



D 2025

INTERPRETABLE MACHINE LEARNING AND ITS APPLICATION TO MEDICAL DECISION SUPPORT SYSTEMS

TIAGO FILIPE SOUSA GONÇALVES

TESE DE DOUTORAMENTO APRESENTADA

À FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO EM
ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Interpretable Machine Learning and its Application to Medical Decision Support Systems

Tiago Filipe Sousa Gonçalves

DOCTORAL THESIS



Doctoral Program in Electrical and Computer Engineering

Supervisor: Jaime dos Santos Cardoso

Friday 11th July, 2025

Interpretable Machine Learning and its Application to Medical Decision Support Systems

Tiago Filipe Sousa Gonçalves

Doctoral Program in Electrical and Computer Engineering

Approved by . . . :

President: Luís Miguel Pinho de Almeida, Full Professor, Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto, Portugal

Referee: Diana Mateus, Full Professor, École Centrale de Nantes, France

Referee: José Luis Alba Castro, Full Professor, Departamento de Teoría do sinal e comunicacións de la Escola De Enxeñaría De Telecomunicación de la Universidade de Vigo, España

Referee: Ana Catarina Fidalgo Barata, Assistant Professor, Departamento de Engenharia Electrotécnica e de Computadores do Instituto Superior Técnico da Universidade de Lisboa, Portugal

Referee: Luís Filipe Pinto de Almeida Teixeira, Associate Professor, Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto

Supervisor: Jaime dos Santos Cardoso, Full Professor, Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto, Portugal

Friday 11th July, 2025

Resumo

Os algoritmos de inteligência artificial permitiram aos computadores aprender diretamente a partir de dados para resolver um problema específico. Com o advento da aprendizagem profunda, o desempenho preditivo destes algoritmos aumentou, especialmente na área de imagem médica. Apesar de atingirem um desempenho preditivo que desafia o ser humano em várias tarefas, estes métodos ainda sofrem de um comportamento opaco devido à sua complexidade. Esta falta de transparência pode comprometer a adoção destas tecnologias em domínios que dependem de decisões de alto risco, em que o processo de raciocínio deve ser transparente e claro. Esta tese centra-se no domínio da saúde, no qual esta tecnologia tem alimentado sistemas de apoio à decisão médica para ajudar os profissionais de saúde nas decisões clínicas. Além disso, sabemos que esta comunidade valoriza a existência de explicações para justificar as suas decisões clínicas. Esta questão motivou a investigação em inteligência artificial explicável, que se centra em desvendar o funcionamento interno dos algoritmos de aprendizagem automática. Embora não se centre numa aplicação médica específica, esta tese baseou-se na avaliação estética dos resultados da cirurgia do cancro da mama e na patologia computacional para ilustrar algumas das metodologias propostas nos trabalhos que fazem contribuições aplicadas. No que diz respeito às contribuições fundamentais, a gama de casos de uso é diversificada, uma vez que este projeto visava abordar alguns dos atuais desafios em aberto no domínio da inteligência artificial explicável para a saúde.

Na primeira fase deste projeto, preocupamo-nos em aprender o que é relevante, estudando o papel dos mecanismos de atenção em algoritmos de aprendizagem profunda para a análise de imagem médica. Apresentamos três casos de estudo diferentes (classificação de imagem médica, pesquisa e obtenção de imagem médica e patologia computacional) em que os mecanismos de atenção desempenham um papel fundamental, procurando compreender o seu impacto no desempenho dos algoritmos e os seus potenciais benefícios para a interpretabilidade, robustez ou transparência. Uma vez que as decisões médicas se baseiam frequentemente em fontes de dados clínicos díspares, a segunda fase deste projeto centra-se em estratégias de aprendizagem automática multimodal que possam tirar partida da informação de vários tipos de dados para uma decisão clínica. Apresentamos três casos de estudo (pesquisa e obtenção multimodal de imagem médica, pesquisa e obtenção multimodal de imagem médica orientadas por mecanismos de atenção e patologia computacional multimodal) em que as estratégias multimodais beneficiam (potencialmente) o poder preditivo dos algoritmos, contribuindo indiretamente para a sua robustez e transparência. Dado que as aplicações em saúde também exigem decisões de alto risco, concordamos que a comunidade deve avançar para o desenvolvimento de algoritmos de aprendizagem automática inerentemente interpretáveis. Assim, na terceira fase deste projeto, procurámos compreender como construir algoritmos que o ser humano consiga compreender e analisámos três casos de estudo que mostram como funcionam os algoritmos inerentemente interpretáveis e mostram as suas aplicações em tarefas médicas (classificação e pesquisa e obtenção de imagem médica com redes B-cos e estudo do potencial da aprendizagem prototípica para abordagens contrafactuais).

Esta tese narra uma história sobre a obtenção de algoritmos transparentes e interpretáveis para

aplicações no domínio da saúde. Discutimos a nossa jornada através de diferentes dimensões do domínio da inteligência artificial explicável, na qual explorámos diversas linhas de investigação e tentámos contribuir para resolver os desafios apresentados pelo estado da arte.

Palavras-chave: aprendizagem automática interpretável, aprendizagem automática multimodal, aprendizagem contrafactual, aprendizagem profunda, aprendizagem prototípica, classificação de imagens médicas, inteligência artificial, inteligência artificial explicável, patologia computacional, pesquisa e obtenção de imagens médicas, sistemas de apoio à decisão médica, visão computacional

Abstract

Artificial intelligence algorithms have enabled computers to learn directly from data to solve a specific problem. With the advent of deep learning, the performance of these algorithms has increased, especially in medical imaging cases. Despite their human-challenging predictive performance in several tasks, these methods still suffer from black-box behaviour due to their complexity. This lack of transparency may jeopardise the adoption of these technologies in fields that rely on high-stakes decisions, where the reasoning process must be transparent and clear. This thesis focuses on the healthcare community, where this technology has powered medical decision support systems to help health professionals in clinical decisions. Besides, this community values the existence of explanations to justify their clinical decisions. This issue motivated the research on explainable artificial intelligence, which focuses on unveiling the inner functioning of machine learning algorithms. Although not focused on a specific medical application, this thesis relied on the aesthetic assessment of breast cancer surgery outcomes and computational pathology to illustrate some of the proposed methodologies in the applied contributions. Concerning the fundamental contributions, the range of use cases is diverse, as this project aimed to tackle some of the current open challenges in explainable artificial intelligence for healthcare.

In the first phase of this project, our main focus lies in learning what is relevant by studying the role of attention mechanisms in deep learning algorithms for medical image analysis. We present three different case studies (medical image classification, medical image retrieval and computational pathology) where attention mechanisms play a role, seeking to understand their impact on the performance of the algorithms and their potential benefits to interpretability, robustness or transparency. As medical decisions often rely on disparate clinical data sources, the second phase of this project focuses on multi-modal machine learning strategies that could leverage the information of several data types towards a clinical decision. We showcase three case studies (multi-modal medical image retrieval, attention-driven multi-modal medical image retrieval and multi-modal computational pathology) where multi-modal strategies (potentially) benefit the predictive power of algorithms, indirectly contributing to their robustness and transparency. Given that healthcare applications also require high-stakes decisions, we agree that the community should move towards developing inherently interpretable machine learning algorithms. Hence, in the third phase of this project, we sought to understand how to build human-understandable algorithms and analyse three case studies that depict how inherently interpretable algorithms work and show their applications in medical tasks (B-cos networks for medical image classification and retrieval and dissect the potential of prototypical learning for counterfactual approaches).

This thesis depicts the narrative of achieving transparent and interpretable algorithms for healthcare applications. We discuss our journey through different dimensions of the realm of explainable artificial intelligence, in which we explored diverse lines of research and tried to make

contributions to the resolution of the challenges presented in the state-of-the-art.

Keywords: artificial intelligence, computational pathology, computer vision, counterfactual learning, deep learning, explainable artificial intelligence, interpretable machine learning, medical decision support systems, medical image classification, medical image retrieval, multi-modal machine learning, prototypical learning

Acknowledgements

If you are reading this, then I have successfully written, defended, and finished my doctoral thesis. Truth is, pursuing a PhD was never one of my plans, but life happens with all its coincidences and randomness. This journey started in 2020 and finished in 2025, meaning that remembering all the people and institutions that contributed to its success may prove to be even more difficult than the PhD itself, but let me try to acknowledge you all properly. Pardon me if I forget anyone or anything.

