital Mammography (DM) DREAM challenge[1] to foster the development of algorithms for the detection of cancer in screening mammography, and to objectively determine

[1] https://www.synapse.org/Digital_Mammography_DREAM_Challenge

by blind evaluation whether machine learning methods applied to data from mammography exams can improve screening accuracy. Similar endeavours in the aesthetic evaluation could have a massive impact in the field.

While the focus has been on the aesthetic outcome evaluation for breast conserving surgery, similar concerns about aesthetic assessment exist for related populations, such as the minority of women who require total mastectomy and may desire reconstruction. Additionally, new surgical techniques, as well as new forms of delivering loco-regional radiation therapy, remain unexplored. Due to the introduction of plastic surgery techniques into the BC surgery arena, a large number of new surgeries have been generalized without a proper evaluation tool. There is an almost absolute lack of knowledge of the aesthetic outcome of these surgeries as well as its impact on patients' QoL. At almost the same time, new radiation therapy techniques were also introduced, and again, there is a heterogeneous evaluation of their impact, either per se or associated with the different surgical operations previously referred. The correlation of this aesthetic outcome with patients' QoL is also not standardized and very difficult to evaluate due to the vast diversity of available tools. Although the different surgical procedures and radiation therapy techniques require specific models to evaluate aesthetics, they all share properties that can be explored to improve the model design. In the machine learning community, transfer and multitask learning focuses on building better predictive models by exploiting knowledge gained in previous/related tasks, which allows for the softening of the traditional supervised learning assumption of having identical train-test distributions [60]. These methodologies can help us to build on top of the present efforts for LR surgeries, adapting the models for the aesthetic evaluation to the vast offer of BC treatments.

Finally, this journey towards more objective methods excluded patient self-assessment from the evaluation process, leading to a division in the community. It is still not totally understood why patients' evaluation is different from expert evaluation and objective evaluation of results. A first reason is related to the fact that aesthetic evaluation is dependent on many individual (psychological, physical, social, and cultural) factors that can impact on how the patients see the results. The other reason, less explored, has to do with the patients' usual absence of knowledge of the resulting outcome. Confronted with the fear of cancer and the fear of losing the breast, they usually tend to be more benevolent with worse aesthetic results and evaluate themselves better. It seems fundamental to unite these two perspectives of QoL, researching methods for the evaluation of the aesthetic outcome of BC treatments integrating objective methods and significant factors derived from patient input.

While the European Union is fighting the battle of Breast Units certification, creating quality indicators mainly related to overall survival (OS) and breast cancer specific survival (BCSS), it seems incongruent not to have a proper evaluation of these important parameters of outcome that will allow comparison of results between Breast Units and correction of factors responsible for worst results and will also allow patients to fight for better care by choosing units with more consistent and favourable results. In the European Society of Mastology (EUSOMA) quality indicators review of 2017, there is still no reference to any form of aesthetic evaluation of results. This is possibly the consequence of none of the discussed methods being recognised as a gold standard,

and the European Organisation for Research and Treatment of Cancer (EORTC) still advises a combination of methods for this evaluation (self-evaluation + subjective panel assessment + objective measurements). Unfortunately, with the current number of incident BC cases, this is neither practical nor feasible. The current rise of AI, propelled by the new paradigm of deep-structured machine learning or deep learning, seems to offer the tools that were missing to achieve the accuracy and automation to convince the community. The striking successes in speech recognition, computer vision, and machine translation – completely taken over by the deep-learning paradigm – give us a set of reasons to approach the task with confidence.

Chapter 6

Keypoint Detection for the Aesthetic Assessment of Breast Cancer Treatment Outcomes

Foreword on Author Contributions

The results of this work have been disseminated in the form of two papers in international conferences, an international journal article and a master's thesis:

- [152] W. Silva, E. Castro, M. J. Cardoso, F. Fitzal and J. S. Cardoso, "Deep Keypoint Detection for the Aesthetic Evaluation of Breast Cancer Surgery Outcomes," in *16th IEEE International Symposium on Biomedical Imaging (ISBI 2019)*, Apr. 2019. doi:10.1109/ISBI.2019.8759331
- [70] T. Gonçalves, W. Silva, and J. S. Cardoso, "Deep Aesthetic Assessment of Breast Cancer Surgery Outcomes," in *XV Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON 2019)*, Sep. 2019. doi:10.1007/978-3-030-31635-8_236
- [71] T. Gonçalves, W. Silva, M. J. Cardoso, and J. S. Cardoso, "A novel approach to keypoint detection for the aesthetic evaluation of breast cancer surgery outcomes," *Health and Technology*, 10 (4), 891-903, 2020. doi:10.1007/s12553-020-00423-8
- [69] T. Gonçalves, "Deep Aesthetic Assessment of Breast Cancer Surgery Outcomes," Master's thesis, Universidade do Porto, Portugal, 2019.

The Master's thesis [69] was supervised by Jaime S. Cardoso and co-supervised by Wilson Silva. The work presented in [69–71] was motivated by the work developed in [152]. My main contributions in these three works [69–71] consisted of the original research ideas and conceptualization of the work.

6.1 Context and Motivation

As discussed in Chapter 5, the subjective assessment of breast cancer treatments poses reproducibility issues. In order to overcome these issues, some objective methods for the assessment of BCCT were introduced, in which the systems developed by Fitzal *et al.* [64] and by Cardoso and Cardoso [23] represent the most relevant works. However, none of the aforementioned works is completely automatic (they require manual annotation of some keypoints), they all apply to only

the classic conservative treatment (leaving out the new surgical techniques) and they have limited performance. Hence, none of them was selected as the gold standard.

Even though [23] also relies on color differences, both referred works [23; 64] point out the relevance of symmetry measurements in the aesthetic assessment, which makes the correct detection of keypoints fundamental. Thus, a first step towards achieving the goal of a completely automatic and objective framework, capable of being selected as a gold standard, is the successful detection of fiducial points. In this sense, the work presented in this chapter focus on the task of detecting keypoints in photographs of women after being subjected to BCCT. Three different approaches were studied, combining different deep learning perspectives of the keypoint detection problem, and combining deep learning and more traditional techniques. All three approaches led to better results than those obtained with the previous state-of-the-art. The deep learning approaches also represent an improvement in terms of inference computation time, which is of extreme relevance when aiming to have these systems deployed in a web application, where they have to run quickly.

6.2 Related Work

Cardoso and Cardoso [22] were the first ones to present an algorithm capable of automatically detecting breast contours in digital photographs. To do this, they compute the gradient of the image and model it as a weighted graph based on pixel gradient, value and position. Assuming that the two endpoints of the breast contour are known, the problem is focused on finding the shortest path between both endpoints that goes through the breast contour. Later on, to help on this task, Sousa *et al.* introduced the use of shape priors to facilitate the process [161]. Following the work on breast contour detection, Cardoso *et al.* proposed a method for the automatic detection of the endpoints [35]. This method assumes that the photo only contains the torso of the patient, which means that the external endpoint of the breast contour can be assumed to be at the point of the body where the arm contour intersects the trunk contour. However, in most of the photographs, patients are in the arms-down position, so the arm's contour is almost indistinguishable from the trunk's contour. This means that the external endpoint of the breast can be defined as the highest point of the breast contour. Figures 6.1 to 6.8 present a graphical sequence of the algorithm proposed by Cardoso *et al.* in [35]. Breast surface is generally characterized as a featureless shape. So, the nipple should be the most prominent feature on it. Taking this into account, Cardoso *et al.* proposed a method [26] that uses a Harris corner descriptor to detect possible nipple locations and then applies a closed contour method to find areola contours around those points (see Fig. 6.9 and Fig. 6.10). High-level features (Harris corner quality factor, the average magnitude of the directional derivative of the contour, shape factor of the contour, equivalent diameter of the contour) are extracted and the best pair candidate/closed contour is selected by a support vector machine (SVM) classifier trained on the extracted features.

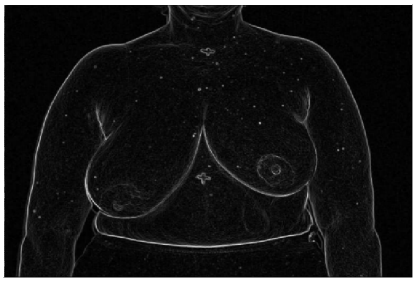


Figure 6.1: Image gradient, from [35].

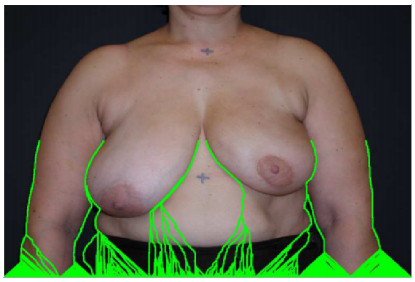


Figure 6.2: Shortest paths from the bottom to the middle.

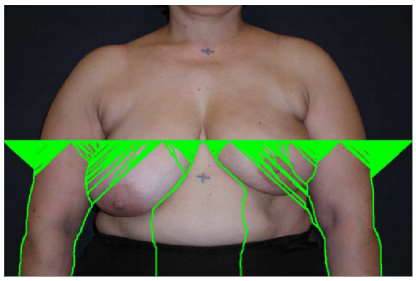


Figure 6.3: Shortest paths from the middle to the bottom.

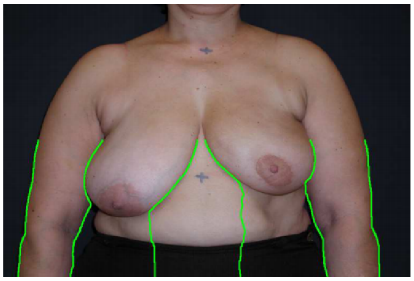


Figure 6.4: Strong paths between middle and bottom.

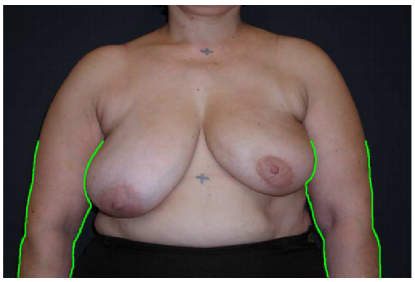


Figure 6.5: Selected shortest paths.

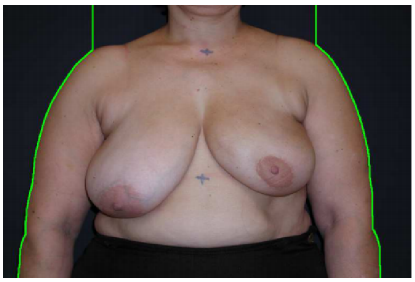


Figure 6.6: Strong paths between top and bottom.

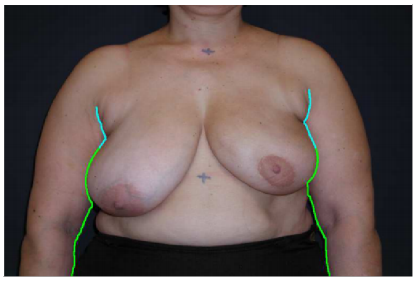


Figure 6.7: The endpoints are the highest points of the shortest path.

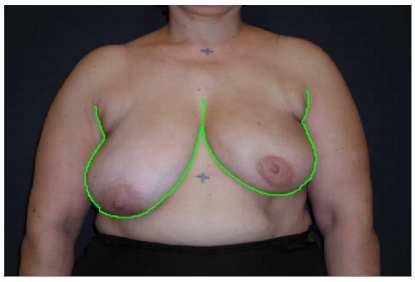


Figure 6.8: Automatic breast contour detection with the shortest path algorithm.

6.3 Methodology

6.3.1 Deep Keypoint Detection Algorithm

Deep Neural Networks (DNN) offer a valuable framework to achieve this integrated learning. However, as in biomedical applications we usually deal with small datasets, DNN tend to overfit

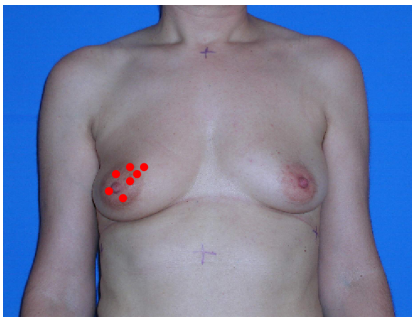


Figure 6.9: Nipple candidates detection [26].

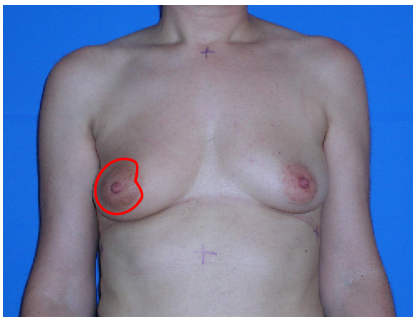


Figure 6.10: Harris corner detection. [26].

severely. There are some approaches that allow to mitigate the effect of overfitting, for example, transfer learning and learning an intermediate representation. This latter idea was explored in other domains in the works done by Belagiannis *et al.* [13] and Cao *et al.* [21]. In both works, an intermediate representation consisting of confidence maps in relation to the location of the keypoints was created. An additional interesting idea also explored in those works was an iterative process of refinement.

Based on the ideas previously mentioned, we have built a DNN to automatically detect keypoints in photographs of patients after being subjected to BCCT (see Figure 6.11). As shown in Figure 6.11, the architecture of the proposed DNN comprises two main modules: regression and refinement of heatmaps, and regression of keypoints.

The first module is what we call Heatmap Regression and Refinement. Here the goal is to generate an intermediate representation consisting of a fuzzy localization for the keypoints we want to detect. This is done in order to help the regularization process of the DNN. Heatmaps are obtained using a fully convolutional neural network. In this work, we used the well-known segmentation model, U-Net [135]. Figure 6.12b presents an example with an image from the dataset and the respective ground truth heatmap super-imposed.

The second module has as input the multiplication of the image with the refined output of the previous module ($Output1^{(n)}$). The regression of the keypoints is composed of three blocks: a convolutional backbone (here, the convolutional blocks of VGG16), some additional convolutional layers and, finally, the classifier composed of fully-connected layers. The first block, VGG16 [157], is pre-trained with ImageNet and then fine-tuned in our dataset. After VGG16, four convolutional layers are added to further increase image processing and decrease the size of feature maps before the dense layers. Finally, three dense layers are used to regress the 74 coordinates, corresponding to the keypoints that make up the breast contour, endpoints, nipples and supra-sternal notch (Figure 6.12a).

The proposed fully supervised learning scheme requires not only the ground truth for the keypoints but also a ground truth for the heatmaps, which is created considering a Gaussian centered at each keypoint and with a pre-defined standard deviation.

Regarding the learning process, we have two different terms in the loss function: heatmap regression, which works here as a regularization term, and keypoints regression, our goal. Thus,

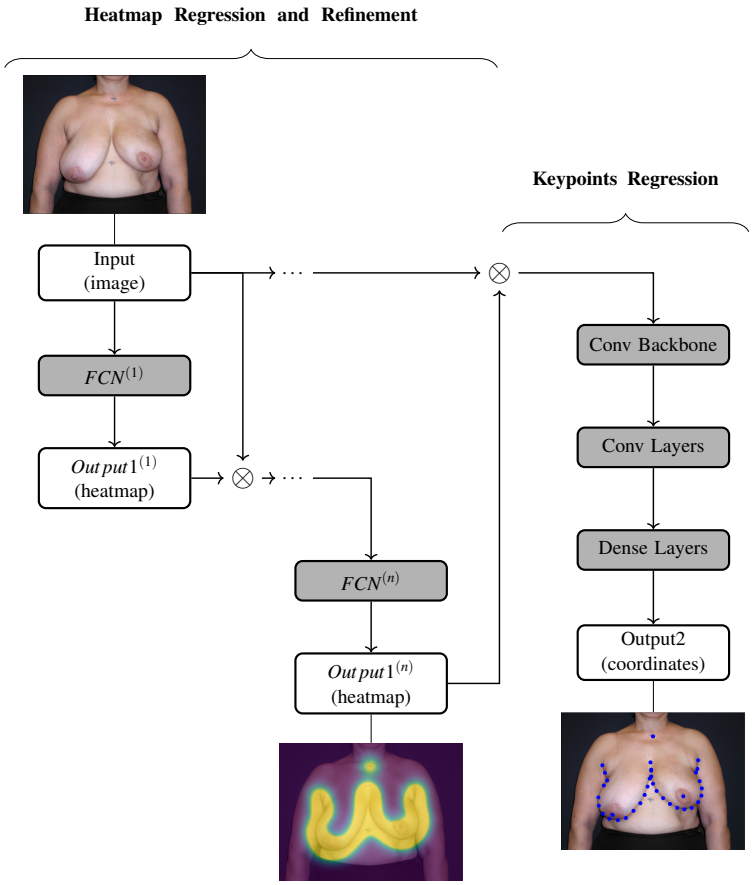
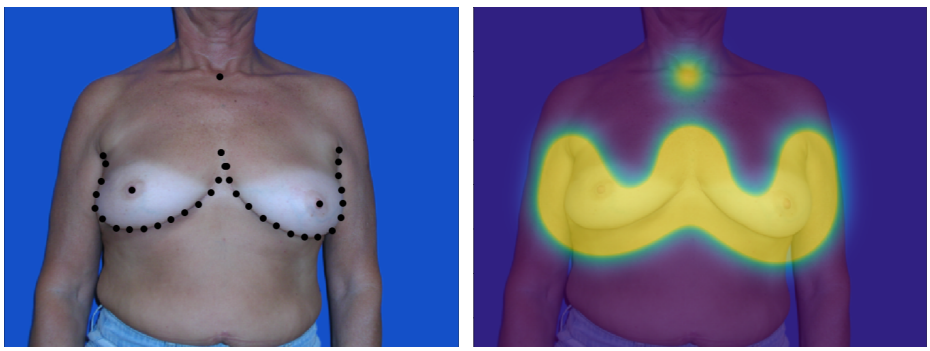


Figure 6.11: Proposed iterative DNN architecture for keypoint detection. FCN stands for fully-convolutional network. Conv Backbone stands for convolutional backbone.