Firstly, I want to thank my supervisor, Professor Jaime S. Cardoso, for his guidance, support, friendship, wisdom and patience. In 2018, you gave me a chance to conduct my MSc dissertation under your supervision, and in 2019, for some reason (I cannot explain), you challenged me to pursue a PhD. After all these years, you never stopped believing in me and guiding me through the peculiar world of research. Even when my attention was split into several other tasks, you knew how to guide me on the right path. A good scientific advisor is extremely important for a successful PhD project. I am glad I had (will always have) a good one.

You cannot conduct your research if you do not have a good research group and infrastructure to support you. In my case, I had the Centre of Telecommunications and Multimedia (CTM) of the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC). Particularly, and more importantly, I had the Visual Computing and Machine Intelligence (VCMI) research group. Being part of VCMI was life-changing. I enjoyed working with outstanding leaders, researchers, colleagues, and friends. To all the incredible senior researchers I had the pleasure of knowing and working with, thank you for sharing your knowledge and leading with an outstanding example. To all the graduate researchers, visiting researchers, PhD students, MSc students, and BSc students I had the pleasure to meet along the way (with an honourable mention for the ones I had the privilege of supervising or co-supervising), thank you for teaching more than I will ever be capable of teaching you. We organised schools, conferences, workshops, games nights, talks and meetings. More than colleagues, I made some good friends. Being part of VCMI allowed me to travel to almost all the continents. For all these things, I am grateful. Also, you are the best at creating a PhD gift. It will not be easy to compete with that.

At INESC TEC, I contributed to the discussion on diversity and inclusion by being part of the Commission for Diversity and Inclusion. To all the beautiful colleagues I have met and worked with, thank you for sparking my mind to topics different from my PhD topic.

During my PhD studies, I had the opportunity to be a member of the Monitoring Committee of PDEEC (my doctoral program). Along with Professor Luís Almeida, Professor José Nuno Fidalgo and my colleague Gowhar, we had the opportunity to debate the program's status and find new solutions to the current problems. After two years, I feel we have done some good things, and I hope the following members keep improving.

I had the opportunity to spend some time in the United States of America with the Quantitative Translational Imaging in Medicine (QTIM) Lab at the Athinoula A. Martinos Center for Biomedical Imaging in Massachusetts. I want to thank Albert Kim and Christopher P. Bridge for answering an email from a motivated Portuguese researcher who wanted to go abroad for his PhD internship. I

hope these last two years of collaboration and friendship have been worthy of how you changed my life. To Dagoberto Pullido, thank you for all the in-house support and for showing me around. To all the folks I have met and the friends I gained, I hope we can meet again somewhen and somewhere in this big (but small) world. To Tracy Washington, my roommate (and now a dear friend), thank you for introducing me to your country and showing me the real things. To Tobi and Aspen, thank you for your friendship; I am glad Tracy took me to your Dia de Los Muertos dinner.

Besides Professor Jaime S. Cardoso, another person has dramatically influenced my academic life. I never thought this friendship would lead to my return to the Netherlands, even briefly. To Wilson Silva, my MSc co-supervisor, thank you for your friendship and all the academic adventures and projects. Your invitation to spend time as a visiting researcher at the Netherlands Cancer Institute and the University of Utrecht played a crucial role in my PhD. While in the Netherlands, I learned with experts, expanded my networking, and made good friends. Thank you for that.

I am from Grijó, a town in Vila Nova de Gaia. To all my friends from there, thank you for making me feel at home, even if I spent most of the time out of town.

To KIZUNA, my second family, who have accompanied me from a very young age and taught me essential values through martial arts, thank you for these years of training, challenges and stories. Once again, I am glad to have you to help me in my battles.

Once again, to a very special group of people entitled *Boémios com Classe*, which has been with me since high school and, hopefully, will be here until the end of times. Thank you for endless nights of board games, debates and crazy ideas. I am still sure that we will change the world.

To all the good friends I made at Universidade do Porto, Universidade Católica Portuguesa or Colégio Internato dos Carvalhos, who span from professors to peers. Each of you has touched my life somehow, and I hope you can still be here to change it more.

To Joana Vale Amaro de Sousa, who showed me that love is an essential dimension of our lives as human beings and taught me (still does) how to love and be loved. You became a fundamental part of my life, and you will always have a place in my heart, independent of what happens in the future.

To my dearest little friend, Maria Emília Campos (Mimi), who sadly passed away before we could celebrate the end of my doctoral path together. You have been around since I was a little kid. I will never forget your presence in my life. I wish we had more time to listen to your stories and adventures.

Last but certainly not least, I want to acknowledge the importance of my family in all the events of my life. My father, Carlos, my mother, Rosa, and my sister, Maria Beatriz, have been crucial in my development as a human being and citizen of the world. During my life, my family did all they could to grant me access to education, culture and sports. In a world where many have nothing, I am lucky to have so much, and I thank you for that. To my grandma Albertina, my grandma Lurdes, and my grandpa Alberto, who have always believed in the power of education in our lives, I hope this milestone proves you right, and I can make you proud. To my aunt Paula, my uncle Paulo and my cousin Joana, thanks for the friendship, companionship and support through these years. Time definitely flies, and we have no time to lose.

To you, dear reader, who came across this document somewhere in time and space, I hope you find this reading worthy of your attention.

Let's see where the randomness of the Universe leads me. Hopefully, we meet will meet again.

“I know not what tomorrow will bring.”

Fernando Pessoa

Contents

1	Introduction	1
1.1	Context and Motivation	1
1.2	United Nations Sustainable Development Goals	2
1.3	Objectives and Document Structure	2
1.4	Contributions	4
1.5	Publications	5
1.6	Involvement in the Scientific Community	10
1.6.1	Collaborations	10
1.6.2	Scientific Events	11
1.6.3	Scientific Projects	12
1.6.4	Supervision of MSc Dissertations	12
2	Background	15
2.1	Deep Learning	15
2.1.1	Origin and Fundamentals	15
2.1.2	Democratisation and Innovation	16
2.2	Explainable Artificial Intelligence	17
2.2.1	Context and Motivation	17
2.2.2	Definitions	18
2.2.3	From Post-model to Inherently Interpretable AI	19
3	Datasets	20
3.1	APTOS2019	20
3.2	CINDERELLA Retrieval	20
3.3	CUB-200-2011	21
3.4	DTU/Herlev Pap Smear Database	21
3.5	International Skin Imaging Collaboration	21
3.5.1	ISIC2018	22
3.5.2	ISIC2020	22
3.6	MIMIC-CXR	22
3.7	PAPILA	22
3.8	PH2	23
3.9	TCGA-BRCA	23
3.10	VinDr-CXR	23

4	Learning What is Relevant: The Role of Attention Mechanisms in Medical Data	25
4.1	Introduction	25
4.2	Background	27
4.2.1	Attention mechanisms for language, text, and speech	29
4.2.2	Attention mechanisms for computer vision	31
4.3	Literature on Attention Mechanisms for Medical Applications	34
4.3.1	Medical Image Classification	34
4.3.2	Medical Image Segmentation	37
4.3.3	Medical Report Understanding	41
4.3.4	Other Tasks	43
4.4	Attention Mechanisms for Medical Image Classification	45
4.4.1	Description and Contributions	45
4.4.2	Data	46
4.4.3	Methodology	46
4.4.4	Results	50
4.4.5	Discussion, Conclusions and Future Work	53
4.5	Attention Mechanisms for Medical Image Retrieval	58
4.5.1	Description and Contributions	58
4.5.2	Background on Medical Image Retrieval	60
4.5.3	Materials and Methods	61
4.5.4	Results and Discussion	67
4.5.5	Conclusions and Future Work	72
4.6	Attention Mechanisms for Computational Pathology	75
4.6.1	Description and Contributions	76
4.6.2	Background on Attention Mechanisms for Computational Pathology . . .	77
4.6.3	Methodology	78
4.6.4	Results and Discussion	79
4.6.5	Conclusions and Future Work	80
4.7	Summary and Conclusions	82
5	Learning from Disparate Sources of Data: Explaining Multi-modal Algorithms	84
5.1	Introduction	84
5.2	Background	85
5.3	Literature Review	87
5.4	Multi-modal Retrieval for the Aesthetic Assessment of Breast Cancer Surgery Outcomes	88
5.4.1	Description and Contributions	88
5.4.2	Methodology	90
5.4.3	Results and Discussion	93
5.4.4	Conclusions and Future Work	94
5.5	Attention-Driven Multi-modal Retrieval for the Aesthetic Assessment of Breast Cancer Surgery Outcomes	95
5.5.1	Description and Contributions	95
5.5.2	Methodology	97
5.5.3	Results and Discussion	97
5.5.4	Conclusions and Future Work	97
5.6	Multi-modal Approaches for Computational Pathology	98
5.6.1	Description and Contributions	98

5.6.2	Brief Review of Multi-modal Machine Learning for Computational Pathology	100
5.6.3	Methodology	100
5.6.4	Results and Discussion	101
5.6.5	Conclusions and Future Work	101
5.7	Summary and Conclusions	102
6	Learning in a Human Understandable Way: Inherently Interpretable Models	106
6.1	Introduction	106
6.2	Background	107
6.3	Literature Review	109
6.3.1	Counterfactual Reasoning	109
6.3.2	Prototype-based Deep Learning	114
6.4	B-cos Networks for Medical Image Classification	119
6.4.1	Description and Contributions	119
6.4.2	Methodology	120
6.4.3	Results and Discussion	122
6.4.4	Conclusions and Future Work	126
6.5	B-cos Networks for Medical Image Retrieval	126
6.5.1	Description and Contributions	126
6.5.2	Methodology	127
6.5.3	Results and Discussion	127
6.5.4	Conclusions and Future Work	127
6.6	Explaining Counterfactuals with Prototypes	129
6.6.1	Description and Contributions	129
6.6.2	Methodology	130
6.6.3	Results and Discussion	135
6.6.4	Conclusions and Future Work	136
6.7	Summary and Conclusions	144
7	Conclusion	146
7.1	Conclusions and Contributions	146
7.2	Final Remarks and Future Work	148
7.3	Funding	149
	References	150

List of Figures

1.1	The United Nations Sustainable Development Goals.	3
2.1	Explainable AI as a three-stage process: pre-model, in-model and post-model methods.	18
4.1	Block diagram of the general attention mechanism for language, text and speech, according to [1].	29
4.2	U-Net architecture with attention blocks, as proposed in [2]. Please note that rectangles represent the feature maps of the encoder and decoder, horizontal arrows represent the regular flow of information, vertical elbow arrows represent the skip connections, and squares represent the attention blocks.	38
4.3	Block diagram of the <i>Squeeze-and-Excitation</i> (SE) attention block, according to [3].	47
4.4	Block diagram of the <i>Convolutional Block Attention Module</i> (CBAM), according to [4].	48
4.5	Block diagram of the <i>Channel Attention Module</i> of the <i>Convolutional Block Attention Module</i> (CBAM), according to [4].	48
4.6	Block diagram of the <i>Spatial Attention Module</i> of the <i>Convolutional Block Attention Module</i> (CBAM), according to [4].	49
4.7	Predictive performance (i.e., accuracy) results of the different models on the AP-TOS2019 data set, using different percentages of training data.	51
4.8	Predictive performance (i.e., accuracy) results of the different models on the ISIC2020 data set, using different percentages of training data.	51
4.9	Predictive performance (i.e., accuracy) results of the different models on the MIMIC-CXR data set, using different percentages of training data.	52
4.10	Pleural Effusion test image (a) and the least (b) and most (c) 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.	59
4.11	Example of test cases and ranking annotation for pleural effusion condition (the top 4 is shown) performed by the radiologist. The numbers on top of the image represent the ranking position of the image in the catalogue (when compared to the query/test image). The green box means pleural effusion case (according to the dataset label).	62

4.12	Example of test cases and ranking annotation for potential pneumonia condition (the top 4 are shown) performed by the radiologist. The numbers on top of the image represent the ranking position of the image in the catalogue (when compared to the query/test image). The green box means pneumonia case (according to the dataset label), the yellow box means radiologist considers the case as potential pneumonia, and the dataset label is no pneumonia.	63
4.13	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 colours 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. The L2 distance is computed between the test image's latent features and each catalogue image's latent features.	64
4.14	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.	68
4.15	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).	69
4.16	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.	71
4.17	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 represents a potential pneumonia case, the red box represents no potential pneumonia, and the orange box represents a disagreement between R1 and the label, with R1 considering the case as potential pneumonia.	73
4.18	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).	74
4.19	Overview of the proposed pipeline.	77
4.20	Attention map obtained using the AM-SB architecture for a WSI predicted to have a high-degree of T-cell mediated cytotoxicity. Red zones and blue zones indicate high and low attention scores, respectively. On the right, we provide two exemplar patches: 1) the high-attention patch illustrates abundant tumour-infiltrating lymphocytes without tumour cells, which are suggestive of high immune activity, and 2) the low-attention region demonstrates areas of tumour necrosis and minimal lymphocytes, consistent with low immune activity.	81

5.1	Illustration of the training and retrieval procedure. The trainable unit is optimised using the triplet loss to promote retrieval according to clinicians' ranking during training. It is worth noting that for images, the trainable unit must incorporate feature extraction, while for tabular data, it can be as simple as Weighted Euclidean Distance (WED). During retrieval, we use the trainable unit to extract the features corresponding to a given query. Afterwards, we use the Euclidean distance (similarity criterion) to calculate the distance between the query and the samples to retrieve the most similar samples. If we aim to improve efficiency, we can store the extracted feature vector of a particular image in a feature database.	91
5.2	Overview of the proposed pipeline.	99
6.1	Overview of ProtoPNet architecture, according to [5]. The model receives an image and tries to identify the parts of that image that are more similar to some prototypical part of some class, generating a vector of similarity scores. Then, it uses a weighted combination of these scores in the classification layer to output the final prediction. All the descendant models derive from this approach.	116
6.2	Comparing explanations from Input $\times$ Gradient [6] (I $\times$ G), Grad-CAM [7] (GC) and B-cos on original (Base.) and B-cos networks. The colour map ranges from blue to red, blue(red) meaning less(more) relevant pixels. Herlev stands for DTU/Herlev Pap Smear Database.	124
6.3	Overview of our proposed framework to obtain image counterfactuals and the open problem of using prototypes to explain that counterfactual and justify the model's prediction. Given a query image, we perform model inference to get the vector of probabilities and then select the index of the second-highest value as the counterfactual label. Afterwards, we use the model to extract features of the query image and all the possible counterfactual images in the database and minimise the Euclidean distance between these possible pairs (i.e., the most similar counterfactual corresponds to the minimum distance). To explain the model's predictions, we retrieve the prototypes of the query and counterfactual images. . .	131
6.4	Example of our proposal using the CUB-200-2011 database and the ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective <i>top-k</i> (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for prototype and c stands for class	137
6.5	Example of our proposal using the CUB-200-2011 database and the Deformable ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective <i>top-k</i> (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for prototype and c stands for class	138
6.6	Example of our proposal using the PAPILA database and the ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective <i>top-k</i> (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for prototype and c stands for class	139

- 6.7 Example of our proposal using the PAPILA database and the Deformable ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective *top-k* (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for **prototype** and c stands for **class**. 140
- 6.8 Example of our proposal using the PH2 database and the ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective *top-k* (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for **prototype** and c stands for **class**. 141
- 6.9 Example of our proposal using the PH2 database and the Deformable ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective *top-k* (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for **prototype** and c stands for **class**. 142