(a) Keypoints (b) Heatmap

Figure 6.12: Example of Ground Truth

the loss function is a linear combination between these two terms, with λ_h and λ_k weighting the importance of each term (Eq. 6.1).

$$\mathscr{L} = \lambda_h \mathscr{L}_{heatmaps} + \lambda_k \mathscr{L}_{keypoints} \tag{6.1}$$

In relation to the regression of the keypoints, the mean squared error (MSE) was the loss

function selected (Eq. 6.2). N_k is the number of coordinates, x_k^{target} the ground truth for a single coordinate and $\hat{x}_k$ the prediction.

$$\mathscr{L}_{keypoints} = \frac{1}{N_k} \sum_{\forall k} (x_k^{target} - \hat{x}_k)^2 \tag{6.2}$$

The heatmaps were also learnt using MSE. However, the heatmaps undergo an iterative process of refinement. Thus, the complete process is defined by Eq. 6.3, where j represents a step in the refinement process and λ_j represents the weight given to that step.

$$\mathscr{L}_{heatmaps} = \sum_{j=1}^{N_h} \lambda_j \mathscr{L}_{heatmap}(j) \tag{6.3}$$

Finally, the loss for the heatmap in each step is defined as follows

$$\mathscr{L}_{heatmap}(j) = \frac{1}{N_p} \sum_{\forall p} (x_p^{target} - \hat{x}_p)^2 \tag{6.4}$$

where N_p corresponds to the number of pixels in the image, and x_p^{target} and $\hat{x}_p$ to the ground truth and prediction for the pixel values, respectively.

6.3.2 Hybrid Keypoint Detection Algorithm

Besides the fully deep learning approach, we also explored the use of a hybrid (deep + traditional) approach to the detection of keypoints (see Fig. 6.13). With this hybrid approach, the endpoints and the nipples are obtained with the use of the previously presented deep learning algorithm, whereas the contour is detected with the conventional shortest path algorithm, which was described in section 6.2 (Related Work). The main difference here is that the endpoints given as input to the shortest path algorithm are obtained with the Deep Keypoint Detection Algorithm, instead of being specified by the user, or detected through traditional approaches. This hybrid algorithm led to an improvement in the results when compared with using the deep learning approach to find all keypoints.

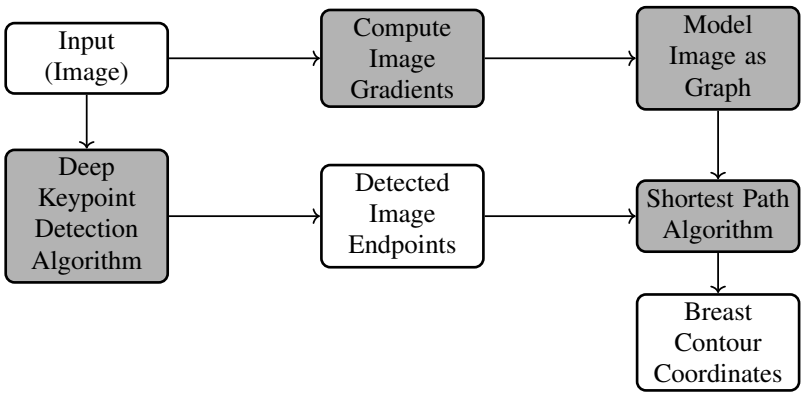


Figure 6.13: Hybrid Keypoint Detection Algorithm.

6.3.3 Keypoint detection through deep segmentation

Our third approach consists on combining image segmentation and the deep keypoint detection algorithm in a pipeline. Since the approach heavily relies on segmentation, the first step was training a deep image segmentation model to achieve good results in breast segmentation. To train this model, it was necessary to generate ground-truth breast masks, which were obtained with the support of the ground-truth keypoints and images. Using the ground-truth keypoints as delimiters of the area occupied by both breasts, the value of 255 was assigned to all pixels inside and the value of 0 to all pixels outside this area. Images were then normalized to have pixel values between 0 and 1. Taking into account previous results with other models [70], for this experiment, it was decided to use the U-Net++ architecture as the deep segmentation model. The official Keras implementation by Zhou *et al.* [1] was trained and fine-tuned (the model has an encoder which is initialized with the ImageNet [53] weights) during 300 epochs with Adadelta [184] as the optimizer. During training, binary cross-entropy was selected as the loss function. Regarding data augmentation, image and mask translations, rotations and flips were employed in an online fashion during training. These hyperparameters were chosen based on model performance. The trained U-Net++ model was then used to generate segmentation masks. From these masks, contours were extracted using the marching squares algorithm (a special case of the marching cubes algorithm [103]), implemented in scikit-image [172]. The intuition behind this contour detection step is that the detected contours will contain the breast contour keypoints, assuming that breast segmentation masks were well generated. As such, this first step outputs a variable number of contour keypoints, some of which are not desired, because they do not belong to what is considered the breast contour. As a post-processing step, the deep keypoint detection algorithm's predicted endpoints are projected onto the mask contours through the minimization of the Euclidean distance between the mask contour keypoint and the predicted keypoint. At the end of this processing step, there are new 34 breast contour keypoints plus the nipples and sternal notch, which were predicted by the deep keypoint detection algorithm alone (see Fig. 6.14). Fig. 6.17 shows a chronological scheme of this algorithm.

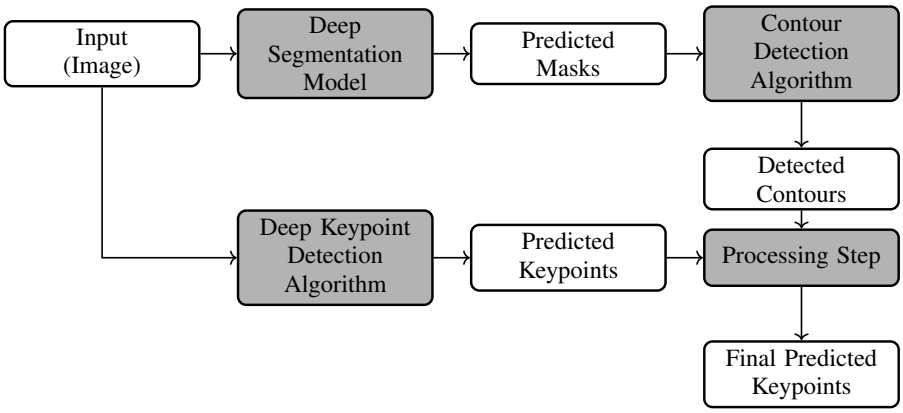


Figure 6.14: A Novel Segmentation-Based Keypoint Detection Algorithm.

[1] See: https://github.com/MrGiovanni/UNetPlusPlus

6.4 Results and Discussion

Traditional approach vs. Deep (+ hybrid) keypoint detection

To assess the performance of the proposed methods and compare them to the baseline algorithm (state-of-the-art) two datasets were considered. The first was the PORTO dataset, which is the standard dataset used in previous works. It is composed of 120 photographs of patients who had undergone BCCT. In these images, the torso of the patient is shown in front of a clean and uniform background. The second was obtained by joining three smaller sets of photographs: the 120 images of the PORTO dataset, 30 other photographs obtained in similar conditions (TSIO dataset) and 71 additional images captured in poorer lighting conditions and without the concern of having a consistent and distinct background (VIENNA dataset, see Figure 6.15). For each image, 37 ground truth points were available (4 endpoints, 30 points along the breast contours, 2 nipples and the supra-sternal notch). However, for comparison with the traditional baseline, the supra-sternal notch was not considered. For both datasets, training and test sets were obtained using 5-fold cross-validation.

For the baseline method, the trunk contour extension parameters were optimized by grid searching on the training data. For nipple detection 10 candidates were considered per breast. SVM hyper-parameters were optimized by grid-search using 3-fold cross-validation. Regarding the DNN model, hyper-parameters controlling dropout rate and strength of data augmentation were optimized using 5-fold cross validation. In relation to the number of iterations for the heatmap refinement, different numbers of iterations were tested but the one that led to the best results was 3. Weights for each sub-loss function were defined as $\lambda_{j=1} = 0.1$, $\lambda_{j=2} = 0.2$ and $\lambda_{j=3} = 0.4$. Iterative refinement of the keypoints regression module was also investigated but no promising results were obtained. The results for the first and second datasets are depicted in Table 6.1 and were computed considering the average error across the five folds. Test images resolution is variable, with the minimum resolution being (1224 $\times$ 1632 pixels) and the maximum resolution being (2592 $\times$ 3888 pixels).

Figure 6.15: VIENNA dataset examples

As shown in Table 6.1, the DNN approach led to an improvement of the results obtained in two tasks for the first dataset and all three tasks for the second. This method is also faster in inference which can be an important attribute for clinical practice. The baseline algorithm's inference time is

120 images dataset - Average error distance (pixels)									
Model	Endpoints			Breast Contour			Nipples		
	mean	std dev	max	mean	std dev	max	mean	std dev	max
Traditional model	51	47	381	16	38	287	79	124	739
DNN model	**28**	**17**	**108**	18	**6**	**39**	**61**	**31**	**147**
Hybrid model	**28**	**17**	**108**	**7**	14	116	**61**	**31**	**147**
221 images dataset - Average error distance (pixels)									
Model	Endpoints			Breast Contour			Nipples		
	mean	std dev	max	mean	std dev	max	mean	std dev	max
Traditional model	85	99	639	33	68	423	111	173	867
DNN model	**39**	**29**	**166**	21	**8**	**49**	**64**	**33**	**158**
Hybrid model	**39**	**29**	**166**	**13**	22	175	**64**	**33**	**158**

Table 6.1: Average error distance for endpoints, breast contours and nipples measured in pixels

in the order of the seconds, whereas DNN model's inference time is almost instantaneous (it takes a few milliseconds). Another important difference between the two approaches is that the error on the DNN is more regular, while in the baseline some examples are missed by a wide margin.

Noticing that the inaccuracy in the breast contour of the baseline method was mainly due to a poor estimation of the endpoints, we tested a Hybrid model which used the endpoints detected by the DNN model along with the breast contour algorithm of the baseline solution. A better mean error was obtained when compared to the other two approaches.

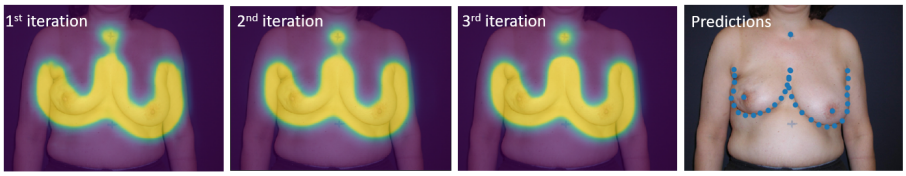


Figure 6.16: Network outputs for a test instance example

Deep keypoint detection vs. Deep segmentation approach

Table 6.2 presents the average error distance (measured in pixels) and the average execution time (measured in seconds) of each model inference on the test set. It can be seen that the Segmentation-Based Keypoint Detection Algorithm surpasses both Deep and Hybrid Keypoint Detection Algorithms from Silva *et al.* in the endpoints and breast contours detection tasks, which were, to our knowledge, the state-of-the-art breast keypoint detection algorithms. Moreover, this novel algorithm achieves lower values of standard deviation and maximum error, which suggests it is even more consistent when compared with the other two. Regarding the study of performance, it can be understood that the Deep Keypoint Detection achieves better execution time, however, it has the highest error for the breast contour. The Segmentation-Based Keypoint Detection Algorithm presents the best balance between time-efficiency and accuracy, being the most accurate model, with a time efficiency comparable to the most time-efficient method.

Table 6.2: Average error distance for endpoints, breast contours and nipples, measured in pixels and average execution time of the models' inferences (on the test set of each cross-validation fold, which has approximately 43 to 45 images). Best results are highlighted in bold. **Note:** STD stands for standard deviation and Max stands for maximum error.

Model	Endpoints			Breast Contours			Nipples			Execution Time (s)
	mean	std dev	max	mean	std dev	max	mean	std dev	max	
Deep keypoint detection algorithm	40	**33**	**182**	21	8	72	**70**	**39**	**218**	**150**
Hybrid keypoint detection algorithm	40	**33**	**182**	13	14	104	**70**	**39**	**218**	1704
Segmentation-based keypoint detection algorithm	**38**	34	195	**11**	**5**	**34**	**70**	**39**	**218**	280

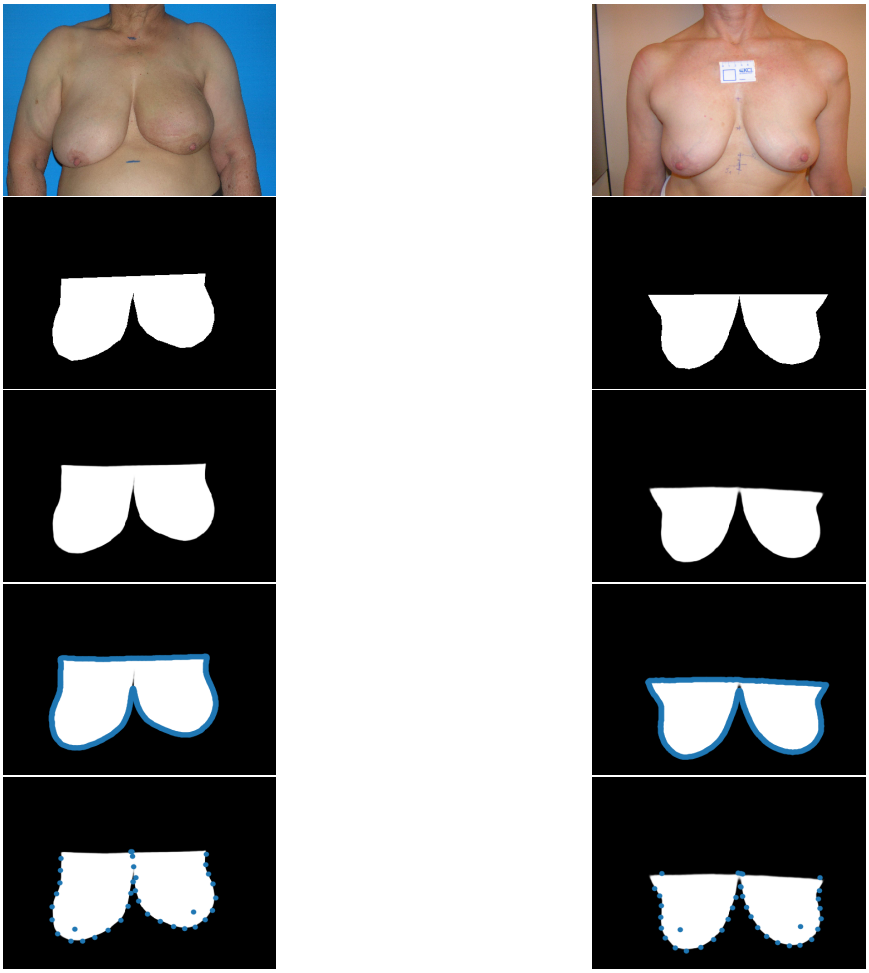


Figure 6.17: Chronological scheme (from top to bottom) of the proposed Segmentation-Based Keypoint Detection Algorithm. Each column represents a single image. The first row is the ground-truth image, the second row is the ground-truth mask, the third row is the U-Net++ predicted mask, the fourth row is the set of all the detected contour keypoints and the fifth row is the set of breast keypoints after the post-processing step.

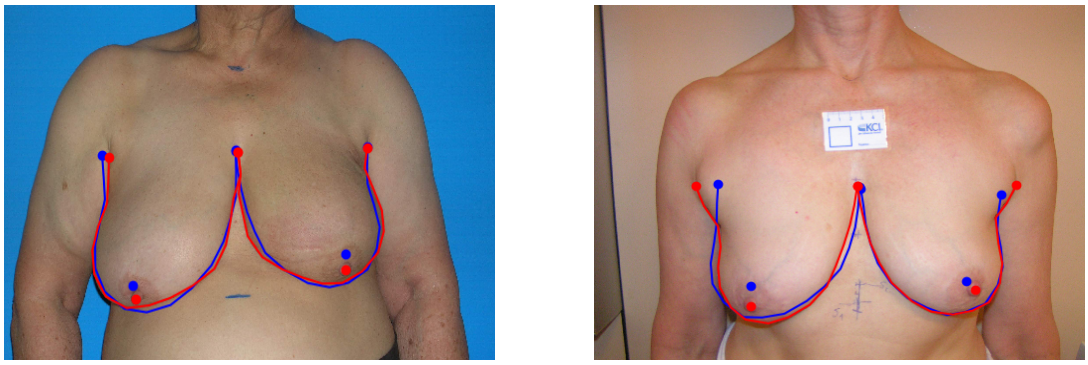


Figure 6.18: Example predictions when using the deep keypoint detection algorithm. Prediction is in blue and ground-truth is in red.

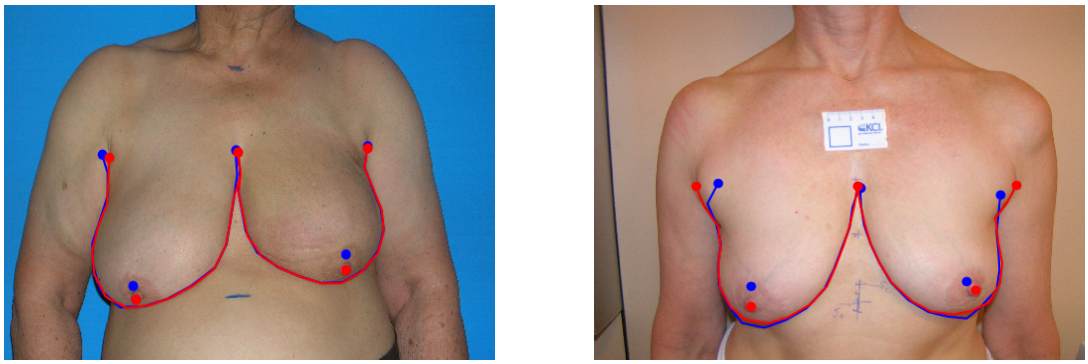


Figure 6.19: Example predictions when using the hybrid keypoint detection algorithm. Prediction is in blue and ground-truth is in red.

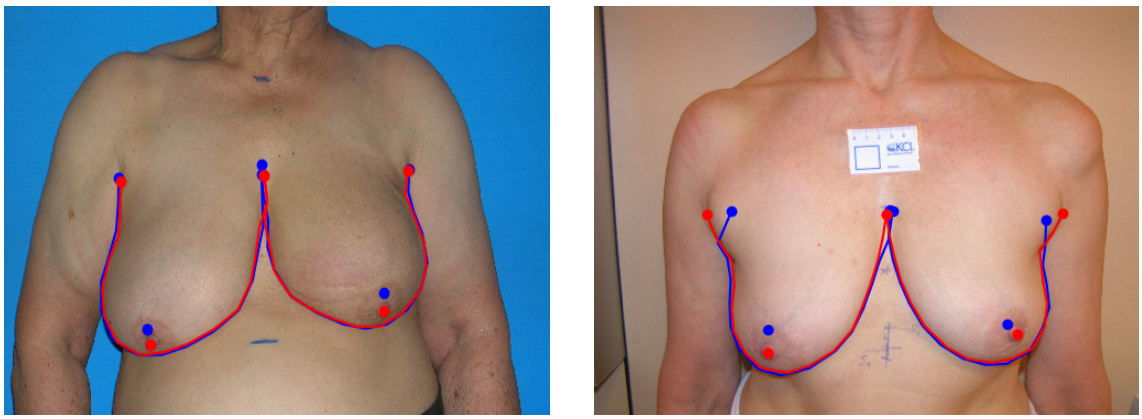


Figure 6.20: Example predictions when using the segmentation-based keypoint detection algorithm. Prediction is in blue and ground-truth is in red.