List of Tables

3.1	This is a summary of the datasets used in this thesis. The dataset names, references, tasks, and data examples are included in this table. Note: AD - Attribute Detection, C - Classification, R - Retrieval, S - Segmentation, MC - Multi-class, ML - Multi-label.	24
4.1	Data splits and meaning of labels for each data set.	46
4.2	Data augmentation parameters used during the training step of all the phases. . .	50
4.3	Set of hyper-parameters used during the training step of all the phases. Please note that we present the batch size as a closed interval since we had to adapt its value during some of the training runs due to computational constraints (e.g., availability of GPU RAM).	50
4.4	Model complexity (i.e., the number of parameters) information for the models used in our experiments.	52
4.5	Example of LRP and DeepLIFT <i>post-hoc</i> saliency maps for an image of the APTOS2019 data set with the label 0 (normal) correctly classified as 0 (normal) by all models.	53
4.6	Example of LRP and DeepLIFT <i>post-hoc</i> saliency maps for an image of the APTOS2019 data set with the label 1 (diabetic retinopathy) correctly classified as 1 (diabetic retinopathy) by all models.	53
4.7	Example of LRP and DeepLIFT <i>post-hoc</i> saliency maps for an image of the APTOS2019 data set with the label 0 (normal) incorrectly classified as 1 (diabetic retinopathy) by all models.	54
4.8	Example of LRP and DeepLIFT <i>post-hoc</i> saliency maps for an image of the ISIC2020 data set with the label 0 (benign) correctly classified as 0 (benign) by all models.	54
4.9	Example of LRP and DeepLIFT <i>post-hoc</i> saliency maps for an image of the ISIC2020 data set with the label 1 (malign) correctly classified as 1 (malign) by all models.	54
4.10	Example of LRP and DeepLIFT <i>post-hoc</i> saliency maps for an image of the ISIC2020 data set with the label 0 (benign) incorrectly classified as 1 (malign) by all models.	55
4.11	Example of LRP and DeepLIFT <i>post-hoc</i> saliency maps for an image of the ISIC2020 data set with the label 1 (malign) incorrectly classified as 0 (benign) by all models.	55
4.12	Example of LRP and DeepLIFT <i>post-hoc</i> saliency maps for an image of the MIMIC-CXR data set with the label 0 (normal) correctly classified as 0 (normal) by all models.	55

4.13	Example of LRP and DeepLIFT <i>post-hoc</i> saliency maps for an image of the MIMIC-CXR data set with the label 1 (pleural effusion) correctly classified as 1 (pleural effusion) by all models.	56
4.14	Example of LRP and DeepLIFT <i>post-hoc</i> saliency maps for an image of the MIMIC-CXR data set with the label 0 (normal) incorrectly classified as 1 (pleural effusion) by all models.	56
4.15	Example of LRP and DeepLIFT <i>post-hoc</i> saliency maps for an image of the MIMIC-CXR data set with the label 1 (pleural effusion) incorrectly classified as 0 (benign) by all models.	56
4.16	AUROC results on the test set, obtained for all model architectures, using ResNet50- and PLIP-derived features. The best results per type of feature-extraction strategy are highlighted in bold. While the AM-SB and AM-MB model architectures outperform TransMIL prediction of all gene expression tasks, the increase in performance was marginal, suggesting that model architecture may not play a critical role in the classification of gene expression.	80
4.17	AUROC results on the test set, obtained for the best model of each feature set. The best-performing feature-extraction strategy is highlighted in bold. Models trained with PLIP-derived features generally perform better than ResNet-derived features, suggesting that using a feature extractor pre-trained on H&E-related data may improve performance.	81
5.1	Retrieval performance metrics for various models categorised by aesthetic quality. Each entry is of the form (Excellent/Good, Fair/Poor). Best results, per model, highlighted in bold.	95
5.2	Retrieval performance metrics for attention-driven models, categorised by aesthetic quality. Each entry is of the form (Excellent/Good, Fair/Poor). Best results, per model, highlighted in bold.	98
5.3	AUROC results on the test set, obtained for the best model of each feature set. The best-performing feature-extraction strategy is highlighted in bold. Models trained with CONCH-derived features generally perform better than ResNet-derived features. When comparing with PLIP-derived features, we also observe that models with CONCH-derived features perform better in most tasks (6 out of 10), suggesting that replacing PLIP with CONCH is a good option regarding the feature extractor.	102
5.4	AUROC results on the test set, obtained for the different clinical-only data machine learning models. The best results are highlighted in bold. Each column presents a specific set of clinical features (selected based on discussions with clinicians). Note: We use (C) for categorical and (1H) for one-hot encoding.	103
5.5	AUROC results on the test set, obtained for the different multi-modal MCAT models. The best results are highlighted in bold. Each column presents the results of the WSI features (CONCH) combined with a specific set of clinical features (selected based on discussions with clinicians). The baseline results correspond to the first column.	104
6.1	Approaches to counterfactual generation through time.	115
6.2	ProtoPNet-based models that resulted from small increments, integrations, or different applications.	117

6.3	Results on model performance and xAI quality assessment obtained for each model across several datasets. For all models we show the results for the best xAI method. For B-cos networks we also show results for the intrinsic B-cos explanations. We compare against the following xAI methods: G - Grad, GC - Grad-CAM [7], I×G - Input×Gradient [6], IG - Integrated Gradients [8]. For each metric ↑ (↓) means higher (lower) is better. Underlined numbers represent the best result for non-B-cos models on a given dataset, while bold numbers highlight the best B-cos results. Herlev stands for DTU/Herlev Pap Smear Database, DN121 stands for DenseNet121 and RN18 IN corresponds to an ImageNet trained ResNet18. . . .	125
6.4	Retrieval performance metrics for B-cos networks (and their original counterparts), categorised by aesthetic quality. Each entry is of the form (Excellent/Good, Fair/Poor). Best results, per model, highlighted in bold. Models with a (P) refer to initialisation with pre-trained weights on ImageNet, and with a (PL) refer to initialisation with pre-trained weights on ImageNet and fine-tuned for a long number of epochs.	128
6.5	Accuracy results on the test set of each database. Best values per backbone highlighted in bold.	135
6.6	Average number of class-identity prototypes per correctly classified image of each database. Best values per backbone highlighted in bold.	136
6.7	Label coherence of the prototypes among backbone models across architectures on each database. Best value highlighted in bold.	136
6.8	Label coherence of the counterfactuals among backbone models across architectures on each database. Best values highlighted in bold.	137

Abbreviations and Symbols

This part presents the abbreviations and symbols used in this thesis, divided by Chapter. Please note that sometimes, we repeat the definition of an abbreviation or symbol in multiple Chapters to ease the reader's comprehension and avoid the need to go backwards or forward in the thesis to recall that information.

Chapter 1

AI	Artificial Intelligence
INVICTA	INvicta school of VISION, Computational intelligence, and patTern Analysis
Martinos	Athinoula A. Martinos Center for Biomedical Imaging
MGH	Massachusetts General Hospital
MICCAI	Medical Image Computing and Computer Assisted Intervention
MIT	Massachusetts Institute of Technology
NKI	Netherlands Cancer Institute
QTIM	Quantitative Translational Imaging in Medicine
RECPAD	Portuguese Conference on Pattern Recognition
SDG	Sustainable Development Goal
UN	United Nations
UNIFESP	Federal University of São Paulo
VCMI	Visual Computing and Machine Intelligence
VISUM	VISION Understanding and Machine intelligence Summer School
xAI	Explainable Artificial Intelligence
xAI4Biometrics	Workshop on Explainable & Interpretable Artificial Intelligence for Biometrics

Chapter 2

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Network
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
Grad-CAM	Gradient Class Activation Mapping
LRP	Layer-wise Relevance Propagation
LSTM	Long Short-Term Memory Network

RNN	Recurrent Neural Network
VAE	Variational Auto-Encoder
ViT	Vision Transformer
xAI	Explainable Artificial Intelligence

Chapter 3

APTOS	Asia Pacific Tele-Ophthalmology Society
COPD	Chronic Obstructive Pulmonary Disease
CXR	Chest X-Ray Radiograph
DICOM	Digital Imaging and Communications in Medicine
DR	Diabetic Retinopathy
GDC	Genomic Data Commons
HIPAA	Health Insurance Portability and Accountability Act
ICDRSS	International Clinical Diabetic Retinopathy Disease Severity Scale
ILD	Interstitial Lung Disease
ISDIS	International Society for Digital Imaging of the Skin
ISIC	International Skin Imaging Collaboration
MIMIC-CXR	MIMIC Chest X-Ray
PA	Postero-Anterior
PHI	Protected Health Information
SIIM	Society for Imaging Informatics in Medicine
TCGA	The Cancer Genome Atlas
TCGA-BRCA	The Cancer Genome Atlas Breast Invasive Carcinoma
TCIA	The Cancer Imaging Archive

Chapter 4

AI	Artificial Intelligence
Adam	Adaptive Moment Estimation
AP	Anterior-Posterior
ATT	Attention Network
AUROC	Area Under the Receiver Operating Characteristic Curve
CAM	Channel Attention Module
CBAM	Convolutional Block Attention Module
CBIR	Content-based Image Retrieval
CLAM	Clustering-constrained Attention Multiple Instance Learning Approach
CLAM-SB	Single-Branch CLAM
CLAM-MB	Multi-Branch CLAM
CNN	Convolutional Neural Network
COVID-19	Coronavirus Disease 2019
CT	Computed Tomography
DCG	Discounted Cumulative Gain
DeepLIFT	Deep Learning Important Features

DeiT	Data-efficient image Transformer
DL	Deep Learning
EMT	Epithelial-Mesenchymal Transition
FCN	Fully-Convolutional Network
GAN	Generative Adversarial Network
GSEA	Gene Set Enrichment Analysis
HIPT	Hierarchical Image Pyramid Transformer
IG-CNN	Interpretability-guided Network
LRP	Layer-wise Relevance Propagation
LSTM	Long Short-Term Memory Network
MIL	Multiple Instance Learning
MLDAM	Multi-Level Dual-Attention Mechanism
MRI	Magnetic Resonance Imaging
nDCG	Normalised Discounted Cumulative Gain
PA	Posterior-Anterior
PACS	Picture Archiving and Communication System
PAM	Position Attention Module
PET	Positron Emission Tomography
PLIP	Pathology Language-Image Pre-training
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SE	Squeeze-and-Excitation
SSIM	Structural Similarity Index
ssGSEA	single-sample GSEA
TME	Tumour Microenvironment
TransMIL	The Transformer based MIL
WSI	Whole Slide Image

Chapter 5

ADCG	Adjusted Discounted Cumulative Gain
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUROC	Area Under the Receiver Operating Characteristic Curve
CAcc	Contrastive Accuracy
CBIR	Content-based Image Retrieval
CFIR	Collaborative Filtering Image Retrieval
CNN	Convolutional Neural Network
CONCH	Contrastive learning from Captions for Histopathology
CT	Computed Tomography
DBN	Deep Belief Network
DBM	Deep Boltzmann Machine
DL	Deep Learning
DT	Decision Tree
EEG	Electroencephalography
fMRI	functional Magnetic Resonance Imaging

GNB	Gaussian Naive Bayes
GPSA	Gated Positional Self-Attention
LR	Logistic Regression
MCAT	Multi-modal Co-Attention Transformer
MIL	Multiple Instance Learning
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
PET	Positron Emission Tomography
RBM	Restricted Boltzmann Machine
SPECT	Single-Photon Emission Computed Tomography
SVM	Support Vector Machine
tanh	Hyperbolic Tangent
TME	Tumour Microenvironment
VT	Visual Transformer
ViT	Vision Transformer
WED	Weighted Euclidean Distance
WSI	Whole Slide Image

Chapter 6

AI	Artificial Intelligence
ADCG	Adjusted Discounted Cumulative Gain
BIN	Born Identity Network
CACC	Contrastive Accuracy
CLEAR	Counterfactual Local Explanations via Regression
CI-Net	Clinical-Inspired Network
CoCoX	Conceptual and Counterfactual Explanations
CX-ToM	Counterfactual Explanations with the Theory-of-Mind
C3LT	Cycle-Consistent Counterfactuals by Latent Transformations
DA-DGCEX	Distribution Aware Deep Guided Counterfactual Explanations
DCFG	Directional Counterfactual Generation
DeDUCE	Deep Deterministic Uncertainty-based Counterfactual Explanations
Deformable ProtoPNet	Deformable Prototypical Part Network
DGCEX	Deep Guided Counterfactual Explanations
DiME	Diffusion Models for Counterfactual Explanations
DISC	Deep Inversion for Synthesising Counterfactuals
DiscGrad	Discriminative Gradients
DiVE	Diverse Valuable Explanations
ECINN	Efficient Counterfactuals from Invertible Neural Networks
EU-GDPR	European Union's General Data Protection Regulation
FRACE	Fast Real-time Counterfactual Explanation
GANs	Generative Adversarial Networks
HPnet	Hierarchical Prototype Network
iFaster-RCNN	Interpretable-Faster-RCNN
Latent-CF	Counterfactual Explanations from a Latent Representation

LEAR	Learn-Explain-Reinforce
PIECE	Plausible Exceptionality-based Contrastive Explanations
ProtoPFormer	Prototypical Part Transformer
ProtoPNet	Prototypical Part Network
ROI	Region of Interest
SA	Score Aggregation
SCOUT	Self-aware Discriminant Counterfactual Explanations
SDEC	Search for Evidence Counterfactual
SDFA	Shallow-Deep Feature Alignment
SLIM	Supersparse Linear Integer Model
STEEEX	Steering Counterfactual Explanations with Semantics
SVM	Support Vector Machine
VQA	Visual Question Answering
xAI	Explainable Artificial Intelligence

Chapter 7

AI	Artificial Intelligence
CNN	Convolutional Neural Network
Deformable ProtoPNet	Deformable Prototypical Part Network
MIL	Multiple Instance Learning
ProtoPNet	Prototypical Part Network
ViT	Vision Transformer
xAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Context and Motivation

The advent of deep learning (i.e., access to more data and computational resources) allowed artificial intelligence (AI) algorithms to achieve remarkable human-like performances in several applications, such as healthcare [9]. However, despite their human-challenging predictive performance in several tasks, these models still suffer from *black-box* behaviour and often fail in seemingly simple situations [10]. These idiosyncrasies may jeopardise the acceptance of tools based on these models, especially in fields that rely on high-stakes decisions (e.g., finance, law, medicine, and politics), where it is of utmost importance that the reasoning process is transparent for both the decision-makers and the ones influenced by such decisions [11]. This context motivated the research on explainable artificial intelligence (xAI), aiming to unveil the inner functioning of machine learning algorithms [12].

Generally, research on xAI can be seen as a three-stage workflow (i.e., pre-model, in-model and post-model) [13]. Pre-model methods focus on understanding the data distribution before building the model through exploratory data analysis [14, 15, 16, 17]. In-model methods seek to integrate interpretability inside the model, either by relying on models based on rules [18, 19], based on cases [20, 21, 22], through the use of regularisation techniques (e.g., sparsity, monotonicity) during training [23, 24], by guiding the neural network into learning relevant concepts [25, 26], or by seeking to integrate causal knowledge into the network [27, 28]. Finally, post-model methods are related to a posterior analysis of the model predictions, either producing saliency maps through gradient information [29, 30, 7], deconvolution [31, 32], optimisation [33], decomposition [34, 35], or through a connection with high-level semantic concepts [36, 37, 25]. As most of the initial works in xAI tried to explain models using a *post-hoc* approach (i.e., explaining a model after training) [38, 39], the community realised that despite the practicality of having model-independent methods to generate explanations, we must take into account that other challenges arise: how to explain such explanations [40] or how to ensure that these explanation-generation methods are robust against adversarial attacks [41]. Therefore, in high-stake decision fields such as healthcare,

it may be preferable to focus on *intrinsically interpretable* models (i.e., models that integrate interpretable rules by design) [42].

Both clinical and technical communities acknowledge that the field of interpretable machine learning for healthcare still has several open challenges that require our attention. This thesis project was a fair opportunity to tackle some of these open challenges and to work towards novel contributions to interpretable algorithms for healthcare. Although we believe that post-hoc methods might serve to check models' behaviour and to validate if the predictions are coherent [43] and robust, we also devoted some efforts to understanding the quality of these xAI models.

1.2 United Nations Sustainable Development Goals

The United Nations (UN) Sustainable Development Goals (SDGs) (see Figure 1.1) work as a global partnership between countries (developed and developing) to end poverty and other deprivations, with strategies that improve health and education, reduce inequality, and promote economic growth while tackling climate change and working to preserve our oceans and forests [44]. They are the core of the 2030 Agenda for Sustainable Development [45], adopted by all UN Member States in 2015. It provides a shared blueprint for peace and prosperity for people and the planet, now and into the future. We consider this thesis to be related to the following SDGs:

- Goal #3 “Ensure healthy lives and promote well-being for all at all ages”: this research work aims to contribute to the development of novel AI algorithms for the medical community that can lead to better and transparent medical decisions while keeping high performance;
- Goal #9 “Build resilient infrastructure, promote inclusive and sustainable industrialisation and foster innovation”: research on AI algorithms can contribute towards making processes more efficient and help stakeholders of public or private sectors to see innovation as a top priority;
- Goal #10 “Reduce inequality within and among countries”: developing transparent and interpretable medical AI algorithms can help us identify and fix structural or cultural biases related to systematic inequalities.