6.5 Summary and Conclusions

This work presented several novel algorithms based on deep learning approaches. Two fully deep learning based and one hybrid (combining deep learning with traditional methods). All proposed

methods surpassed the state-of-the-art algorithms present in the literature. Furthermore, a comparative study regarding algorithms performance has been done to assess which one would fit better a web-based application for the aesthetic assessment of BCCT. The deep learning models revealed themselves as the best in terms of executing time, while being very competitive in terms of keypoint prediction when compared with the hybrid approach, and especially when compared with the traditional approach. In future work, we will focus on improving results in terms of nipple detection. The integration and full deployment of one of these algorithms in a web-application is also planned, currently being in a test phase. Finally, integrating these models into a single pipeline for the aesthetic evaluation of breast cancer treatments is a natural next step.

Chapter 7

Deep Aesthetic Assessment and Retrieval of Breast Cancer Treatment Outcomes

Foreword on Author Contributions

The results of this work have been disseminated in the form of a paper in an international conference:

- [155] W. Silva, M. Carvalho, C. Mavioso, M. J. Cardoso, and J. S. Cardoso, "Deep Aesthetic Assessment and Retrieval of Breast Cancer Treatment Outcomes," in *10th Iberian Conference on Pattern Recognition and Image Analysis (IbPria 2022)*, May 2022 [Best Student Paper Award]. doi.org/10.1007/978-3-031-04881-4_9

7.1 Context and Motivation

Quality of life after breast cancer treatment is increasingly gaining more attention from the medical community, particularly with regard to aesthetic outcomes. As previously mentioned in Chapter 5, there is no accepted gold standard method for evaluating the aesthetic outcome of a treatment, with subjective evaluation by one or more observers being the most commonly used form of assessment. This subjective evaluation presents a challenge in terms of trustworthiness and reproducibility. To overcome these issues, objective methods for the assessment of BCCT were introduced. However, none was able to replace the subjective assessment completely due to requiring manual identification of fiducial points (problem tackled in Chapter 6), being only applicable to classic conservative treatments, and having limited performance [152].

Apart from the fact that there is no gold standard for the evaluation of aesthetic outcomes, patients' expectations are not properly managed, which, combined with some objective deficiencies, results in nearly 30% of patients undergoing breast cancer treatment being dissatisfied with the results obtained [125]. Therefore, it is very important that patients are aware of realistic outcomes and feel engaged with those results. To fulfil this goal, the presentation of photographs of

breast cancer treatment outcomes from patients with similar characteristics is of utmost importance. Even though the final goal is to perform this search starting from preoperative photographs, in this work, we only focus on postoperative pictures due to the lack of available matching data and preliminary work. To automatically find the most similar images, we need to develop content-based image retrieval systems [154], and, ideally, adapt them, to retrieve a new, generated image, that retains the realistic aesthetic outcome but shares the biometric characteristics of the patient requiring treatment. However, this will only be possible if we have a model for the aesthetic assessment of breast cancer treatments that can be integrated into this ideal end-to-end model for generating and retrieving realistic probable outcomes.

In this work, we propose a deep neural network that performs the aesthetic evaluation automatically and therefore does not require any manual annotation. Moreover, the network also retrieves the most similar past cases from the dataset, by searching in a highly semantic space previous to classification. We also analyse the interpretability saliency maps generated by Layer-wise Relevance Propagation (LRP) [7] to find out whether the model is robust and trustworthy.

7.2 Methodology

The state-of-the-art in the aesthetic evaluation of breast cancer treatments is the method proposed by Cardoso and Cardoso [23], which uses an SVM as the machine learning method to perform the classification. However, the SVM requires a first step involving a semi-automatic annotation of keypoints (such as breast contour and nipple positions) and a computation of asymmetry features (based on dimension and colour differences). Therefore, there is a need for human intervention. Moreover, it also can't be integrated into an end-to-end image generation model for the retrieval of biometrically-morphed probable outcomes.

Our proposed method (Fig. 7.1), which is inspired by the ideas described in [27], uses a highly regularized deep neural network to assess the aesthetic outcome automatically. It is highly regularized because the dataset dimension is very small, and the aesthetic result is very subjective. After having performed experiments with standard CNN networks, such as DenseNet-121 [86] and ResNet50 [83], we concluded that we had to considerably reduce the number of parameters to learn to overcome the intense overfitting. Thus, we designed a much simpler deep neural network, following the traditional "conv-conv-pooling" scheme, totalizing 262,908 learnable parameters (already including the fully-connected layers). However, reducing the number of parameters was not enough to prevent overfitting, leading us to the introduction of intermediate supervision in regard to the detection of important keypoints, and to the integration of pre-defined functions to translate those keypoints to asymmetry measures, namely, LBC (difference between the lower breast contours), BCE (difference between inframammary fold distances), UNR (difference between nipple levels), and BRA (breast retraction assessment). All these functions were integrated into the network by the use of "Lambda" layers (FTS Computation Functions in Fig. 7.1).

The training process was divided into two steps. First, we train the CNN model to learn to detect the keypoints (8 coordinates describing the positions of the nipples, levels of inferior breast

contour, and sternal notch). The loss function being used is the one present in Eq. 7.1, i.e., the mean-squared error of the keypoint coordinates.

$$\mathscr{L}_{model} = \mathscr{L}_{keypoints} \tag{7.1}$$

Afterwards, we train the CNN model in a multitask fashion, simultaneously optimizing keypoint detection and classification performance, which can be represented by Eq. 7.2, where λ_k and λ_c weight the different losses, $\mathscr{L}_{keypoints}$ represents the mean-squared error loss for keypoint detection, and $\mathscr{L}_{classification}$ represents the binary cross-entropy loss for classification.

$$\mathscr{L}_{model} = \lambda_k \mathscr{L}_{keypoints} + \lambda_c \mathscr{L}_{classification} \tag{7.2}$$

All images were pre-processed using the same procedure as for the ImageNet data. The model was first trained for 350 epochs (with the last fully-connected layers frozen), following the loss function presented in Eq. 7.1. Afterwards, the model was trained for 250 epochs (with the first convolutional layers frozen), following the loss function presented in Eq. 7.2. In both steps, we used the Adadelta optimizer [184], and a batch size of 16. The model at the end of the first step was the one that led to the lowest mean-squared error in the validation data. The final model was selected based on the binary classification performance in the validation data.

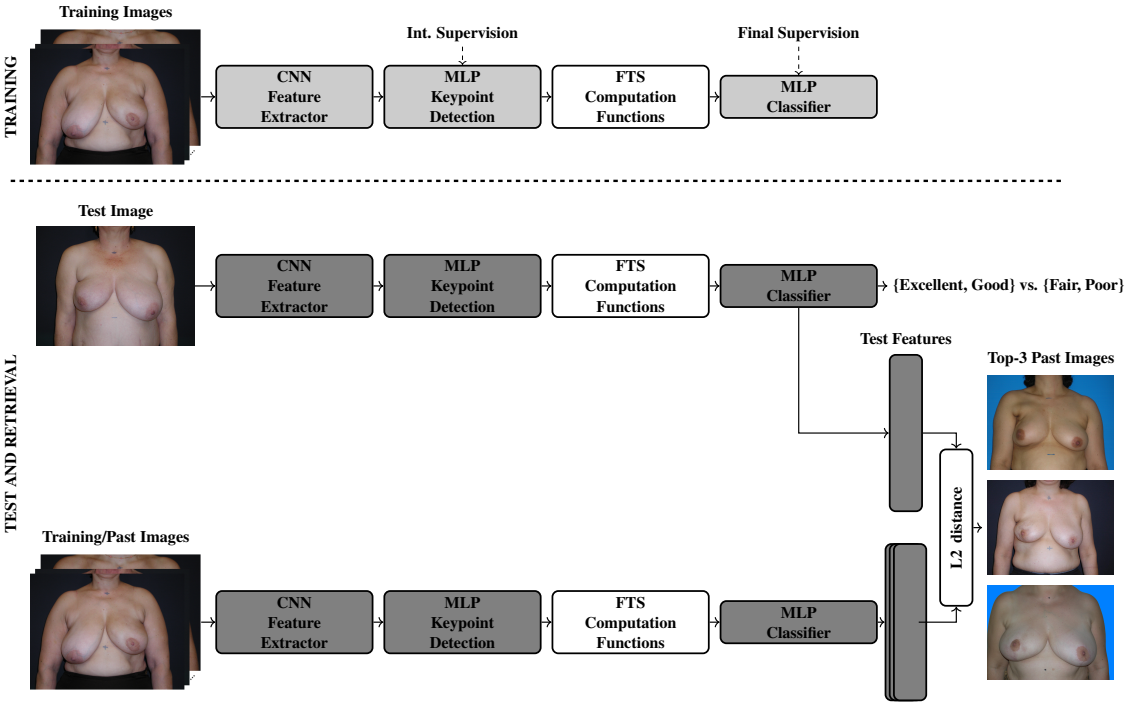
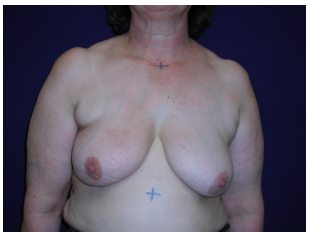


Figure 7.1: Overview of the proposed approach. Blocks in light gray mean deep neural networks are being trained (i.e., weights are being updated), whereas blocks in dark gray represent trained deep neural networks (i.e., weights are fixed). The block in white means that no weights are learnt. The "L2 distance" is computed based on the features from the previous to the last layer of the network, i.e., exactly before the classification decision.

7.3 Experimental Setup

7.3.1 Data

For the experiments, we used 143 photographs of women who had undergone breast cancer conservative treatment. These 143 images came from two previously acquired datasets (PORTO and TSIO) [152]. These images were then graded by a highly experienced breast surgeon in terms of aesthetic outcome into one of four classes: Excellent, Good, Fair, and Poor. In this work, due to the small dimension of the dataset, we only used the binary labels ({Excellent, Good} vs. {Fair, Poor}) to evaluate classification performance; this work is the first to explore the use of a deep neural network to solve this problem. Nevertheless, the original four classes were used to evaluate the quality of the retrieval. Of the original 143 images, we used 80% for training and model/hyperparameter selection, and 20% for test. To reduce the complexity of the data, the images were resized to (384 $\times$ 256), while retaining the original three-channel RGB nature.



(a) Photography from PORTO dataset

(b) Photography from TSIO dataset

Figure 7.2: Example of images used in this work.

7.3.2 Evaluation

Baseline:

As baselines we considered four SVM models, such as the one used in Cardoso and Cardoso [23]. These four SVM models resulted from variations in the inputs and kernels being used. We performed the experiments using both a linear and an RBF kernel, and giving as input to the SVM either the entire set of symmetry features, or only the four features being implicitly used by our deep neural network (i.e., LBC, BCE, UNR, and BRA).

Performance Assessment:

We considered two types of evaluation, one for classification performance using accuracy and balanced accuracy (due to the imbalanced nature of the data), and another for retrieval, where we checked whether the top-3 retrieved images belonged to the same class or to a neighbouring class (here, considering the original four classes). In addition, we generated saliency maps for the test images to understand the origin of the deep model's decisions.

7.4 Results and Discussion

In Table 7.1, we present the results in terms of accuracy and balanced accuracy for all models considered, i.e., SVM baselines and our proposed model. For all SVM models, the parameter C, which weights the trade-off between misclassifying the data and the achieved margin, was optimized using 5-fold cross-validation, following the same search space as originally used by Cardoso and Cardoso [23] (i.e., exponentially growing sequences of C: $C = 1.25^{-1}, 1.25^{0}, ..., 1.25^{30}$). The *gamma* value for the SVM with *RBF* kernel was set to 3, also as done in Cardoso and Cardoso [23]. For all SVM models, class weights were set to "balanced", meaning that the misclassifications were weighted by the inverse of the class frequency. For our proposed CNN model, we used data augmentation (horizontal flips and translations), and also weighted the misclassifications by the inverse of the class frequency.

As can be seen in Table 7.1, our proposed model outperforms all SVM models both in terms of accuracy and balanced accuracy. Only regarding the SVM models, the ones that used all asymmetry features (7 fts) were able to achieve a higher performance. When comparing SVMs with different kernels, there was only a slight improvement with the RBF kernel in terms of balanced accuracy.

Table 7.1: Results for the test set. Linear and RBF represent the SVM kernel, while 4 and 7 represent the number of symmetry features given as input to the SVM.

Model/Metrics	Accuracy ↑	Balanced Accuracy ↑
SVM (Linear, 4 fts)	0.79	0.80
SVM (Linear, 7 fts)	0.83	0.83
SVM (RBF, 4 fts)	0.79	0.82
SVM (RBF, 7 fts)	0.83	0.84
CNN (Proposed)	**0.86**	**0.89**

Besides comparing our model with the state-of-the-art, we were also interested in exploring the retrieval quality of the model. To measure that, we looked for the top-3 most similar past cases (from the training set) to the query case (from the test set). Even though the model was only trained in the binary setting ({Excellent, Good} vs. {Fair, Poor}), by observing the retrieval results, it seems the model was able to acquire a correct notion of severity.

In Fig. 7.3, we present a query example from the test set that belongs to the binary class {Excellent, Good}, and being labelled by the breast surgeon as having an Excellent aesthetic outcome. The LRP saliency map demonstrates that the algorithm is paying attention to a region of interest (breast contour and nipple), which increases trust in the model. All the top-3 images retrieved were labelled as either Excellent or Good (i.e., the same class or a neighbouring class).

In Fig. 7.4, we also present a query example from the test set belonging to the binary class {Excellent, Good}, but this time having being labelled by the breast surgeon as having a Good aesthetic outcome. The LRP saliency map presented also demonstrates the algorithm is paying

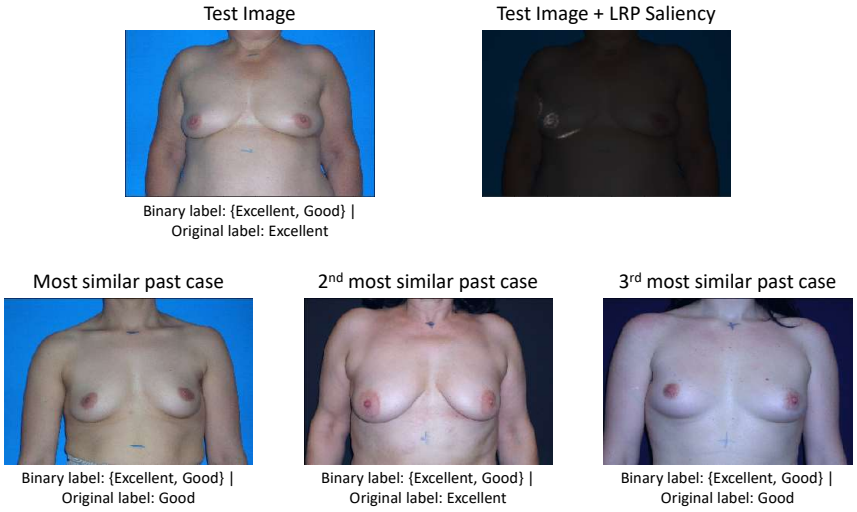


Figure 7.3: Example of query and retrieved images for an Excellent aesthetic outcome. Binary label means class belongs to set {Excellent, Good}. Original label is the ordinal label previous to binarization (Excellent, Good, Fair, or Poor). LRP saliency map is also shown for the test image.

attention to a region of interest. All the top-3 images retrieved were labelled with the same class of the query (i.e., Good).

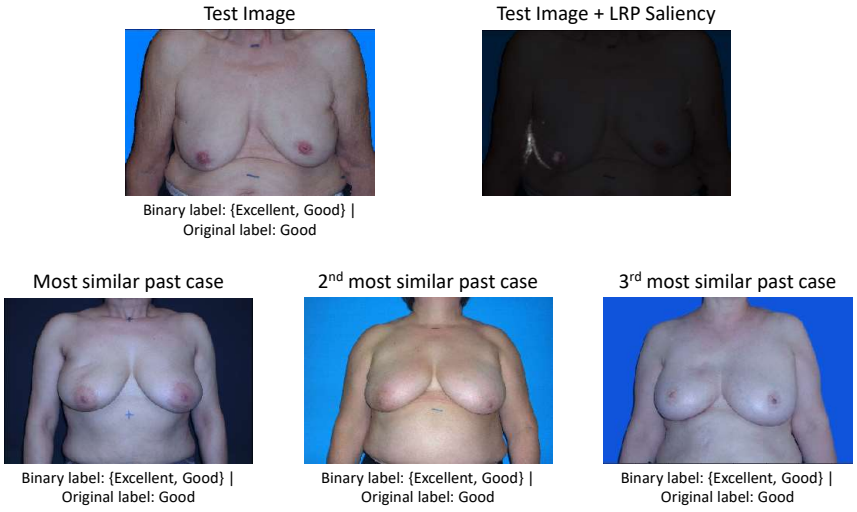


Figure 7.4: Binary label means class belongs to set {Excellent, Good}. Original label is the ordinal label previous to binarization (Excellent, Good, Fair, or Poor). LRP saliency map is also shown for the test image.

A query example from the test set belonging to the binary class {Fair, Poor}, and having been

labelled as Fair is presented in Fig. 7.5. This time, the LRP saliency map highlights the breast contour of both breasts, which makes sense as the difference between the two breasts is what is impacting more the lack of aesthetic quality in this particular case. The top-3 images retrieved were labelled as either Fair or Poor (i.e., same class or neighbouring class).