1.3 Objectives and Document Structure

This thesis aims to study and develop interpretable computer vision and machine learning algorithms for medical decision support systems. The attentive reader will notice that we do not specify a single medical application. However, this thesis relies on the aesthetic assessment of breast cancer surgery outcomes and computational pathology use cases to illustrate some of the proposed methodologies in the applied contributions (for the fundamental contributions, the range of use cases is diverse). In this thesis, we approach the field of interpretable machine learning from different dimensions and



Figure 1.1: The United Nations Sustainable Development Goals.

try to identify new applications or details of improvement. Moreover, the experimental Chapters of this document often look at the same (or similar) tasks from different algorithmic perspectives and discuss the benefits and limitations of each. Although each Chapter is self-contained (given the disparate lines of research discussed in this thesis), the reader will notice that there is a fluid storyline, giving them the necessary intuition to relate each Chapter to the next (or the previous). Thus, the following constitute the main scientific and technical objectives of the project:

- Assessing the quality of current techniques that aim to promote interpretability in the medical context;
- Investigating novel algorithms and applications of inherently interpretable artificial intelligence for medical applications;
- Understanding the role of diverse data modalities (i.e., multi-modal medical data) and different data fusion approaches in the predictive power of algorithms.

Including this Introduction, this thesis contains seven Chapters:

- Chapter 2, which contextualises the topics described in this thesis and contains theoretical background on the fundamentals of deep learning and explainable artificial intelligence;
- Chapter 3, which describes the datasets used in the experimental part of the works reported in this thesis;
- Chapter 4, which focuses the discussion on the study of deep learning algorithms based on attention mechanisms and their application to medical data, including classification, retrieval and multiple instance learning (digital pathology);

- Chapter 5, which explores the potential of multi-modal machine learning in medical data, especially in medical image retrieval, and on multiple instance learning (digital pathology);
- Chapter 6, which dives into the realm of inherently interpretable machine learning and presents disparate applications of this field in medical data, including works related to different deep learning architectures (i.e., B-cos networks), medical image retrieval and prototypical learning;
- Chapter 7, which provides the concluding remarks of this thesis, reflects on the current challenges of this field and suggests future work directions.

1.4 Contributions

This doctoral thesis starts with the hypothesis that xAI approaches are essential (or even mandatory) to the medical domain. Following the literature and all the claims that the trade-off between interpretability and predictive performance was not relevant (i.e., not significant), we hypothesised that, for the medical domain (independently of the application), these algorithms would lead us to the right path, without harming (significantly) performance or optimisation. To test this hypothesis, we tackled the field of xAI (for the medical domain) from three points of view:

1. **Learning What is Relevant:** we made fundamental and applied contributions to the topic of attention mechanisms for medical image classification, medical image retrieval and computational pathology (see Chapter 4). We assessed the potential benefits of using attention mechanisms (for medical image classification), found that attention-driven models can learn embedding spaces that contain more meaningful information about the pathology (for medical image retrieval), and confirmed their quality as feature extraction and score aggregation operators (for computational pathology).
2. **Learning from Disparate Sources of Data:** we made applied contributions to the topic of multi-modal machine learning for medical image retrieval and computational pathology (see Chapter 5). We proposed a multi-modal framework for medical image retrieval, extended this study with attention mechanisms (establishing the connection to the previous topic), and increased the work on computational pathology studying the impact of multi-modal approaches in algorithms' predictive performance.
3. **Learning in a Human Understandable Way:** we made fundamental and applied contributions to the topic of inherently interpretable machine learning for medical image classification, medical image retrieval and counterfactual learning with prototypes (see Chapter 6). We studied the suitability of B-cos networks for medical applications (classification and retrieval) and dissected the paradigm of prototypical learning and its potential use in counterfactual learning and retrieval.

1.5 Publications

During his doctoral studies, the author collaborated in research in computer vision and interpretable machine learning applied to healthcare and biometrics. In this Section, we list all the publications that resulted (directly and indirectly) from the doctoral research described in this thesis.

The core work of this thesis, described in Chapters 4 – 6, resulted in five articles in international journals and conference proceedings. The publications are enumerated below, clustered by type and in reverse chronological order:

5. Mohammad Hossein Zolfagharnasab and Nuno Freitas and Tiago Goncalves and Eduard Bonci and Carlos Mavioso and Maria J. Cardoso and Helder P. Oliveira and Jaime S. Cardoso, “Predicting Aesthetic Outcomes in Breast Cancer Surgery: A Multimodal Retrieval Approach,” in *Proceedings of the Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care (in MICCAI 2024)*, 2024 [46]
4. I. Rio-Torto, T. Gonçalves, J. S. Cardoso, and L. F. Teixeira, “On the Suitability of B-cos Networks for the Medical Domain,” in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2024 [47]
3. T. Gonçalves, D. Pulido-Arias, J. Willett, K. V. Hoebel, M. Cleveland, S. R. Ahmed, E. Gerstner, J. Kalpathy-Cramer, J. S. Cardoso, C. P. Bridge, and A. E. Kim, “Deep Learning-based Prediction of Breast Cancer Tumor and Immune Phenotypes from Histopathology,” 2024 [48]
2. 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*, vol. 12, no. 1, p. 20732, 2022 [49]
1. T. Gonçalves, I. Rio-Torto, L. F. Teixeira, and J. S. Cardoso, “A Survey on Attention Mechanisms for Medical Applications: Are We Moving Toward Better Algorithms?,” *IEEE Access*, vol. 10, pp. 98909–98935, 2022 [50]

At a national level, this work resulted in ten short papers presented at the Portuguese Conference on Pattern Recognition (RECPAD) from 2021 through 2024 (four editions). The publications are enumerated below, clustered by type and in reverse chronological order:

10. T. Gonçalves, L. Capozzi, A. Rebelo, and J. S. Cardoso, “Deep Image Segmentation based on Mutual Information: A Study,” in *30th Portuguese Conference in Pattern Recognition (RECPAD)*, (Covilhã, Portugal), 2024 [51]
9. L. Capozzi, T. Gonçalves, J. S. Cardoso, and A. Rebelo, “Towards Fingerprint Presentation Attack Generation using Generative Adversarial Networks,” in *30th Portuguese Conference in Pattern Recognition (RECPAD)*, (Covilhã, Portugal), 2024 [52]

8. M. H. Zolfagharnasab, N. Freitas, T. Gonçalves, E. Bonci, C. Mavioso, M. J. Cardoso, H. P. Oliveira, and J. S. Cardoso, “Multimodal Deep Learning for Predicting Aesthetic Outcomes in Breast Cancer Surgery,” in *30th Portuguese Conference in Pattern Recognition (RECPAD)*, (Covilhã, Portugal), 2024 [53]
7. T. Gonçalves, L. Capozzi, A. Rebelo, and J. S. Cardoso, “Optimisation of Deep Neural Networks using a Genetic Algorithm: A Comparative Study,” in *29th Portuguese Conference in Pattern Recognition (RECPAD)*, (Coimbra, Portugal), 2023 [54]
6. L. Capozzi, T. Gonçalves, J. S. Cardoso, and A. Rebelo, “Analysis of Neurons’ Information in Deep Spiking Neural Networks using Information Theory,” in *29th Portuguese Conference in Pattern Recognition (RECPAD)*, (Coimbra, Portugal), 2023 [55]
5. I. Rio-Torto, C. Patrício, H. Montenegro, T. Gonçalves, and J. S. Cardoso, “Detecting Concepts and Generating Captions from Medical Images,” in *29th Portuguese Conference in Pattern Recognition (RECPAD)*, (Coimbra, Portugal), 2023 [56]
4. H. Montenegro, P. Neto, C. Patrício, I. Rio-Torto, T. Gonçalves, and L. F. Teixeira, “Evaluating Privacy on Synthetic Images Obtained using Deep Generative Models,” in *29th Portuguese Conference in Pattern Recognition (RECPAD)*, (Coimbra, Portugal), 2023 [57]
3. E. Caldeira, E. M. Castro, and T. Gonçalves, “From Easy to Hard: A Curriculum Learning Approach for Breast Lesion Classification,” in *28th Portuguese Conference in Pattern Recognition (RECPAD)*, (Leiria, Portugal), 2022 [58]
2. R. C. Diana Teixeira e Silva and T. Gonçalves, “Preliminary Study on Deep Iterative Semantic Segmentation,” in *28th Portuguese Conference in Pattern Recognition (RECPAD)*, (Leiria, Portugal), 2022 [59]
1. T. Gonçalves and J. S. Cardoso, “Preliminary Study on the Impact of Attention Mechanisms for Medical Image Classification,” in *27th Portuguese Conference in Pattern Recognition (RECPAD)*, (Évora, Portugal), 2021 [60]