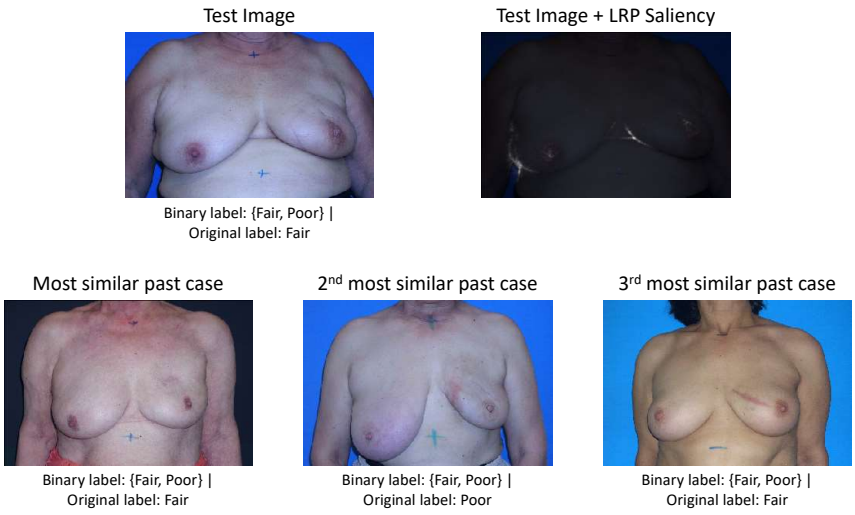


Figure 7.5: Example of query and retrieved images for a Fair aesthetic outcome. Binary label means class belongs to set {Fair, Poor}. Original label is the ordinal label previous to binarization (Excellent, Good, Fair, or Poor). LRP saliency map is also shown for the test image.

The last query example we present in this work belongs to the binary class {Fair, Poor}, and was labelled as Poor, meaning the worst aesthetic outcome. Quite interestingly, the LRP saliency map points to the breast retraction more than to the breast contour, which also makes clinical sense, as it is the most determinant factor for the poor aesthetic outcome. The top-3 images retrieved were labelled as either Poor or Fair (i.e., same class or neighbouring class).

Even though we only presented four examples of query images and their respective top-3 retrieved most similar past cases, the results were similar for all the other query/test images, in the sense that all saliency maps were focused on clinically relevant regions (breast, breast contour, nipples), and that all the identified most similar past cases belonged to one of the neighbouring original classes.

7.5 Summary and Conclusions

We have proposed to improve the automatic assessment of the aesthetic outcome of breast cancer treatments using deep neural networks. In addition to improving performance, the use of a deep neural network allows a natural semantic search for similar cases and an easy integration into future image generation models. As presented in Table 7.1, our proposed model outperforms state-of-the-art methods for aesthetic evaluation and does not require manual or semi-automatic

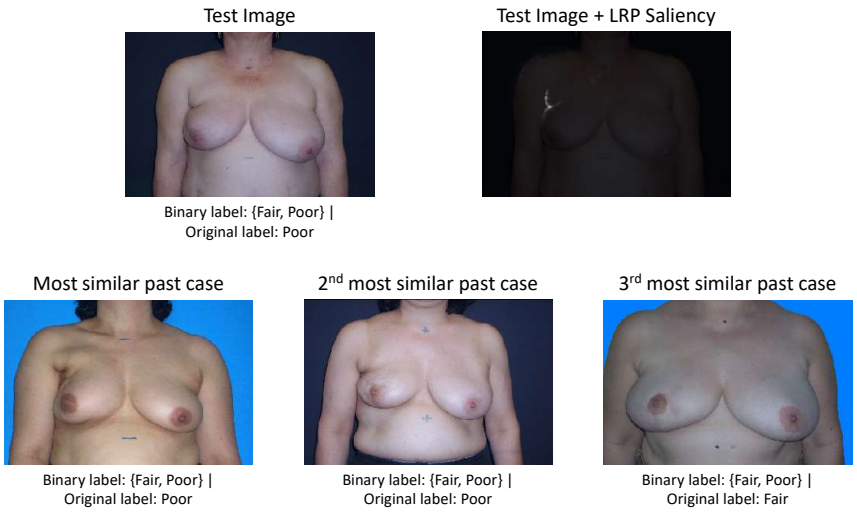


Figure 7.6: Example of query and retrieved images for a Poor aesthetic outcome. Binary label means class belongs to set {Fair, Poor}. Original label is the ordinal label previous to binarization (Excellent, Good, Fair, or Poor). LRP saliency map is also shown for the test image.

preprocessing during inference. Furthermore, as illustrated by Figures 7.3, 7.4, 7.5, and 7.6, it has the capacity to find meaningful past cases, which can be extremely useful for teaching purposes (for instance, new breast surgeons, or nurses) and for management of expectations (for the patients).

Besides being the first work to explore image retrieval in the aesthetic evaluation context, it was also the first work to explore the potential of deep neural networks in the assessment of the aesthetic outcome of breast cancer treatments. In future work, we plan to extend this model to the original ordinal scenario in order to completely replace the SVM models currently being used, and integrate it in a web application that will be openly accessible to any breast unit in the world. Moreover, we want to deepen the explainability of the model, exploring the inherent interpretability generated by the intermediate supervision and representation, in order to provide multimodal explanations (by complementing the retrieval with the importance given by the high-level concepts learnt in the semantic space, similarly to what is done in Silva *et al.* [150]). Finally, we will explore image generation techniques currently used for privacy-preserving case-based explanations [113; 114] to adapt the retrieved past cases to the biometric characteristics of the query image in order to maximize patient engagement and acceptance.

Chapter 8

Assessment of Ordinal Classifiers and its application to Aesthetic Evaluation

Foreword on Author Contributions

Part of the results of this work have been disseminated in the form of one paper in an international conference:

- [151] W. Silva, J. R. Pinto, and J. S. Cardoso, "A Uniform Performance Index for Ordinal Classification with Imbalanced Classes," in *International Joint Conference on Neural Networks (IJCNN 2018)*, Jul. 2018. doi:10.1109/IJCNN.2018.8489327

8.1 Context and Motivation

The previous chapter tackled a simplified version of the aesthetic evaluation problem, where we only dealt with binary classification (Excellent and Good vs Fair and Poor). In a more complete setting, we want to classify each image into one of the original four classes, meaning we have a multi-class problem. Nonetheless, this is not a typical multi-class problem where classes are unrelated. In this scenario, although there is a finite set of possible labels like in any classification task, the labels present a natural inherent order among themselves like in regression problems [25; 65; 79].

It is incautious to objectively state that there is a natural definitive order among cats, dogs, and koalas, but it is undeniable that excellent represents a better aesthetic outcome than good, fair and poor. While the first example pertains to a nominal classification task, the second illustrates the nature of ordinal classification problems, where labels typically present relationships of superiority and inferiority between them.

This generates extraordinary requirements for classifiers on ordinal contexts. Recalling the example above, misclassifying a cat as a dog or a koala is equally undesirable, but it is much worse to misclassify excellent as fair than to attribute them a good aesthetic result. This means misclassifications should not be treated equally, and their influence should relate to the natural order between classes [79].

Similarly, if we attribute good to an excellent case, it would be more adequate and fair to misclassify a good as fair, than to give it excellent. This reveals another property of ordinal classifiers: misclassifications that preserve the natural order of the labels are more desirable than misclassifications that infringe it.

A good ordinal classifier should address these concerns [24; 48; 65], and a suitable ordinal classification metric should be able to adequately capture the degree to which the classifiers comply with them. Furthermore, the metric should also be robust against common classification issues, such as imbalanced classes [49]. Due to the natural order between classes, imbalanced classes are even more common in ordinal settings [130], with the first and last classes generally being under-represented in samples/datasets (as it is the case in the aesthetic evaluation of breast cancer treatments, with excellent and poor being the most under-represented classes).

Furthermore, ordinal classification problems do not boil down to the aesthetic evaluation of breast cancer treatments and are currently present in all fields of research, from computer vision to social sciences [2], which magnifies the importance of adequate performance measurement. In this work, we aim to fill this void with two variants of a novel index, for performance assessment and comparison of ordinal classification in imbalanced settings, that more closely follow the explained desirable behavior.

Several metrics are currently used for the measurement of the performance of ordinal classifiers. However, each one presents its own weaknesses when dealing with this very specific and demanding scenario.

One of such metrics is the Misclassification Error Rate (MER) (8.1). Despite considering the accuracy of the predictions, it fails to account for the natural order of the classes by attributing equal cost to all misclassifications, which is undesirable for performance assessment in ordinal classification tasks.

$$MER = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \hat{y}_i) \tag{8.1}$$

Mean Squared Error (MSE) (8.2) or Mean Absolute Error (MAE) (8.3) are two of the most common, where higher numerical differences between the actual and predicted labels are reflected on the error, resulting in higher penalization of bigger mistakes (such as estimating class $\hat{y} = 5$ to an object of true class $y = 1$) over smaller mistakes (attributing $\hat{y} = 2$ for the same object). The error sum is then averaged over all N observations.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{8.2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{8.3}$$

Nevertheless, both metrics present a significantly disadvantageous dependence on the numbers arbitrarily assigned to each class. This can be fixed by defining the classes by their indexes on a

confusion matrix, but MSE and MAE will still equally penalise "forwards" (estimating a following class) and "backwards" errors (estimating a previous class). In ordinal classification problems, where ranking plays a major role, this lack of distinction between errors is a significant flaw.

To attend to the relevance of ranking in ordinal classification, one common metric is the Spearman's rank correlation coefficient R_S [162], based on two rank vectors p and q, of length N, associated with the variables y and $\hat{y}$:

$$R_s = \frac{\sum(p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum(p_i - \bar{p})^2 \sum(q_i - \bar{q})^2}}. \tag{8.4}$$

However, as verifiable on (8.4), the Spearman's coefficient is still dependent on the values chosen for the ranks to represent the classes.

Kendall's τ_b [88], in turn, also takes into account ranking in the measurement of classification performance, but is independent from the values used to represent each class:

$$\tau_b = \frac{\sum q_{ij} p_{ij}}{\sqrt{\sum q_{ij}^2 \sum p_{ij}^2}}, \tag{8.5}$$

where q_{ij} behaves as follows:

$$\begin{cases} q_{ij} = 1, \text{ if } q_i > q_j \\ q_{ij} = 0, \text{ if } q_i = q_j \\ q_{ij} = -1, \text{ if } q_i < q_j \end{cases}, \tag{8.6}$$

and the same is true for p_{ij}.

In the same line of thought, and taking into consideration the high number of cases in which a tie happens, Pinto da Costa *et al.* [51] introduced r_{int}:

$$r_{int} = -1 + 2\frac{\text{card}(S_1 \cap S_2)}{\sqrt{\text{card}(S_1)\text{card}(S_2)}}. \tag{8.7}$$

Defining n_{ij} as the number of observations whose true class is y_i and whose predicted class is y_j, the total number of observations whose true class is y_i, $n_{i\bullet}$, is given by $\sum_{j=1}^{K} n_{ij}$, and the total number of observations whose predicted class is y_j, $n_{\bullet j}$, is given by $\sum_{i=1}^{K} n_{ij}$, and we get:

$$\text{card}(S_1) = \sum_{i=1}^{K} \sum_{j=i}^{K} n_{i\bullet} n_{j\bullet} - n, \tag{8.8}$$

$$\text{card}(S_2) = \sum_{i=1}^{K} \sum_{j=i}^{K} n_{\bullet i} n_{\bullet j} - n, \tag{8.9}$$

$$\text{card}(S_1 \cap S_2) = \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{i'=i}^{K} \sum_{j'=j}^{K} n_{ij} n_{i'j'} - n. \tag{8.10}$$

All three indices of similarity, R_S, τ_b and r_{int}, vary between -1 and 1.

However, it is fair to affirm that both Kendall's τ_b and r_{int}, by assuming that the only thing that matters is the order relation between classes, go too far in their quest for abstraction from class labels. The reliance on relative order is beneficial for robust ranking error measurement, but causes critical loss of information on absolute classification error.

The ideal solution would consider both the natural ranking between classes and the absolute classification accuracy on the performance assessment. Considering this, the Ordinal Classification Index was proposed by Cardoso and Sousa [25], fitted for accounting for both absolute classification error and ranking error. With r denoting a row and c a column of the considered confusion matrix, the OC_β^γ was defined as:

$$\begin{aligned} OC_\beta^\gamma &= \min\left\{1 - \frac{\text{benefit(path)}}{N+M} + \beta(\text{penalty(path)})\right\} \\ &= \min\left\{1 - \frac{\sum_{(r,c)\in\text{path}} n_{r,c}}{N + \left(\sum_{\forall(r,c)} n_{r,c}|r-c|^\gamma\right)^{1/\gamma}} \right. \\ &\qquad \left. + \beta \sum_{(r,c)\in\text{path}} n_{r,c}|r-c|^\gamma \right\}, \end{aligned} \tag{8.11}$$

where the minimization is performed over the set of all consistent paths that can be traced over the confusion matrix, from entry $(1,1)$ to entry (K,K). As defined in [25], a path is consistent if every pair of nodes is nondiscordant, which in turn means that the relative order of the true classes for that pair is not opposite to the relative order of the predicted classes.

Each path is characterized by a benefit and a penalty. The benefit will give advantage to paths that include the largest entries on the confusion matrix, rewarding paths that better follow the natural class order. The penalty will punish paths as they deviate from the main diagonal, effectively acting as a regularizer and including classification accuracy on the performance assessment. The parameter β will weight the benefit and penalty, allowing the metric to focus more on accuracy or ranking.

However, this metric suffers from two main setbacks. First, the freely tunable parameter, β, generates ambiguity as it allows users to choose its value for their own benefit. Second, and similarly to all aforementioned metrics, it is sensitive to imbalanced classes: the influence of each class is not necessarily uniform, and is instead linked to the number of instances of each on the considered population sample.

This implies that, if a class is significantly better represented in the sample than the others, it will have a much higher impact on the metric than the remaining classes, which is generally undesirable. To address this issue, some alternative metrics have been proposed.

If the imbalanced classification problem at hand is binary, then two metrics are commonly used: the F_1 (8.12) and the G-mean (8.13). However, they are largely limited by being solely applicable to binary classification.

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{8.12}$$

$$\text{G-mean} = \sqrt{\frac{TP}{TP+FN}\left(1-\frac{FP}{FP+TN}\right)} \tag{8.13}$$

For imbalanced problems with $K > 2$ classes, several adaptations of MAE have been proposed to more robustly assess the performance of the classifiers. Namely, Maximum (MMAE) [50], and Average Mean Absolute Error (AMAE) [6], presented in (8.14) and (8.15), respectively. With these, the error is determined separately for each class, and the metric assumes the maximum or average of the values obtained, effectively enforcing uniform influence of all classes, independently of possible imbalanced representations.

$$MMAE = \max\{MAE_k; k = 1, ..., K\} \tag{8.14}$$

$$AMAE = \frac{1}{K}\sum_{k=1}^{K} MAE_k \tag{8.15}$$

Nevertheless, these metrics, more devoted to imbalanced problems, present the major weakness of overlooking the importance of ranking for ordinal classification. Furthermore, the adaptations of Mean Absolute Error still depend on the values chosen as labels for each class.

Considering this current panorama in ordinal imbalanced metrics here presented, it is possible to conclude that there is still no metric that can adequately combine classification accuracy and ranking in the same metric, while remaining robust against the influence of imbalanced classes. In the following sections, two variants of a new index, based on the aforementioned Ordinal Classification Index, are formulated and proposed to fill this void.

8.2 Methodology

8.2.1 Conceptual formulation

The proposed index results of the adaptation of the aforementioned Ordinal Classification Index, OC_β^γ, proposed by Cardoso and Sousa [25], and aims towards the achievement of robustness against imbalanced classes, and the suppression of the freely tunable parameter β. First, to fix the weakness related to imbalanced classes, we start by re-interpreting the OC_β^γ in a stochastic formulation. Towards that goal, a simple algebraic manipulation of OC_β^γ gives:

$$OC_\beta^\gamma = \min\left\{1 - \frac{\sum_{(r,c)\in\text{path}} \frac{n_{r,c}}{N}}{\sum_{\forall(r,c)} \frac{n_{r,c}}{N} + \left(\sum_{\forall(r,c)} \frac{n_{r,c}}{N}|r-c|^\gamma\right)^{1/\gamma}} + N\beta \sum_{(r,c)\in\text{path}} \frac{n_{r,c}}{N}|r-c|^\gamma\right\}, \tag{8.16}$$

where $\beta \in \mathbb{R}_{\geq 0}$ and $\gamma \in \mathbb{R}_{>0}$.

Interpreting Y and $\hat{Y}$ as random variables, the normalized confusion matrix with entries $\frac{n_{r,c}}{N}$ can be understood as an approximation of the joint probability function between Y and $\hat{Y}$, $p(y,\hat{y})$.

Adopting this stochastic view, OC_β^γ can be written as:

$$OC_\beta^\gamma = \min\left\{1 - \frac{\sum_{(y,\hat{y})\in \text{path}} p(y,\hat{y})}{\sum_{\forall(y,\hat{y})} p(y,\hat{y}) + \left(\sum_{\forall(y,\hat{y})} p(y,\hat{y})|y-\hat{y}|^\gamma\right)^{1/\gamma}} + N\beta \sum_{(y,\hat{y})\in \text{path}} p(y,\hat{y})|y-\hat{y}|^\gamma\right\}, \tag{8.17}$$

where $(y,\hat{y})$ is equivalent to the notation (r,c), used for confusion matrices.