The work related to collaborations or other research activities (e.g., challenges, competitions, hackathons, and student supervision) also played an essential role in developing the main storyline of this thesis, resulting in nineteen articles and nine abstracts in international journals and conference proceedings. The publications are enumerated below, clustered by type and in reverse chronological order:

- Articles in International Journals and Conference Proceedings:

19. T. Gonçalves, A. Hedström, A. Pahud de Mortanges, X. Li, H. Müller, J. S. Cardoso, and M. Reyes, “Chapter 15 - Interpretable AI for medical image analysis: methods, evaluation, and clinical considerations,” in *Trustworthy AI in Medical Imaging* (M. Lorenzi

- and M. A. Zuluaga, eds.), The MICCAI Society book Series, pp. 315–346, Academic Press, 2025 [61]
18. M. Vieira, T. Gonçalves, W. Silva, and A. F. Sequeira, “An End-to-End Framework to Classify and Generate Privacy-Preserving Explanations in Pornography Detection,” in *2024 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–6, 2024 [62]
 17. M. M. Beirão, J. Matos, T. Gonçalves, C. Kase, L. F. Nakayama, D. de Freitas, and J. S. Cardoso, “Classification of Keratitis from Eye Corneal Photographs using Deep Learning,” in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3060–3065, 2024 [63]
 16. P. C. Neto, R. M. Mamede, C. Albuquerque, T. Gonçalves, and A. F. Sequeira, “Massively Annotated Datasets for Assessment of Synthetic and Real Data in Face Recognition,” 2024 [64]
 15. Inês Martins and João Matos and Tiago Goncalves and Leo A. Celi and An-Kwok Ian Wong and Jaime S. Cardoso, “Evaluating the Impact of Pulse Oximetry Bias in Machine Learning under Counterfactual Thinking,” in *Proceedings of the Third Workshop on Applications of Medical Artificial Intelligence (AMAI) (in MICCAI 2024)*, 2024 [65]
 14. Eduarda Caldeira and Pedro C. Neto and Tiago Gonçalves and Naser Damer and Ana F. Sequeira and Jaime S. Cardoso, “Disentangling Morphed Identities for Face Morphing Detection,” *Science Talks*, vol. 10, p. 100331, 2024 [66]
 13. R. Cruz, D. T. e. Silva, T. Gonçalves, D. Carneiro, and J. S. Cardoso, “Two-Stage Framework for Faster Semantic Segmentation,” *Sensors*, vol. 23, no. 6, 2023 [67]
 12. E. Castro, P. M. Ferreira, A. Rebelo, I. Rio-Torto, L. Capozzi, M. F. Ferreira, T. Gonçalves, T. Albuquerque, W. Silva, C. Afonso, *et al.*, “Fill in the blank for fashion complementary outfit product Retrieval: VISUM summer school competition,” *Machine Vision and Applications*, vol. 34, no. 1, p. 16, 2023 [68]
 11. O. Kaidar-Person, M. Antunes, J. S. Cardoso, O. Ciani, H. Cruz, R. Di Micco, O. D. Gentilini, T. Gonçalves, P. Gouveia, J. Heil, P. Kabata, D. Lopes, M. Martinho, H. Martins, C. Mavioso, M. Mika, H. Montenegro, H. P. Oliveira, A. Pfob, N. Rotmensz, T. Schinköthe, G. Silva, R. Tarricone, M.-J. Cardoso, and on behalf of the CINDERELLA Consortium, “Evaluating the ability of an artificial-intelligence cloud-based platform designed to provide information prior to locoregional therapy for breast cancer in improving patient’s satisfaction with therapy: The CINDERELLA trial,” *PLOS ONE*, vol. 18, pp. 1–12, 08 2023 [69]
 10. P. Serrano e Silva, R. Cruz, A. S. M. Shihavuddin, and T. Gonçalves, “Interpretability-Guided Human Feedback During Neural Network Training,” in *Pattern Recognition and Image Analysis* (A. Pertusa, A. J. Gallego, J. A. Sánchez, and I. Domingues, eds.), (Cham), pp. 276–287, Springer Nature Switzerland, 2023 [70]

9. E. Caldeira, P. C. Neto, T. Gonçalves, N. Damer, A. F. Sequeira, and J. S. Cardoso, “Unveiling the Two-Faced Truth: Disentangling Morphed Identities for Face Morphing Detection,” in *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 955–959, 2023 [71]
8. H. Montenegro, P. Neto, C. Patrício, I. Rio-Torto, T. Gonçalves, and L. F. Teixeira, “Evaluating Privacy on Synthetic Images Generated using GANs: Contributions of the VCMi Team to ImageCLEFmedical GANs 2023,” *Challenge*, p. 8, 2023 [72]
7. I. Rio-Torto, C. Patrício, H. Montenegro, T. Gonçalves, and J. S. Cardoso, “Detecting concepts and generating captions from medical images: Contributions of the VCMi team to Image-CLEFmedical caption 2023,” in *CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Thessaloniki, Greece, 2023* [73]
6. D. T. e Silva, R. Cruz, T. Gonçalves, and D. Carneiro, “Two-stage semantic segmentation in neural networks,” in *Fifteenth International Conference on Machine Vision (ICMV 2022)* (W. Osten, D. P. Nikolaev, and J. J. Zhou, eds.), vol. 12701, p. 127010G, International Society for Optics and Photonics, SPIE, 2023 [74]
5. M. Huber, F. Boutros, A. T. Luu, K. Raja, R. Ramachandra, N. Damer, P. C. Neto, T. Gonçalves, A. F. Sequeira, J. S. Cardoso, J. Tremço, M. Lourenço, S. Serra, E. Cermño, M. Ivanovska, B. Batagelj, A. Kronovšek, P. Peer, and V. Štruc, “SYN-MAD 2022: Competition on Face Morphing Attack Detection Based on Privacy-aware Synthetic Training Data,” in *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10, 2022 [75]
4. P. C. Neto, T. Gonçalves, M. Huber, N. Damer, A. F. Sequeira, and J. S. Cardoso, “OrthoMAD: Morphing Attack Detection Through Orthogonal Identity Disentanglement,” in *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–5, 2022 [76]
3. I. Rio-Torto, C. Patrício, H. Montenegro, and T. Gonçalves, “Detecting Concepts and Generating Captions from Medical Images: Contributions of the VCMi Team to ImageCLEFmedical 2022 Caption,” 2022 [77]
2. P. Maia, J. Morgado, T. Gonçalves, and T. Albuquerque, “Applying Machine Learning for Traffic Forecasting in Porto, Portugal,” in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (M. Kamp, I. Koprinska, A. Bibal, T. Bouadi, B. Frénay, L. Galárraga, J. Oramas, L. Adilova, Y. Krishnamurthy, B. Kang, C. Largeron, J. Lijffijt, T. Viard, P. Welke, M. Ruocco, E. Aune, C. Gallicchio, G. Schiele, F. Pernkopf, M. Blott, H. Fröning, G. Schindler, R. Guidotti, A. Monreale, S. Rinzivillo, P. Biecek, E. Ntoutsis, M. Pechenizkiy, B. Rosenhahn, C. Buckley, D. Cialfi, P. Lanillos, M. Ramstead, T. Verbelen, P. M. Ferreira, G. Andresini, D. Malerba, I. Medeiros, P. Fournier-Viger, M. S. Nawaz, S. Ventura, M. Sun, M. Zhou, V. Bitetta, I. Bordino, A. Ferretti, F. Gullo, G. Ponti, L. Severini, R. Ribeiro,