Since, for any pair of random variables A and B, the joint probability can be written as $p(a,b) = p(a)p(b|a)$, we have:

$$OC_\beta^\gamma = \min\left\{1 - \frac{\sum_{y:(y,\hat{y})\in \text{path}} \sum_{\hat{y}} p(y)p(\hat{y}|y)}{\sum_y \sum_{\hat{y}} p(y)p(\hat{y}|y) + \left(\sum_y \sum_{\hat{y}} p(y)p(\hat{y}|y)|y-\hat{y}|^\gamma\right)^{1/\gamma}} + N\beta \sum_{y:(y,\hat{y})\in \text{path}} \sum_{\hat{y}} p(y)p(\hat{y}|y)|y-\hat{y}|^\gamma\right\}. \tag{8.18}$$

This is equivalent to:

$$OC_\beta^\gamma = \min\left\{1 - \frac{\sum_{y:(y,\hat{y})\in \text{path}} p(y) \sum_{\hat{y}} p(\hat{y}|y)}{\sum_y p(y) \sum_{\hat{y}} p(\hat{y}|y) + \left(\sum_y p(y) \sum_{\hat{y}} p(\hat{y}|y)|y-\hat{y}|^\gamma\right)^{1/\gamma}} + N\beta \sum_{y:(y,\hat{y})\in \text{path}} p(y) \sum_{\hat{y}} p(\hat{y}|y)|y-\hat{y}|^\gamma\right\}. \tag{8.19}$$

In (8.19), the dependency of OC_β^γ on the class distribution $p(y)$ is evident. When classes are highly imbalanced, classes with high probability dominate the result. Like AMAE brings robustness to the MAE metric in imbalance settings by replacing the original $p(y)$ distribution with uniform probabilities $1/K$ for each class, we propose to modify OC_β^γ using the same strategy. Thus, we propose the first variant of our index, the Uniform Ordinal Classification Index, UOC_β^γ, as:

$$UOC_\beta^\gamma = \min\left\{1 - \frac{\sum_{y:(y,\hat{y})\in \text{path}} \frac{1}{K} \sum_{\hat{y}} p(\hat{y}|y)}{\sum_y \frac{1}{K} \sum_{\hat{y}} p(\hat{y}|y) + \left(\sum_y \frac{1}{K} \sum_{\hat{y}} p(\hat{y}|y)|y-\hat{y}|^\gamma\right)^{1/\gamma}} + N\beta \sum_{y:(y,\hat{y})\in \text{path}} \frac{1}{K} \sum_{\hat{y}} p(\hat{y}|y)|y-\hat{y}|^\gamma\right\}, \tag{8.20}$$

which can be simplified to:

$$UOC_\beta^\gamma = \min\left\{1 - \frac{\sum_{(y,\hat{y})\in \text{path}} p(\hat{y}|y)}{K + \frac{K}{K^\gamma}\left(\sum_{\forall(y,\hat{y})} p(\hat{y}|y)|y-\hat{y}|^\gamma\right)^{1/\gamma}} + \frac{N}{K}\beta \sum_{(y,\hat{y})\in \text{path}} p(\hat{y}|y)|y-\hat{y}|^\gamma\right\}. \tag{8.21}$$

As β is, still, a user defined constant, it is possible to recast $N\beta$ as β, finally giving the proposed formulation to UOC:

$$UOC_{\beta}^{\gamma} = \min\left\{1 - \frac{\sum_{(y,\hat{y})\in \text{path}} p(\hat{y}|y)}{K + \frac{K}{K^{\gamma}}\left(\sum_{\forall(y,\hat{y})} p(\hat{y}|y)|y-\hat{y}|^{\gamma}\right)^{1/\gamma}} + \frac{\beta}{K}\sum_{(y,\hat{y})\in \text{path}} p(\hat{y}|y)|y-\hat{y}|^{\gamma}\right\}. \tag{8.22}$$

Following a procedure similar to [25], it is possible to show that for $\beta \geq 1$, UOC_{β}^{γ} in (8.22) results in a metric and the optimal path is always over the main diagonal. Thus, considering $\beta \in [0,1]$, for specific settings that require especial emphasis in either ranking error or instance-based error, the variant UOC_{β}^{γ} can be used with a user-defined β in the lower or higher end, respectively, of its range.

Nevertheless, the existence of β and γ remains a source of ambiguity in most applications. For γ, we propose the value 1, as used for OC, as the Minkowski distance is generally used for the values of 1, 2, or infinity, and the variation of the results with different γ values will not be significant [25]. For β, we propose its elimination through the formulation of a second variant, with the integration of UOC_{β}^{1} along β's aforementioned range of values, through:

$$A_{UOC} = \int_0^1 UOC_{\beta}^{1}\, d\beta \tag{8.23}$$

8.2.2 Application from estimates in a confusion matrix

When applied to a real scenario, $p(y,\hat{y})$ and $p(y)$ can easily be estimated through Maximum Likelihood Estimation, and UOC_{β}^{γ} can be applied from a confusion matrix.

$$UOC_{\beta}^{\gamma} = \min\left\{1 - \frac{\sum_{(y,\hat{y})\in \text{path}} \hat{p}(\hat{y}=c|y=r)}{K + \frac{K}{K^{\gamma}}\left(\sum_{\forall(y,\hat{y})} \hat{p}(\hat{y}=c|y=r)|y-\hat{y}|^{\gamma}\right)^{1/\gamma}} + \frac{\beta}{K}\sum_{(y,\hat{y})\in \text{path}} p(\hat{y}=c|y=r)|y-\hat{y}|^{\gamma}\right\} \tag{8.24}$$

$$UOC_{\beta}^{\gamma} = \min\left\{1 - \frac{\sum_{(r,c)\in \text{path}} n_{r,c}/N_r}{K + \frac{K}{K^{\gamma}}\left(\sum_{\forall(r,c)} (n_{r,c}/N_r)|r-c|^{\gamma}\right)^{1/\gamma}} + \frac{\beta}{K}\sum_{(r,c)\in \text{path}} \frac{n_{r,c}}{N_r}|r-c|^{\gamma}\right\} \tag{8.25}$$

With (8.25), it is also easy to integrate and obtain A_{UOC} from a confusion matrix. In Fig. 8.1, we illustrate A_{UOC} and the values of UOC_{β}^{1} obtained from an example confusion matrix.

8.2.3 Handling unobserved classes

One particular issue can arise from the application of UOC_{β}^{γ} to confusion matrices: it is not guaranteed that every class will be observed in the considered set/sample, especially if the latter is

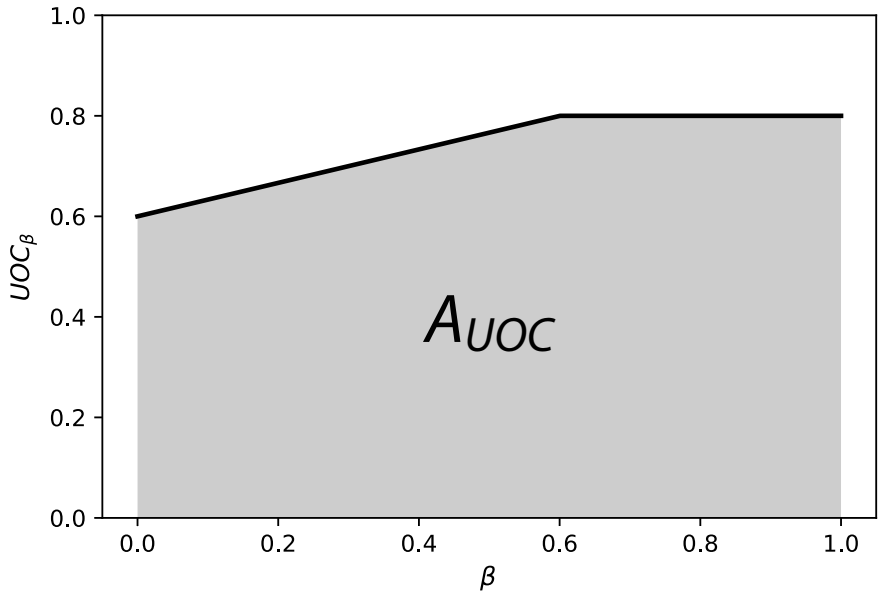


Figure 8.1: Illustration of A_{UOC} and the values of UOC^1_β obtained from an example confusion matrix.

small, and N_r can, in some cases, be zero. We propose to generalise (8.22) to attend to these situations.

Let $\mathbb{1}_{\mathscr{O}}$ be the indicator function of the set $\mathscr{O}$ of the observed classes:

$$\mathbb{1}_{\mathscr{O}}(y) = \begin{cases} 1 & y \in \mathscr{O} \\ 0 & y \notin \mathscr{O} \end{cases}, \tag{8.26}$$

and $K' \leq K$ the cardinality of $\mathscr{O}$ (the number of observed classes). In order to make (8.19) robust in imbalanced settings and to address unobserved classes, we propose to fix the probability distribution for y to an uniform distribution over the observed classes only, with $p(y) = \frac{1}{K'}\mathbb{1}_{\mathscr{O}}(y)$. Introducing this proposed distribution in (8.19) and simplifying as before, one obtains

$$UOC^\gamma_\beta = \min\left\{1 - \frac{\sum_{(y,\hat{y})\in \text{path}} p(\hat{y}|y)\mathbb{1}_{\mathscr{O}}(y)}{K' + \frac{K'}{K'^\gamma}\left(\sum_{\forall(y,\hat{y})} p(\hat{y}|y)\mathbb{1}_{\mathscr{O}}(y)|y-\hat{y}|^\gamma\right)^{1/\gamma}} + \frac{\beta}{K'}\sum_{(y,\hat{y})\in\text{path}} p(\hat{y}|y)\mathbb{1}_{\mathscr{O}}(y)|y-\hat{y}|^\gamma\right\}. \tag{8.27}$$

For real scenarios, in place of (8.25), we can rewrite (8.27) in order to make it robust against

unobserved classes while using estimates from a confusion matrix, and we obtain:

$$UOC_{\beta}^{\gamma} = \min\left\{1 - \frac{\sum_{(r,c)\in \text{path}} (n_{r,c}/N_r)\mathbb{1}_{\mathscr{O}}(r)}{K' + \frac{K'}{K'^{\gamma}}\left(\sum_{\forall (r,c)} (n_{r,c}/N_r)\mathbb{1}_{\mathscr{O}}(r)|r-c|^{\gamma}\right)^{1/\gamma}} + \frac{\beta}{K'}\sum_{(r,c)\in \text{path}} \frac{n_{r,c}}{N_r}\mathbb{1}_{\mathscr{O}}(r)|r-c|^{\gamma}\right\}, \quad (8.28)$$

that can be similarly used in (8.23) for settings that do not present a preferential value for β.

All this effectively amounts to ignore classes that are not observed in the considered sample. Alternatives such as considering uniform conditional probabilities on those cases presents the disadvantage of consistently penalizing performance because of each unobserved class. On the other hand, assuming perfect performance for each unobserved class is overly optimistic, as it rarely will be true. Our proposal avoids generating tendencies to either benefit or penalise performance due to unobserved classes, and instead bases it entirely on the classes that are observed.

8.3 Experimental Setup

Single Sample and Tridiagonal Matrices

As stated by Cardoso and Sousa [25], one of the weaknesses between τ_b, R_s, or r_{int} and MER, MAE, or MSE, is that the former are not applicable to performance assessment with a single observation.

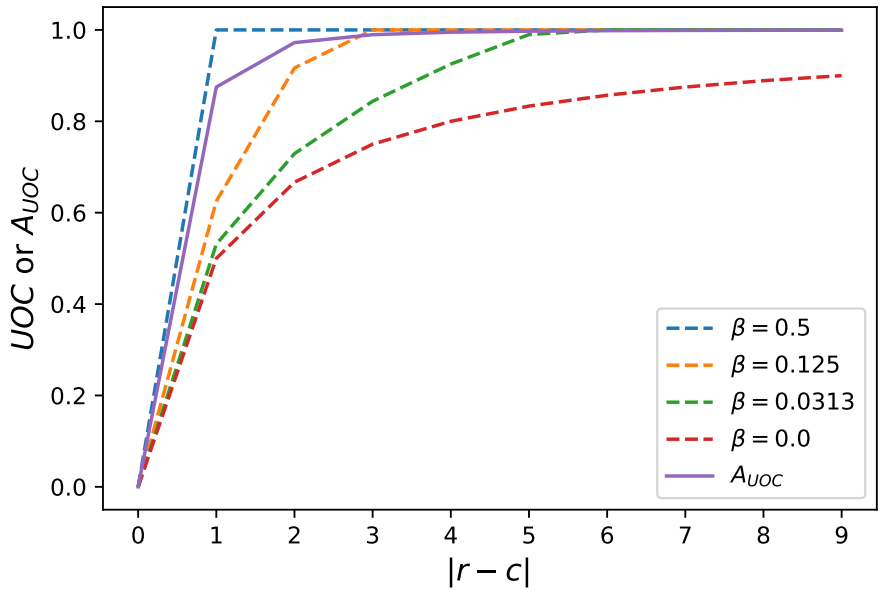


Figure 8.2: Values of UOC, for several β values, and A_{UOC}, obtained with a confusion matrix with a single sample, according to its distance to the diagonal $|r-c|$.

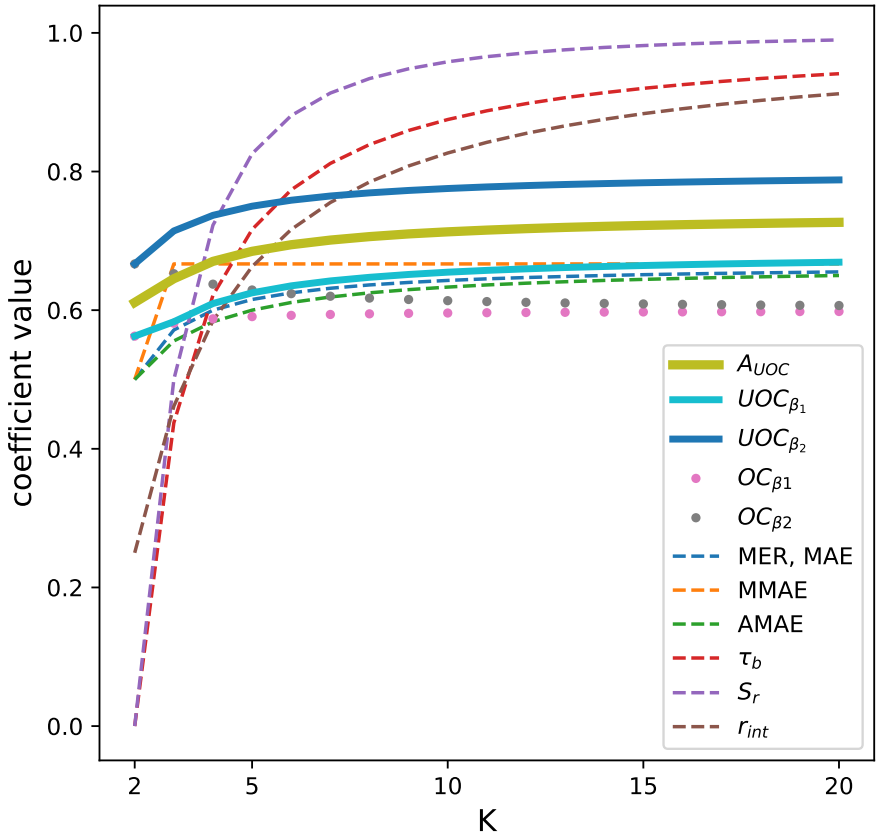


Figure 8.3: Evolution of metric values with the total number of classes K, when applied to tridiagonal matrices.

Similarly to OC_β^γ, UOC_β is applicable to single observations, and its value increases monotonically from 0 to 1 with the increase of the sample's distance to the diagonal, and the rate is dependent from the chosen value of β. The value of A_{UOC}, although independent from β, presents similar behavior (cf. Fig. 8.2).

One other issue of r_{int}, R_s, and τ_b, is the result of their application to tridiagonal matrices (cf. Fig. 8.3). These confusion matrices present zeros on all entries except their three main diagonals, where the entries are 1. As K, the number of classes, increases, the three aforementioned metrics converge to 1. To affirm a certain score is the most appropriate for this situation would be reckless, as the relevance on the performance of the two diagonals (other than the main one) is subjective. Nevertheless, attributing a near-perfect performance to classifiers that, simultaneously, present a MER of $2/3$, is clearly inappropriate. The proposed variants, UOC_β^γ and A_{UOC}, present an intermediate behavior between the remaining metrics, while steering away from the undesirable behavior of r_{int}, R_s, and τ_b.

Table 8.1: Results for the simulated confusion matrices, with $\beta_1 = 0.25$ and $\beta_2 = 0.75$

Classifier	**Accuracy-focused**					**Ranking-Focused**			**Mixed Focus**				
	Sensitive to imbalance			***Robust to imbalance***					***Sensitive to imbalance***		***Robust to imbalance***		
	MER	MSE	MAE	MMAE	AMAE	R_s	τ_b	r_{int}	$OC^1_{\beta_1}$	$OC^1_{\beta_2}$	UOC_{β_1}	UOC_{β_2}	A_{UOC}
A	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00
B	0.56	0.56	0.56	1.00	0.50	0.90	0.86	0.86	0.40	0.50	0.46	0.67	0.56
C	0.56	1.22	0.78	2.00	0.75	0.67	0.61	0.69	0.50	0.63	0.62	0.71	0.65
D	0.56	0.56	0.56	1.00	0.50	0.73	0.60	0.74	0.53	0.58	0.56	0.67	0.61
E	0.77	0.77	0.77	1.00	0.50	0.24	0.11	0.53	0.65	0.72	0.68	0.80	0.74
F	0.85	0.85	0.85	1.00	0.50	0.29	0.23	0.79	0.58	0.71	0.56	0.67	0.61

Simulated Examples, Missing, and Imbalanced Classes

To show that the proposed index variants combine both accuracy and ranking in the performance assessment, while remaining robust to missing classes and imbalance, the following simulated confusion matrices ($K = 4$), each one representing the behavior of a classifier, were considered. A comparison between the different metrics is also presented in Table 8.1.

$$CM_A = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} \quad CM_B = \begin{bmatrix} 0 & 4 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

$$CM_C = \begin{bmatrix} 0 & 0 & 4 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} \quad CM_D = \begin{bmatrix} 0 & 4 & 0 & 0 \\ 6 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

$$CM_E = \begin{bmatrix} 0 & 4 & 0 & 0 \\ 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} \quad CM_F = \begin{bmatrix} 0 & 40 & 0 & 0 \\ 6 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

All compared metrics attribute a perfect performance score to classifier A. This is expectable, as a perfect accuracy implies the absence of ranking error. However, the classifiers B, C, and D present classification errors. In an accuracy perspective, B, C, and D present equal MER, but classifier C has higher MSE and MAE. Regarding ranking, B clearly resembles more closely the true order of the classes than the other two classifiers. Thus, the ranking-focused metrics, r_{int}, R_s, and τ_b, attribute worse performance scores to D than B, and τ_b goes even further and gives the lowest score to C as it presents lower ranking error than D, despite the lower accuracy. All accuracy-focused metrics disregard ranking and give equal scores to B and D, and a lower score to C. UOC retains some similarity to OC, as both have the flexibility to resemble ranking-focused metrics for low-range values of β, and to focus on accuracy with higher β values. As expectable, A_{UOC} presents an intermediate behavior. Finally, E and F can be considered similar to classifier D. However, classifier E was tested without objects of class 3, and the dataset used to evaluate classifier F is highly imbalanced. Most existing metrics present sensitivity to imbalanced classes, as they do not attribute equal scores to classifiers D and F, as they should. The exceptions are MMAE, AMAE, and the proposed variants UOC and A_{UOC}. Nevertheless, while other metrics, including the proposed ones, will penalize classifier E due to the missing class, MMAE and AMAE assume an optimistic scenario (no error on the missing classes), which may be rarely true. The proposed index, as stated before, deals with missing classes in a balanced fashion, by ignoring them completely. In this case, this results on a slight performance penalisation, as two thirds of the classes do not conform to ranking order or accuracy, while for D it is only one half.

To showcase the behavior of the proposed index on real situations, and compare it with the aforementioned state-of-the-art alternatives, we trained a Support Vector Machine, a k-Nearest Neighbors, and a Random Forest classifier on 70% of the data of two real public datasets of ordinal classification problems, with imbalanced classes. The predictions of each classifier on the remaining 30% of each dataset were used to build confusion matrices (cf. Figures 8.4 and 8.5) and compute the metrics (cf. Table 8.2). Here, our goal was not to assert the superiority or inferiority of any classifier over the others, but to showcase how each metric allows us to measure and compare their performances based on the resulting confusion matrices.

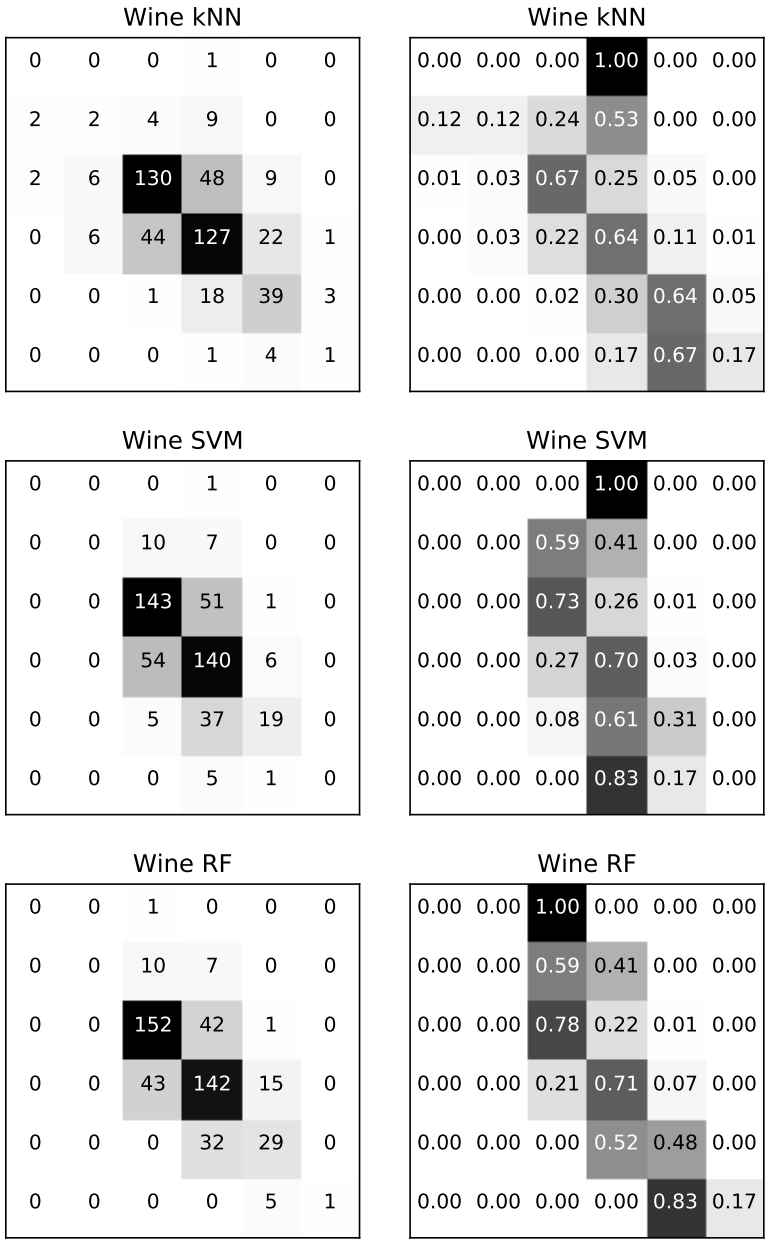


Figure 8.4: Confusion matrices, regular and normalized, for the classifiers kNN, SVM, and Random Forest, used on the wine quality dataset.

Table 8.2: Results for the classifiers in real datasets, with $\beta_1 = 0.25$ and $\beta_2 = 0.75$

Classifier	**Accuracy-focused**					**Ranking-Focused**			**Mixed Focus**				
	Sensitive to imbalance			***Robust to imbalance***					***Sensitive to imbalance***		***Robust to imbalance***		
	MER	MSE	MAE	MMAE	AMAE	R_s	τ_b	r_{int}	$OC^1_{\beta_1}$	$OC^1_{\beta_2}$	UOC_{β_1}	UOC_{β_2}	A_{UOC}
Wine kNN	0.38	0.58	0.44	3.00	1.10	0.57	0.52	0.64	0.46	0.48	0.76	0.82	0.79
Wine SVM	0.37	0.50	0.41	3.00	1.27	0.53	0.50	0.65	0.42	0.44	0.82	0.87	0.84
Wine RF	0.33	0.38	0.34	2.00	0.88	0.66	0.62	0.71	0.37	0.39	0.70	0.81	0.75
ESL kNN	0.37	0.39	0.37	1.00	0.50	0.91	0.84	0.81	0.39	0.40	0.54	0.72	0.63
ESL SVM	0.31	0.80	0.40	5.00	1.11	0.83	0.79	0.78	0.37	0.38	0.75	0.80	0.78
ESL RF	0.37	0.46	0.40	2.00	0.61	0.91	0.83	0.81	0.41	0.42	0.63	0.73	0.68

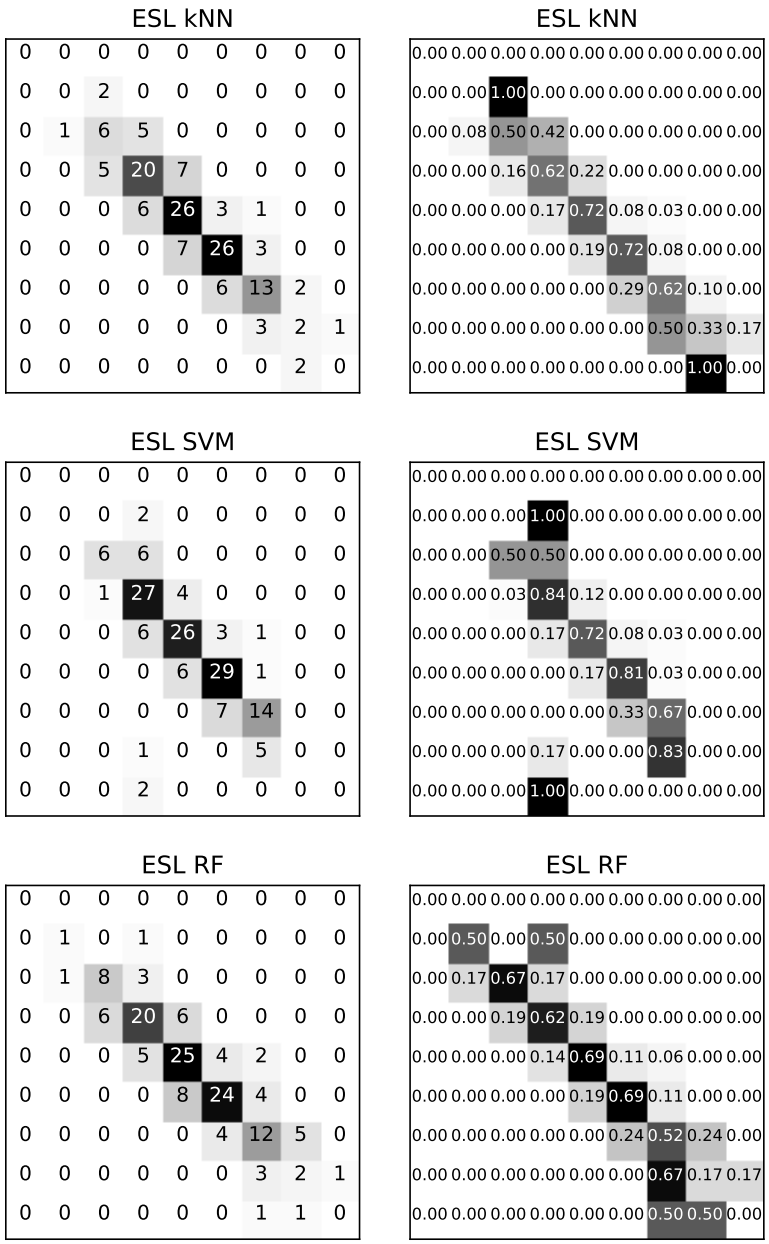


Figure 8.5: Confusion matrices, regular and normalized, for the classifiers kNN, SVM, and Random Forest, used on the ESL dataset.

Wine Quality Dataset

This dataset relate eleven numerical features (such as acidity, sulphates, density, pH, and residual sugar) of Portuguese red wines with its quality ranking [47]. The dataset includes 1599 samples for quality classes three to eight, and is available on Kaggle datasets[1].

[1]Red Wine Quality - Kaggle datasets. Available at: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009.

Analysing the regular confusion matrices of the classifiers, all three appear to have a very similar performance. Nevertheless, when inspecting the normalized confusion matrices, it can be clearly concluded that the predictions of the Random Forest classifier resembles much more closely the true ranking order of the classes, and the overall accuracy inside each class is higher. SVM clearly presented the worst results, while kNN showed an intermediate performance.

MER, by considering equal all misclassifications, attributed similar scores to all classifiers, with Random Forest only slightly better. MSE and MAE, sensitive to imbalance, attributed a worse score to kNN than SVM. The same was verified for the ranking-focused r_{int} and the mixed-focus metrics $OC^1_{\beta_1}$, and $OC^1_{\beta_2}$, denoting that this undesirable behavior is probably due to their sensitivity to imbalanced datasets.

MMAE presented flaws in its claim to be robust against imbalanced classes, since the scores it presents are clearly a major result of the worse represented class 1, with one object that is misclassified by all classifiers. On the other hand, AMAE shows the desired behavior, but the non-normalised values it takes are not fit for absolute performance assessment, and unfortunately limit its use to the relative comparison of classifiers in equal settings.

Both proposed variants UOC_β and A_{UOC} present the desired behavior. UOC_β presents the advantage of flexibility: with the lower value of β, the index favored ranking and the difference between the classifiers' performance scores was amplified; and with the higher value of β, the focus on accuracy increased and the behavior of UOC approached that of an accuracy-focused metric robust to imbalance. On the other hand, A_{UOC} presented an intermediate behavior, ideal for situations where neither ranking nor accuracy should be especially favored.

Employee Selection (ESL) Dataset

The ESL dataset includes numerical evaluations of 488 job applicants in four relevant psychometric parameters, and a final ordinal classification of the applicants according to their fit to the job (from 1 up to 9). The dataset belongs to the Business Administration School of the Tel Aviv University, and is available at Weka datasets[2].

Again, looking at the regular confusion matrices of the classifiers gives an impression of similarity between performances. However, the normalized confusion matrices dissipate this idea. Regarding the true ranking order of the classes, it is clear the distinction between SVM and the other two classifiers. In terms of pure ranking, kNN and RF have similar performance and both present better results than SVM, which has the best result when using MER, a completely accuracy-focused metric. Furthermore, MAE was not able to differentiate between the SVM and the RF.

Mixed-focus metrics $OC^1_{\beta_1}$ and $OC^1_{\beta_2}$ erroneously consider the SVM as the best classifier. This happens due to a class imbalance in the ESL dataset. On the contrary, UOC_{β_1} and UOC_{β_2}, being robust to class imbalance, acknowledge the best performance of the kNN. A_{UOC}, which represents

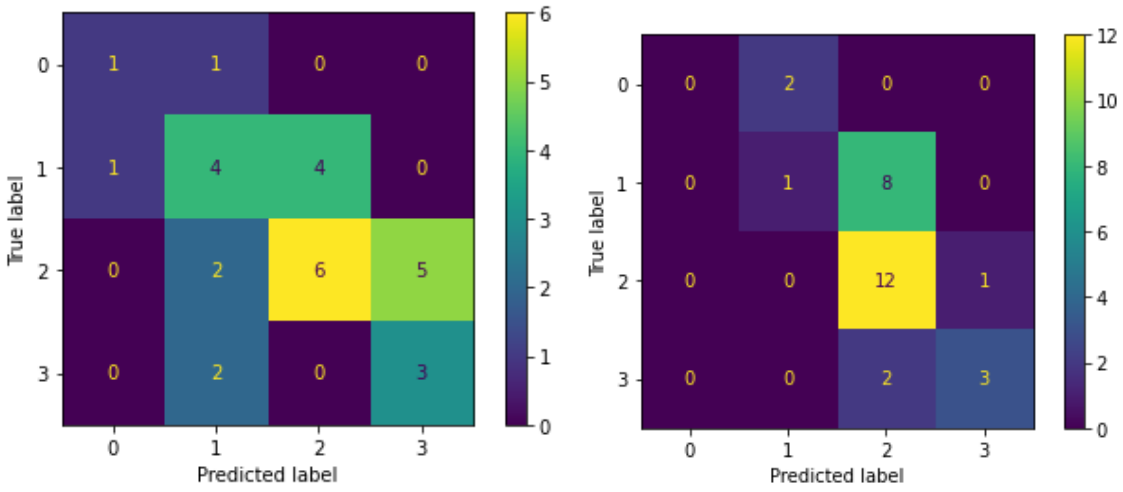
[2]Dr. Arie Ben David ordinal datasets - Weka datasets. Available at: http://weka.wikispaces.com/Datasets.

an equilibrium between ranking and accuracy, is also capable of distinguishing the performances between the three classifiers and is in agreement with UOC_{β_1} and UOC_{β_2}.

8.4 Ordinal Assessment of Aesthetic Evaluation Classifiers

After demonstrating the usefulness of the proposed metrics in general ordinal and imbalanced problems, we were interested in analyzing the impact of this new metric for the machine learning models developed in the context of our clinical problem. To exemplify, we will focus on using only the traditional SVM classifier and the data used in Chapter 7. However, this time taking into account the original four classes. The only difference was the use of all 23 high-level features as input.

Since we are interested in comparing the performance with regards to the ordinal performance in the aesthetic evaluation of breast cancer treatments, we compared an SVM classifier trained in a standard multi-class setting, and an SVM trained following an ordinal approach [65].



(a) Multi-class (b) Frank and Hall

Figure 8.6: Confusion matrices for the SVM classifier. In (a) by training as a typical multi-class problem. In (b) using the Frank and Hall approach.

In Fig. 8.6, we present the resulting confusion matrices for the test data. As it is clear from the confusion matrices, the ordinal approach generates a confusion matrix with most values either on the main diagonal or neighboring the main diagonal, demonstrating a higher focus in ranking than the common multi-class classifier. However, it never predicted one of the classes (the minority class), showing weaknesses in terms of tackling the imbalance. Those exact details can then be observed in a condensed form in Table 8.3. It is clear that the Frank and Hall approach leads to a better accuracy performance. However, at the expense of poor performance for the minority classes (even though we have used balanced weights for each binary problem). On the other hand, the standard multi-class approach was able to pay attention to all classes, but at the expense of general accuracy performance. Besides this information, the proposed metric A_{UOC} also pays

attention to the ranking performance, providing a more complete evaluation. Even though the Frank and Hall approach captures much better the ranking nature of the problem, it is highly penalized by the poor performance in the minority classes, obtaining an overall performance a bit poorer than the traditional multi-class approach.

Table 8.3: Results for the test set. Multi-class and Frank Hall represent the two approaches for training the SVM model. Linear kernel was used since it generated the best results.

Model/Metrics	Accuracy ↑	Balanced Accuracy ↑	A_{UOC} ↓
SVM (Multi-class)	0.48	**0.50**	**0.62**
SVM (Frank and Hall)	**0.55**	0.41	0.63

8.5 Summary and Conclusions

In this work, two variants of a novel index for performance assessment of ordinal classification in imbalanced settings are proposed. The first, UOC_{β}^{γ}, is tunable to give preference to ranking or accuracy error, and thus allows for tailored performance assessment to fit settings that present such preferences. For a fixed, parameter-free performance assessment, A_{UOC} presents an intermediate behavior.

The proposed index was evaluated and compared with state-of-the-art alternatives in several simulated and real scenarios. The results show that its variants, unlike most other alternatives, are able to capture both ranking and instance-based error in the performance assessment, while remaining impervious against imbalanced classes, and presenting a desirable behavior when faced with unobserved classes.

Thus, it can be concluded that the proposed metrics are suitable to be applied in the complete absolute assessment of classification performance, as well as the adequate and robust relative comparison between sets of ordinal classifiers in imbalanced settings. Meaning that now it is possible to effectively compare the performance of different algorithms for the general problem of the aesthetic evaluation of breast cancer treatments.

Part IV

Conclusions

Chapter 9

Conclusion

Breast cancer is the most frequently diagnosed cancer in women worldwide. Breast cancer survival rates have increased substantially in recent years, shifting the clinical focus from survival alone to survival and improved quality of life after treatment. This reason alone, plus the availability of different types of treatment, are calling for an objective method to assess the aesthetic outcome of treatments, which is directly correlated with feelings of well-being and thus to quality of life after treatment. The present thesis described an effort to develop machine learning methods to assist the clinicians/breast surgeons in the aesthetic evaluation of breast cancer treatments, helping to define a personalized ideal treatment, assessing surgery's quality and allowing a fair comparison between different breast units.

Motivated by the central clinical application driving this thesis, we addressed both fundamental and applied topics. The two main fundamental topics in the area of machine learning were explainable artificial intelligence, and ordinal classification learning. These fundamental topics are summarized mainly in part II of this thesis (see Chapters 2- 4) but also in part III (see Chapter 8). Then, we presented a literature review regarding the aesthetic evaluation of breast cancer treatments (see Chapter 5) and developed methods to solve application-specific tasks, which are detailed in part III (see Chapters 6, and 7).

The research outcomes of this thesis have been published in numerous international conferences and journals. This work benefited from collaborations with local (INESC TEC) colleagues, but also with other research teams, national and international (e.g., from Austria, Switzerland, and United States of America). Currently, international collaborations are being fostered in a Carnegie Mellon Portugal project entitled TAMI ("TAMI - Transparent Artificial Medical Intelligence"), and will be further increased within a new European project recently approved entitled CINDERELLA ("CINDERELLA - Clinical Validation of an AI-based approach to improve the shared decision-making process and outcomes in Breast Cancer Patients proposed for Locoregional treatment").

9.1 Fundamental Contributions

A significant part of this work was devoted to fundamental contributions in Machine Learning and Computer Vision, mainly regarding explainable artificial intelligence but also ordinal learning. Our fundamental contributions were the following:

- Generation of diverse and complementary explanations (Chapter 2, [150; 153]). We proposed architectures/approaches to make decisions and, at the same time, be viable to produce diverse types of explanations. We argue that different people think in different ways (more visually, or more verbally), and that a decision-support system should have the capability to convince different types of explanation consumers. We also present a framework to evaluate the quality of the explanations produced ("3Cs of interpretability") and a method based on one of these metrics (correctness) to extract a single explanation representing an ensemble decision. These approaches were validated in three different problems (one of each, the aesthetic evaluation), with the ensemble model leading to the best classification performance while generating diverse explanations, and having the more representative explanation with very high performance in terms of explanation quality.

- Improvement of content-based medical image retrieval by exploring the potential of interpretability post-model techniques to focus the attention of the retrieval in clinically-relevant regions (Chapter 3, [154; 156]). We proposed an interpretability-guided content-based medical image retrieval approach, focusing the attention of the retrieval in regions of clinical interest without requiring additional supervision (based on saliency maps generated from the classification decisions). This method was validated in Chest X-ray data, leading to a retrieval process more similar to the one of an experienced board-certified radiologist than by using state-of-art approaches. Moreover, the retrieval performance of our method lies within the inter-variability of board-certified radiologists for the same task.

- Development of privacy-preserving machine learning models to anonymize medical case-based explanations (Chapter 4, [113–115]). Case-based explanations are of extreme value in clinical contexts. However, in some circumstances, their use is prohibited due to privacy concerns. We studied state-of-the-art privacy-preserving machine learning models to anonymize medical case-based explanations, and proposed alterations to the most promising model in order to fully comply with the requirements of such a system (i.e., privacy, explanatory value, and realism). We mainly validated our methods in a medical/biometrics dataset of eye images, also performing experiments in Chest X-ray data.

- Proposal of new metrics for performance assessment of ordinal classification with imbalanced classes (Chapter 8, [151]). The aesthetic evaluation of breast cancer treatments is an inherently ordinal problem, with a considerably imbalanced scenario (most cases fit within the two intermediate classes, i.e., fair and good). In order to properly assess the quality of the algorithms developed, one had to have a metric specifically designed for that same

scenario. We addressed this need with the proposal of UOC^{γ}_{β}, and a variant of it, which does not require any hyper-parameter selection, A_{UOC}.

9.2 Applied Contributions

Besides our contributions to explainable artificial intelligence and ordinal learning, we also explored more application-specific tasks. We performed a literature review on the aesthetic evaluation of breast cancer treatments (Chapter 5, [27]), discussing the main topics of research, limitations of current techniques (clinical and scientific), and presenting open challenges that still need to be addressed. In addition to this review, we worked on two main application-oriented tasks:

- Detection of keypoints in photographs of women's torso after being subjected to breast cancer treatments (Chapter 6, [70; 71; 152]). We introduced deep learning methodologies to this topic, proposing three different approaches, two fully deep learning-based and one combining deep learning and traditional methods. All three proposed approaches led to an improvement in the results when compared with the previous state-of-the-art. Moreover, at inference, these methods are much quicker than the previous ones, which is crucial for their deployment in web applications to aid the clinicians.

- Aesthetic evaluation of breast cancer treatments (Chapter 7, [155]). The proposed methodology heavily relies on regularization strategies to fight the inherent overfitting to a problem that has only a small dataset available and deals with a considerably subjective task. Besides surpassing the previous state-of-the-art in terms of binary classification, this model can also be used to find the most semantically-similar past cases, working as case-based explanations for the decision.

9.3 Final Remarks and Future Work

As described above and detailed in the chapters of this thesis, several contributions were made, updating the state-of-the-art in a diversity of domains, both fundamental and applied. Nonetheless, the objective aesthetic evaluation of breast cancer treatments is far from being solved.

Part of the efforts to continue research on this topic will be related to the recently approved European project entitled "CINDERELLA: Clinical Validation of an AI-based approach to improve the shared decision-making process and outcomes in Breast Cancer Patients proposed for Locoregional treatment". The project's main goal is to visually simulate the aesthetic outcome of a certain surgery, requiring additional research efforts in terms of aesthetic evaluation of breast cancer treatments, medical image retrieval and explainable AI, topics intensely explored in the context of this thesis. Moreover, research efforts in transfer and multitask learning, and engineering efforts by developing web applications and cloud services are also necessary.

Even though all the machine learning topics addressed in this thesis were motivated by the aesthetic evaluation of breast cancer treatments, some research lines are more general and could be

further explored in wider contexts. Indeed, they were already explored in the biometrics context, as my collaborations with the VCMI's biometrics subgroup demonstrate [117; 143; 144].

9.4 Funding

This work was financed by the Portuguese science and technology foundation, Fundação para a Ciência e a Tecnologia (FCT) and co-financed by the European Social Fund through the North Regional Operational Programme (NORTE 2020), under the grant SFRH/BD/139468/2018.

References

[1] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans. innvestigate neural networks! *J. Mach. Learn. Res.*, 20(93):1–8, 2019.

[2] K. Antoniuk, V. Franc, and V. Hlavác. Interval insensitive loss for ordinal classification. In *Asian Conference on Machine Learning*, pages 189–204, 2015.

[3] S. Ayyachamy, V. Alex, M. Khened, and G. Krishnamurthi. Medical image retrieval using resnet-18. In *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*, volume 10954, page 1095410. International Society for Optics and Photonics, 2019.

[4] N. A. Azeez and C. V. der Vyver. Security and privacy issues in e-health cloud-based system: A comprehensive content analysis. *Egyptian Informatics Journal*, 20(2):97 – 108, 2019. ISSN 1110-8665. doi: https://doi.org/10.1016/j.eij.2018.12.001. URL http://www.sciencedirect.com/science/article/pii/S1110866517302797.

[5] J. Ba and R. Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems 27*, pages 2654–2662. Curran Associates, Inc., 2014.

[6] S. Baccianella, A. Esuli, and F. Sebastiani. Evaluation measures for ordinal regression. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 283–287, Nov 2009. doi: 10.1109/ISDA.2009.230.

[7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[8] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MÃžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

[9] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[10] A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, and C. Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12):1061–1070, 2021.

[11] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. *arXiv preprint arXiv:1704.05796*, 2017.

[12] G. F. Beadle, S. Come, I. Henderson, B. Silver, S. Hellman, and J. R. Harris. The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer. *International Journal of Radiation Oncology*Biology*Physics*, 10(11): 2131 – 2137, 1984. ISSN 0360-3016. doi: 10.1016/0360-3016(84)90213-X.

[13] V. Belagiannis and A. Zisserman. Recurrent human pose estimation. In *International Conference on Automatic Face and Gesture Recognition*. IEEE, 2017.

[14] L. Biganzoli, L. Marotti, C. D. Hart, L. Cataliotti, B. Cutuli, T. Kuhn, R. E. Mansel, A. Ponti, P. Poortmans, P. Regitnig, J. A. van der Hage, Y. Wengstrom, and M. Rosselli Del Turco. Quality indicators in breast cancer care: An update from the eusoma working group. *European Journal of Cancer*, 86:59 – 81, 2017. doi: 10.1016/j.ejca.2017.08.017.

[15] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph.*, 27(3):1–8, 2008. ISSN 0730-0301. doi: 10.1145/1360612.1360638. URL https://doi.org/10.1145/1360612.1360638.

[16] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 2018. doi: 10.3322/caac.21492.

[17] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[18] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[19] J. Bromley, J. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Sackinger, and R. Shah. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7:25, 08 1993. doi: 10.1142/S0218001493000339.

[20] Y. Cai, Y. Li, C. Qiu, J. Ma, and X. Gao. Medical image retrieval based on convolutional neural network and supervised hashing. *IEEE Access*, 7:51877–51885, 2019.

[21] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[22] J. S. Cardoso and M. J. Cardoso. Breast contour detection for the aesthetic evaluation of breast cancer conservative treatment. In *Springer Lecture Notes in Computer Science, Advances in Soft Computing 45, Computer Recognition Systems 2 CORES 2007: 5th International Conference on Computer Recognition Systems*, pages 518–525, October 2007. doi: 10.1007/978-3-540-75175-5_65.

[23] J. S. Cardoso and M. J. Cardoso. Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. *Artificial intelligence in medicine*, 40 (2):115–126, 2007.

[24] J. S. Cardoso and J. F. P. da Costa. Learning to classify ordinal data: the data replication method. *Journal of Machine Learning Research*, 8:1393–1429, 2007.

[25] J. S. Cardoso and R. Sousa. Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(8):1173–1195, 2011.

[26] J. S. Cardoso, I. Domingues, and H. P. Oliveira. Closed shortest path in the original coordinates with an application to breast cancer. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(01):1555002, 2015.

[27] J. S. Cardoso, W. Silva, and M. J. Cardoso. Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment. *The Breast*, 49:123–130, 2020.

[28] M. J. Cardoso, J. S. Cardoso, A. C. Santos, H. Barros, and M. C. Oliveira. Interobserver agreement and consensus over the esthetic evaluation of conservative treatment for breast cancer. *The Breast*, 15:52–57, february 2006. doi: 10.1016/j.breast.2005.04.013.

[29] M. J. Cardoso, J. S. Cardoso, N. Amaral, I. Azevedo, L. Barreau, M. Bernardo, D. Christie, S. Costa, F. Fitzal, J. L. Fougo, J. Johansen, D. Macmillan, M. P. Mano, L. Regolo, J. Rosa, L. F. Teixeira, C. Vrieling, and J. Zgajnar. Turning subjective into objective: The bcct.core software for evaluation of cosmetic results in breast cancer conservative treatment. *Breast*, 16:456–461, 2007. doi: 10.1016/j.breast.2007.05.002.

[30] M. J. Cardoso, A. Magalhaes, T. Almeida, S. Costa, C. Vrieling, D. Christie, J. Johansen, and J. S. Cardoso. Is face-only photographic view enough for the aesthetic evaluation of breast cancer conservative treatment? *Breast Cancer Research and Treatment*, 112:565–568, 2008. doi: 10.1007/s10549-008-9896-5.

[31] M. J. Cardoso, J. S. Cardoso, C. Vrieling, D. Macmillan, D. Rainsbury, J. Heil, E. Hau, and M. Keshtgar. Recommendations for the aesthetic evaluation of breast cancer conservative treatment. *Breast Cancer Research and Treatment*, 135:629–637, 2012. doi: 10.1007/s10549-012-1978-8.

[32] M. J. Cardoso, H. P. Oliveira, and J. S. Cardoso. Assessing cosmetic results after breast conserving surgery. *Journal of Surgical Oncology*, 110(1):37–44, 2014. doi: 10.1002/jso.23596.

[33] M. J. Cardoso, J. S. Cardoso, H. P. Oliveira, and P. Gouveia. The breast cancer conservative treatment. cosmetic results - bcct.core - software for objective assessment of aesthetic outcome in breast cancer conservative treatment: a narrative review. *Computer Methods and Programs in Biomedicine*, 2016. doi: 10.1016/j.cmpb.2015.11.010.

[34] M. J. Cardoso, C. Vrieling, J. S. Cardoso, H. P. Oliveira, N. R. Williams, J. M. Dixon, the PICTURE Project Clinical Trial Team, and the PICTURE Project Delphi Panel. The value of 3d images in aesthetic evaluation of breast cancer conservative treatment. results from a prospective multicentric clinical trial. *The Breast*, 41:19–24, 2018. doi: 10.1016/j.breast.2018.06.008.

[35] Cardoso et al. Automatic breast contour detection in digital photographs. In *Proceedings of the First International Conference on Health Informatics*, pages 91–98, Funchal, Madeira, Portugal, 2008. SciTePress - Science and and Technology Publications. ISBN 978-989-8111-16-6. doi: 10.5220/0001039500910098.

[36] D. C. Castro, I. Walker, and B. Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

[37] H. Charfare, E. MacLatchie, C. Cordier, M. Bradley, C. Eadie, A. Byrtus, K. Burnet, D. Chapman, G. C. Wishart, and A. D. Purushotham. A comparison of different methods of assessing cosmetic outcome following breast-conserving surgery and factors influencing cosmetic outcome. *British Journal of Medical Practitioners*, 3, 2010.

[38] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–32, 2021.

[39] C. Chen and A. Ross. An explainable attention-guided iris presentation attack detector. In *WACV (Workshops)*, pages 97–106, 2021.

[40] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[41] J. Chen, J. Konrad, and P. Ishwar. Vgan-based image representation learning for privacy-preserving facial expression recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), CV-COPS Workshop*, 2018.

[42] Z. Chen, Y. Bei, and C. Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

[43] D. Cho, J. H. Lee, and I. H. Suh. Cleanir: Controllable attribute-preserving natural identity remover. *Applied Sciences*, 10(3):1120, Feb 2020. ISSN 2076-3417. doi: 10.3390/app10031120. URL http://dx.doi.org/10.3390/app10031120.

[44] F. Chollet *et al.* Keras. https://keras.io, 2015.

[45] D. Christie, M. O'brien, J. Christie, T. Kron, S. Ferguson, C. Hamilton, and J. Denham. A comparison of methods of cosmetic assessment in breast conservation treatment. *The Breast*, 5(5):358–367, 1996.

[46] L. Z. Cordova, D. J. Hunter-Smith, and W. M. Rozen. Patient reported outcome measures (proms) following mastectomy with breast reconstruction or without reconstruction: a systematic review. *Gland Surgery*, 8(4), 2019.

[47] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2009.05.016.

[48] J. Costa and J. S. Cardoso. oAdaBoost: An AdaBoost variant for Ordinal Classification. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pages 68–76, 2015.

[49] R. Cruz, K. Fernandes, J. S. Cardoso, and J. F. P. Costa. Tackling class imbalance with ranking. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2182–2187, 2016.

[50] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero, and P. Gutiérrez. Metrics to guide a multi-objective evolutionary algorithm for ordinal classification. *Neurocomputing*, 135:21 – 31, 2014. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2013.05.058. URL http://www.sciencedirect.com/science/article/pii/S0925231213011399.

[51] J. F. P. da Costa, H. Alonso, and J. S. Cardoso. The unimodal model for the classification of ordinal data. *Neural Networks*, 21:78–91, 2008.

[52] P. Das and A. Neelima. An overview of approaches for content-based medical image retrieval. *International journal of multimedia information retrieval*, 6(4):271–280, 2017.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[54] C. E. DeSantis, F. Bray, J. Ferlay, J. Lortet-Tieulent, B. O. Anderson, and A. Jemal. International variation in female breast cancer incidence and mortality rates. *Cancer Epidemiology and Prevention Biomarkers*, 2015.

[55] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning, 2017.

[56] A. Dosovitskiy and T. Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4829–4837, 2016.

[57] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, Aug. 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

[58] B. Eiben, R. Lacher, V. Vavourakis, J. H. Hipwell, D. Stoyanov, N. R. Williams, J. Sabczynski, T. Bulow, D. Kutra, K. Meetz, S. Young, H. Barschdorf, H. P. Oliveira, J. S. Cardoso, J. P. Monteiro, H. Zolfagharnasab, R. Sinkus, P. Gouveia, G.-J. Liefers, B. Molenkamp, C. J. van de Velde, D. J. Hawkes, M. J. Cardoso, and M. Keshtgar. Breast conserving surgery outcome prediction: A patient-specific, integrated multi-modal imaging and mechano-biological modelling framework. In *Proceedings of the 13th International Workshop on Breast Imaging (IWBI)*, pages 274–281, 2016. URL http://www.inescporto.pt/~jsc/publications/conferences/2016EibenIWBI.pdf.

[59] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, 2017. doi: 10.1038/nature21056.

[60] K. Fernandes and J. S. Cardoso. Hypothesis transfer learning based on structural model similarity. *Neural Computing and Applications*, 31(8):3417–3430, 2019.

[61] K. Fernandes, J. S. Cardoso, and B. Astrup. A deep learning approach for the forensic evaluation of sexual assault. *Pattern Analysis and Applications*, 2018.

[62] FICO. Explainable machine learning challenge. https://community.fico.com/s/explainable-machine-learning-challenge. 2018.

[63] FICO. Introduction to Scorecard for FICO Model Builder. Technical report, 2006.

[64] F. Fitzal, W. Krois, H. Trischler, L. Wutzel, O. Riedl, U. Kühbelböck, B. Wintersteiner, M. J. Cardoso, P. Dubsky, M. Gnant, *et al.* The use of a breast symmetry index for objective evaluation of breast cosmesis. *The breast*, 16(4):429–435, 2007.

[65] E. Frank and M. Hall. A simple approach to ordinal classification. In *Machine Learning: ECML 2001*, pages 145–156. Springer, 2001.

[66] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent. Large-scale privacy protection in google street view. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2373–2380, 2009.

[67] K. Gan, A. Li, Z. Lipton, and S. Tayur. Causal inference with selectively deconfounded data. In *International Conference on Artificial Intelligence and Statistics*, pages 2791–2799. PMLR, 2021.

[68] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2 (11):665–673, 2020.

[69] T. Gonçalves. Deep Aesthetic Assessment of Breast Cancer Surgery Outcomes. Master's thesis, Universidade do Porto, Portugal, 2019.

[70] T. Gonçalves, W. Silva, and J. Cardoso. Deep aesthetic assessment of breast cancer surgery outcomes. In *Mediterranean Conference on Medical and Biological Engineering and Computing*, pages 1967–1983. Springer, 2019.

[71] T. Gonçalves, W. Silva, M. J. Cardoso, and J. S. Cardoso. A novel approach to keypoint detection for the aesthetic evaluation of breast cancer surgery outcomes. *Health and Technology*, 10(4):891–903, 2020.

[72] M. Gong, J. Liu, H. Li, Y. Xie, and Z. Tang. Disentangled representation learning for multiple attributes preserving face deidentification. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2020. doi: 10.1109/TNNLS.2020.3027617.

[73] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[74] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[75] R. K. Grace, R. Manimegalai, and S. S. Kumar. Medical image retrieval system in grid using hadoop framework. In *2014 international conference on computational science and computational intelligence*, volume 1, pages 144–148. IEEE, 2014.

[76] R. Gross, E. Airoldi, B. Malin, and L. Sweeney. Integrating utility into face de-identification. In *Privacy Enhancing Technologies*, pages 227–242, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[77] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5769–5779, 2017.

[78] M. Gupta, A. Cotter, J. Pfeifer, K. Voevodski, K. Canini, A. Mangylov, W. Moczydlowski, and A. van Esbroeck. Monotonic calibrated interpolated look-up tables. *Journal Machine Learning Research (JMLR)*, 2016. URL http://jmlr.org/papers/volume17/15-243/15-243.pdf.

[79] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez. Ordinal Regression Methods: Survey and Experimental Study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, Jan 2016.

[80] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.

[81] M. H. Haloua, N. M. A. Krekel, G. J. A. Jacobs, B. Zonderhuis, M.-B. Bouman, M. E. Buncamper, F. B. Niessen, H. A. H. Winters, C. Terwee, S. Meijer, and M. P. van den Tol. Cosmetic outcome assessment following breast-conserving therapy: A comparison between bcct.core software and panel evaluation. *International journal of breast cancer*, 2014:716860–716860, 2014.

[82] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–318, 1986.

[83] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[84] C. Hill-Kayser, C. Vachani, G. Di Lullo, and J. Metz. Cosmetic outcomes and complications reported by patients having undergone breast-conserving treatment. *Int J Radiat Oncol Biol Phys*, 83, 2012. doi: 10.1016/j.ijrobp.2011.08.013.

[85] J. Hofmanninger and G. Langs. Mapping visual features to semantic profiles for retrieval in medical imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 457–465, 2015.

[86] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[87] G. Kaissis, M. Makowski, D. Rückert, and R. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2, 06 2020. doi: 10.1038/s42256-020-0186-1.

[88] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93, June 1938. doi: 10.2307/2332226.

[89] B. Kim, C. Rudin, and J. Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 1952–1960, Cambridge, MA, USA, 2014. MIT Press.

[90] B. Kim, J. A. Shah, and F. Doshi-Velez. Mind the gap: A generative approach to interpretable feature selection and extraction. 2015.

[91] B. Kim, R. Khanna, and O. O. Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

[92] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, *et al.* Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[93] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[94] D. P. Kingma and M. Welling. Auto-encoding variational bayes. 2014.

[95] J. Konecny, B. McMahan, and D. Ramage. Federated optimization:distributed optimization beyond the datacenter. In *Proceedings of 8th NIPS Workshop on Optimization for Machine Learning*, 2015.

[96] M. Lagendijk, L. S. E. van Egdom, C. Richel, N. van Leeuwen, C. Verhoef, H. F. Lingsma, and L. B. Koppert. Patient reported outcome measures in breast cancer patients. *European Journal of Surgical Oncology*, 44(7):963–968, Jul 2018. ISSN 0748-7983.

[97] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14539. URL http://www.nature.com/articles/nature14539.

[98] C. S. Lee, P. G. Nagy, S. J. Weaver, and D. E. Newman-Toker. Cognitive and system factors contributing to diagnostic errors in radiology. *American Journal of Roentgenology*, 201 (3):611–617, Sep 2013. ISSN 0361-803X. doi: 10.2214/AJR.12.10375. URL https://doi.org/10.2214/AJR.12.10375.

[99] J. Lee, E. Kim, G. P. Reece, M. A. Crosby, E. K. Beahm, and M. K. Markey. Automated calculation of ptosis on lateral clinical photographs. *Journal of evaluation in clinical practice*, 21(5):900–910, 2015. doi: 10.1111/jep.12397.

[100] O. Li, H. Liu, C. Chen, and C. Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *AAAI*, 2018.

[101] Z. Li, X. Zhang, H. Müller, and S. Zhang. Large-scale retrieval for medical image analytics: A comprehensive review. *Medical image analysis*, 43:66–84, 2018.

[102] E. V. Limbergen, E. van der Schueren, and K. V. Tongelen. Cosmetic evaluation of breast conserving treatment for mammary cancer. 1. proposal of a quantitative scoring system. *Radiotherapy and Oncology*, 16(3):159 – 167, 1989. ISSN 0167-8140. doi: 10.1016/0167-8140(89)90016-9.

[103] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques - SIGGRAPH '87*, pages 163–169, Not Known, 1987. ACM Press. ISBN 978-0-89791-227-3. doi: 10.1145/37401.37422. URL http://portal.acm.org/citation.cfm?doid=37401.37422.

[104] L. Ma, X. Liu, Y. Gao, Y. Zhao, X. Zhao, and C. Zhou. A new method of content based medical image retrieval and its applications to ct imaging sign retrieval. *Journal of biomedical informatics*, 66:148–158, 2017.

[105] A. Mbilinyi and H. Schuldt. Retrieving chest x-rays for differential diagnosis: A deep metric learning approach. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE, 2021.

[106] R. J. McDonald, K. M. Schwartz, L. J. Eckel, F. E. Diehn, C. H. Hunt, B. J. Bartholmai, B. J. Erickson, and D. F. Kallmes. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic Radiology*, 22(9):1191–1198, Sep 2015. ISSN 1076-6332. doi: 10.1016/j.acra.2015.05.007. URL https://doi.org/10.1016/j.acra.2015.05.007.

[107] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira. PH2 - a dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5437–5440, July 2013.

[108] R. Merie, L. Browne, J. S. Cardoso, M. J. Cardoso, Y. Chin, C. Catherine, P. Graham, A. Szwajcer, and E. Hau. A proposal for a gold standard for cosmetic evaluation after breast conserving therapy: results from the st george and wollongong breast boost trial. *Journal of Medical Imaging and Radiation Oncology*, 2017. doi: 10.1111/1754-9485.12645.

[109] A. E. Minarno, K. M. Ghufron, T. S. Sabrila, L. Husniah, and F. D. S. Sumadi. Cnn based autoencoder application in breast cancer image retrieval. In *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 29–34. IEEE, 2021.

[110] S. S. Mishra, B. Mandal, and N. B. Puhan. Multi-level dual-attention based cnn for macular optical coherence tomography classification. *IEEE Signal Processing Letters*, 26(12):1793–1797, 2019.

[111] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[112] H. Montenegro. A privacy-preserving framework for case-based interpretability in machine learning. Master's thesis, Universidade do Porto, Portugal, 2021.

[113] H. Montenegro, W. Silva, and J. S. Cardoso. Privacy-preserving generative adversarial network for case-based explainability in medical image analysis. *IEEE Access*, 9:148037–148047, 2021.

[114] H. Montenegro, W. Silva, and J. S. Cardoso. Towards privacy-preserving explanations in medical image analysis. In *1st Workshop on Interpretable Machine Learning in Healthcare at ICML (IMLH 2021)*, 2021.

[115] H. Montenegro, W. Silva, A. Gaudio, M. Fredrikson, A. Smailagic, and J. S. Cardoso. Privacy-preserving case-based explanations: Enabling visual interpretability by protecting privacy. *IEEE Access*, 2022.

[116] B. Mustafa, A. Loh, J. Freyberg, P. MacWilliams, M. Wilson, S. M. McKinney, M. Sieniek, J. Winkens, Y. Liu, P. Bui, *et al.* Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913*, 2021.

[117] P. C. Neto, T. Gonçalves, J. R. Pinto, W. Silva, A. F. Sequeira, A. Ross, and J. S. Cardoso. Explainable biometrics in the age of deep learning. *arXiv preprint arXiv:2208.09500*, 2022.

[118] E. M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005. doi: 10.1109/TKDE.2005.32.

[119] M. Noguchi, Y. Saito, Y. Mizukami, A. Nonomura, N. Ohta, N. Koyasaki, T. Taniya, and I. Miyazaki. Breast deformity, its correction, and assessment of breast conserving surgery. *Breast Cancer Research and Treatment*, 18(2):111–118, 1991. doi: 10.1007/BF01980973.

[120] J. Nowaková, M. Prílepok, and V. Snášel. Medical image retrieval using vector quantization and fuzzy s-tree. *Journal of medical systems*, 41(2):1–16, 2017.

[121] W. Oleszkiewicz, T. Wlodarczyk, K. Piczak, T. Trzcinski, P. Kairouz, and R. Rajagopal. Siamese generative adversarial privatizer for biometric data. 04 2018.

[122] H. P. Oliveira, J. S. Cardoso, A. Magalhaes, and M. J. Cardoso. Simultaneous detection of prominent points on breast cancer conservative treatment images. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 2841–2844, 2012. doi: 10.1109/ICIP.2012.6467491.

[123] H. P. Oliveira, J. S. Cardoso, A. Magalhaes, and M. J. Cardoso. Methods for the aesthetic evaluation of breast cancer conservation treatment: A technological review. *Current Medical Imaging Reviews*, 9(1):32–46, 2013. doi: 10.2174/1573405611309010006.

[124] H. P. Oliveira, J. S. Cardoso, A. Magalhaes, and M. J. Cardoso. A 3d low-cost solution for the aesthetic evaluation of breast cancer conservative treatment. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 2(2):90–106, 2014. doi: 10.1080/21681163.2013.858403.

[125] D. Ovadia. Ai will help cinderella to see herself in the mirror. URL https://cancerworld.net/ai-will-help-cinderella-to-see-herself-in-the-mirror/.

[126] M. Owais, M. Arsalan, J. Choi, and K. R. Park. Effective diagnosis and treatment through content-based medical image retrieval (cbmir) by using artificial intelligence. *Journal of clinical medicine*, 8(4):462, 2019.

[127] H. Pashler, M. McDaniel, D. Rohrer, and R. Bjork. Learning styles: Concepts and evidence. *Psychological Science in the Public Interest*, 9(3):105–119, 2008.

[128] R. D. Pezner, M. P. Patterson, L. Hill, N. Vora, K. R. Desai, J. O. Archambeau, and J. A. Lipsett. Breast retraction assessment: an objective evaluation of cosmetic results of patients treated conservatively for breast cancer. *International Journal of Radiation Oncology*Biology*Physics*, 11(3):575–578, 1985. ISSN 0360-3016. doi: 10.1016/0360-3016(85)90190-7.

[129] B. Pierquin, J. Huart, M. Raynal, Y. Otmezguine, E. Calitchi, J.-J. Mazeron, G. Ganem, J.-P. L. Bourgeois, G. Marinello, M. Julien, B. Brun, and F. Feuilhade. Conservative treatment for breast cancer: long-term results (15 years). *Radiotherapy and Oncology*, 20(1):16–23, 1991. doi: 10.1016/0167-8140(91)90107-R.

[130] M. Pérez-Ortiz, P. A. Gutiérrez, C. Hervás-Martínez, and X. Yao. Graph-Based Approaches for Over-Sampling in the Context of Ordinal Regression. *IEEE Transactions on Knowledge and Data Engineering*, 27(5):1233–1245, May 2015.

[131] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid. Medical image retrieval using deep convolutional neural network. *Neurocomputing*, 266:8–20, 2017.

[132] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1530–1538. JMLR.org, 2015.

[133] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[134] R. L. Rivest. Learning decision lists. *Machine Learning*, 2(3):229–246, Nov. 1987. ISSN 0885-6125, 1573-0565. doi: 10.1007/BF00058680. URL http://link.springer.com/10.1007/BF00058680.

[135] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[136] M. Rosselli Del Turco, A. Ponti, U. Bick, L. Biganzoli, G. Cserni, B. Cutuli, T. Decker, M. Dietel, O. Gentilini, T. Kuehn, M. Mano, P. Mantellini, L. Marotti, P. Poortmans, F. Rank, H. Roe, E. Scaffidi, J. van der Hage, G. Viale, C. Wells, M. Welnicka-Jaskiewicz, Y. Wengstom, and L. Cataliotti. Quality indicators in breast cancer care. *European Journal of Cancer*, 46(13):2344 – 2356, 2010. doi: 10.1016/j.ejca.2010.06.119.

[137] H. R. Roth, K. Chang, P. Singh, N. Neumark, W. Li, V. Gupta, S. Gupta, L. Qu, A. Ihsani, B. C. Bizzo, and et al. Federated learning for breast density classification: A real-world implementation. *Lecture Notes in Computer Science*, page 181–191, 2020. ISSN 1611-3349. doi: 10.1007/978-3-030-60548-3_18. URL http://dx.doi.org/10.1007/978-3-030-60548-3_18.

[138] C. Rudin and B. Ustun. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, 48(5):449–466, 2018.

[139] V. Sacchini, A. Luini, S. Tana, L. Lozza, V. Galimberti, M. Merson, R. Agresti, P. Veronesi, and M. Greco. Quantitative and qualitative cosmetic evaluation after conservative treatment for breast cancer. *European Journal of Cancer and Clinical Oncology*, 27(11):1395 – 1400, 1991. ISSN 0277-5379. doi: 10.1016/0277-5379(91)90019-A.

[140] C. G. Schwarz, W. K. Kremers, T. M. Therneau, R. R. Sharp, J. L. Gunter, P. Vemuri, A. Arani, A. J. Spychalla, K. Kantarci, D. S. Knopman, R. C. Petersen, and J. Jack, C. R. Identification of anonymous mri research participants with face-recognition software. In *The New England journal of medicine*, volume 381, pages 1684–1686, 2019. doi: 10.1056/NEJMc1908881.

[141] K. Seetharaman and S. Sathiamoorthy. A unified learning framework for content based medical image retrieval using a statistical model. *Journal of King Saud University-Computer and Information Sciences*, 28(1):110–124, 2016.

[142] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[143] A. F. Sequeira, W. Silva, J. R. Pinto, T. Gonçalves, and J. S. Cardoso. Interpretable biometrics: Should we rethink how presentation attack detection is evaluated? In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2020.

[144] A. F. Sequeira, T. Gonçalves, W. Silva, J. R. Pinto, and J. S. Cardoso. An exploratory study of interpretability for face presentation attack detection. *IET Biometrics*, 10(4):441–455, 2021.

[145] M. Sheller, B. Edwards, G. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. Colen, and S. Bakas. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10, 07 2020. doi: 10.1038/s41598-020-69250-1.

[146] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[147] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

[148] J. Sill. Monotonic networks. In *Advances in neural information processing systems*, pages 661–667, 1998.

[149] P. F. Silva and J. S. Cardoso. Differential scorecards for binary and ordinal data. *Intelligent data analysis*, 19(6):1391–1408, 2015.

[150] W. Silva, K. Fernandes, M. J. Cardoso, and J. S. Cardoso. Towards complementary explanations using deep neural networks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 133–140. Springer, 2018.

[151] W. Silva, J. R. Pinto, and J. S. Cardoso. A uniform performance index for ordinal classification with imbalanced classes. In *2018 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2018.

[152] W. Silva, E. Castro, M. J. Cardoso, F. Fitzal, and J. S. Cardoso. Deep keypoint detection for the aesthetic evaluation of breast cancer surgery outcomes. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1082–1086. IEEE, 2019.

[153] W. Silva, K. Fernandes, and J. S. Cardoso. How to produce complementary explanations using an ensemble model. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[154] W. Silva, A. Poellinger, J. S. Cardoso, and M. Reyes. Interpretability-guided content-based medical image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–314. Springer, 2020.

[155] W. Silva, M. Carvalho, C. Mavioso, M. J. Cardoso, and J. S. Cardoso. Deep aesthetic assessment and retrieval of breast cancer treatment outcomes. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 108–118. Springer, 2022.

[156] W. Silva, T. Gonçalves, K. Härmä, E. Schröder, V. C. Obmann, M. C. Barroso, A. Poellinger, M. Reyes, and J. S. Cardoso. Computer-aided diagnosis through medical image retrieval in radiology. *Scientific Reports*, 12(20732), 2022.

[157] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[158] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014.

[159] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[160] E. Snell. Utilizing cloud computing for stronger healthcare data security, Nov 2019. URL https://healthitsecurity.com/features/utilizing-cloud-computing-for-stronger-healthcare-data-security.

[161] R. Sousa, J. S. Cardoso, J. P. Da Costa, and M. J. Cardoso. Breast contour detection with shape priors. In *2008 15th IEEE International Conference on Image Processing*, pages 1440–1443. IEEE, 2008.

[162] C. Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[163] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[164] M. Srinivas, R. R. Naidu, C. S. Sastry, and C. K. Mohan. Content based medical image retrieval using dictionary learning. *Neurocomputing*, 168:880–895, 2015.

[165] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern recognition*, 40(12):3358–3378, 2007.

[166] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[167] H. R. Tizhoosh. Barcode annotations for medical image retrieval: A preliminary investigation. In *2015 IEEE international conference on image processing (ICIP)*, pages 818–822. IEEE, 2015.

[168] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Assessment of iris recognition reliability for eyes affected by ocular pathologies. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, 2015. doi: 10.1109/BTAS.2015.7358747.

[169] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Implications of ocular pathologies for iris recognition reliability. *Image and Vision Computing*, 58:158–167, 2017. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2016.08.001. URL https://www.sciencedirect.com/science/article/pii/S0262885616301251.

[170] L. I. Tsouskas and I. S. Fentiman. Breast compliance: a new method for evaluation of cosmetic outcome after conservative treatment of early breast cancer. *Breast Cancer Research and Treatment*, 15:185–190, 1990. doi: 10.1007/BF01806355.

[171] J. W. Tukey. *Exploratory data analysis*, volume 2. Reading, Mass., 1977.

[172] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. scikit-image: image processing in Python. *PeerJ*, 2:e453, June 2014. ISSN 2167-8359. doi: 10.7717/peerj.453. URL https://peerj.com/articles/453.

[173] K. R. Varshney, J. C. Rasmussen, A. Mojsilovic, M. Singh, and J. M. DiMicco. Interactive visual salesforce analytics. In *ICIS*, 2012.

[174] J. Vosshenrich, P. Brantner, J. Cyriac, D. T. Boll, E. M. Merkle, and T. Heye. Quantifying radiology resident fatigue: Analysis of preliminary reports. *Radiology*, 298(3):632–639, 2021.

[175] C. Vrieling, L. Collette, E. Bartelink, J. H. Borger, S. J. Brenninkmeyer, J.-C. Horiot, M. Pierart, P. M. Poortmans, H. Struikmans, E. Van der Schueren, J. A. Van Dongen, E. Van Limbergen, and H. Bartelink. Validation of the methods of cosmetic assessment after breast-conserving therapy in the eortc “boost versus no boost” trial. *International Journal of Radiation Oncology * Biology * Physics*, 45(3):667–676, Oct 1999.

[176] F. Wang and D. M. Tax. Survey on the attention based rnn model and its applications in computer vision. *arXiv preprint arXiv:1601.06823*, 2016.

[177] T. Wang and C. Rudin. Causal rule sets for identifying subgroups with enhanced treatment effect. *arXiv preprint arXiv:1710.05426*, 2017.

[178] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille. A bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research*, 18(1):2357–2393, 2017.

[179] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

[180] J. Weatheritt, D. Rueckert, and R. Wolz. Transfer learning for brain segmentation: Pre-task selection and data limitations. In *Annual Conference on Medical Image Understanding and Analysis*, pages 118–130. Springer, 2020.

[181] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum. Automatic coronary calcium scoring in cardiac ct angiography using convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 589–596. Springer, 2015.

[182] Y. Wu, F. Yang, Y. Xu, and H. Ling. Privacy-protective-gan for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 34:47–60, 01 2019. doi: 10.1007/s11390-019-1898-8.

[183] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[184] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[185] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[186] L. Zhang, Z. Zhu, B. Yang, W. Liu, H. Zhu, and M. Zou. Medical image encryption and compression scheme using compressive sensing and pixel swapping based permutation approach. 2015. doi: 10.1155/2015/940638.

[187] Y. Zhuang, N. Jiang, Z. Wu, Q. Li, D. K. Chiu, and H. Hu. Efficient and robust large medical image retrieval in mobile cloud computing environment. *Information Sciences*, 263:60–86, 2014.

[188] A. Ziller, D. Usynin, R. Braren, M. Makowski, D. Rueckert, and G. Kaissis. Medical imaging deep learning with differential privacy. *Scientific Reports*, 11(1):1–8, 2021.