J. Gama, R. Gavalda, L. Cooper, N. Ghazaleh, J. Richiardi, D. Roqueiro, D. Saldana Miranda, K. Sechidis, and G. Graça, eds.), (Cham), pp. 299–308, Springer International Publishing, 2021 [78]

1. 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*, vol. 10, no. 4, pp. 441–455, 2021 [43]

- Abstracts in International Journals and Conference Proceedings:

9. T. Schinköthe, E. Bonci, K. Orit, H. Cruz, R. Di Micco, O. Gentilini, J. Heil, P. Kabata, M. Romariz, T. Gonçalves, *et al.*, “The CINDERELLA APPROach: Future Concepts for Patient Empowerment in Breast Cancer Treatment with Artificial Intelligence-Driven Healthcare Platform,” *European Journal of Cancer*, vol. 200, 2024 [79]
8. E.-A. Bonci, O. Kaidar-Person, M. Antunes, O. Ciani, H. Cruz, R. Di Micco, O. Gentilini, P. Gouveia, J. Heil, P. Kabata, *et al.*, “CINDERELLA Trial: validation of an artificial-intelligence cloud-based platform to improve the shared decision-making process and outcomes in breast cancer patients proposed for locoregional treatment,” *European Journal of Surgical Oncology*, vol. 50, no. 2, 2024 [80]
7. A. Pfob, E. Bonci, O. Kaidar-Person, M. Antunes, O. Ciani, H. Cruz, R. Di Micco, O. Gentilini, J. Heil, P. Kabata, *et al.*, “108TiP CINDERELLA clinical trial: Using artificial intelligence-driven healthcare to enhance breast cancer locoregional treatment decisions,” *ESMO Open*, vol. 9, 2024 [81]
6. E.-A. Bonci, O. Kaidar-Person, M. Antunes, O. Ciani, H. Cruz, R. D. Micco, O. D. Gentilini, N. Rotmensz, P. Gouveia, J. Heil, P. Kabata, N. Freitas, T. Gonçalves, M. Romariz, H. Montenegro, H. P. Oliveira, J. S. Cardoso, H. Martins, D. Lopes, M. Martinho, L. Borsoi, E. Listorti, C. Mavioso, M. Mika, A. Pfob, T. Schinköthe, G. Silva, and M.-J. Cardoso, “Abstract PO3-19-11: CINDERELLA Clinical Trial (NCT05196269): using artificial intelligence-driven healthcare to enhance breast cancer locoregional treatment decisions,” *Cancer Research*, vol. 84, pp. PO3-19-11–PO3-19-11, 05 2024 [82]
5. E.-A. Bonci, O. Kaidar-Person, M. Antunes, O. Ciani, H. Cruz, R. Di Micco, O. D. Gentilini, J. Heil, P. Kabata, M. Romariz, T. Gonçalves, H. Martins, L. Borsoi, M. Mika, A. Pfob, N. Romem, T. Schinkoethe, G. Silva, M. Bobowicz, M.-J. Cardoso, and n. null, “CINDERELLA clinical trial: Using artificial intelligence-driven healthcare to enhance breast cancer locoregional treatment decisions,” *Journal of Clinical Oncology*, vol. 42, no. 16_suppl, pp. TPS621–TPS621, 2024. PMID: [83]
4. Shreyas Bhat Brahmavar and Tiago Goncalves and Tobias R. Bodenmann and Syed Rakin Ahmed and Jay B. Patel and Praveer Singh and Katharina V. Hoebel and Mason C. Cleveland and Felix Dorfner and Dagoberto Pulido-Arias and Philipp Hahnel and Jaime S. Cardoso and Jayashree Kalpathy-Cramer and Priscilla Brastianos and Elizabeth

- Gerstner and Albert E. Kim and Christopher P. Bridge, “Deep Learning-based Non-Invasive Molecular Profiling of Brain Metastases from MR Imaging,” 2024 [84]
3. I. Martins, J. Matos, T. Goncalves, L. A. Celi, A. I. Wong, and J. S. Cardoso, “Thermometry Bias in Critical Care and Impact in Downstream Machine Learning Tasks,” 2024 [85]
 2. M. Beirao, J. Matos, T. Goncalves, C. Kase, L. F. Nakayama, and J. S. Cardoso, “A machine learning system for keratitis diagnosis from cornea images,” 2024 [86]
 1. D. P. Arias, T. Goncalves, J. Kalpathy-Cramer, E. Gerstner, J. S. Cardoso, A. Kim, and C. P. Bridge, “Dissecting the Impact of Data Augmentation on Whole Slide Image Classification,” 2024 [87]

1.6 Involvement in the Scientific Community

This Section summarises the participation of the author in the Scientific Community.

1.6.1 Collaborations

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

- Visual Computing and Machine Intelligence (VCMI) Research Group of the Centre of Telecommunications and Multimedia at INESC TEC: Ana Filipa Sequeira, Ana Rebelo, Hélder P. Oliveira, Luís F. Teixeira, Ricardo Cruz, Tânia Pereira, Wilson Silva, Cristiano Patrício, Helena Montenegro, Isabel Rio-Torto, Leonardo Capozzi, Miguel Romariz, Mohammad Hossein Zolfagharnasab, Pedro C. Neto, Pedro Fernandes Sousa, Rafael Mamede, Sara Pires Oliveira, Vitória Cruz, Leonardo Machado Ferreira, Eduardo Meca Castro and João Ribeiro Pinto – the author collaborated with several researchers of the VCMI group in topics related his thesis (e.g., computer vision, xAI for healthcare) and in other research lines or projects of interest to the group (e.g., biometrics, challenges, hackathons, software development, synthetic data);
- Champalimaud Foundation: Maria João Cardoso, Carlos Mavioso, Eduard-Alexandru Bonci and other members of the CINDERELLA project’s consortium – the author collaborated with researchers and clinicians of the Champalimaud Foundation in projects related to breast cancer;
- Quantitative Translational Imaging in Medicine (QTIM) Lab of the Athinoula A. Martinos Center for Biomedical Imaging (Martinos) at the Massachusetts General Hospital (MGH): Christopher Bridge, Albert Kim, Dagoberto Pulido-Arias and other members of the QTIM Lab – supervised by Professors Christopher Bridge and Albert Kim, the author collaborated

in projects related to medical imaging and interpretable digital pathology (topic directly related to this thesis);

- Radiology AI Lab at the Netherlands Cancer Institute (NKI) and AI Technology for Life Lab at the University of Utrecht: Wilson Silva and other members of the labs – supervised by Wilson Silva, the author collaborated in projects related to medical imaging and xAI, devoting some efforts to explainable multi-label and multi-modal architectures (not published nor included in this thesis);
- Massachusetts Institute of Technology (MIT) and Duke University: Leo Anthony Celi, An-Kwok Ian Wong, João Carlos Gonçalves Matos – thanks to a long-term friendship with João Carlos Gonçalves Matos, the author worked with these researchers on topics related to health equity and bias (not included in this thesis);
- Federal University of São Paulo (UNIFESP): Luis Filipe Nakayama – thanks to João Carlos Gonçalves Matos, who met Luis during their stays at MIT, the author worked with these researchers on topics related to ophthalmologic diseases, xAI and bias (not included in this thesis);
- Mauricio Reyes (University of Bern, University Hospital Bern), Henning Müller (HES-SO Valais-Wallis, University of Geneva), Xiaoxiao Li (University of British Columbia), Aurélie Pahud de Mortanges (University of Bern, University Hospital Bern) and Anna Hedström (Technical University Berlin) – with his supervisor, Jaime S. Cardoso, the author collaborated with these researchers in the book chapter “Interpretable AI for Medical Image Analysis: Methods, Evaluation and Clinical Considerations”, present in the book “Trustworthy AI in Medical Imaging” [61].

1.6.2 Scientific Events

During the time of his PhD, the author was involved in the organisation of the following scientific events:

- INVICTA school of Vision, Computational intelligence, and patTern Analysis (INVICTA) Spring School – the author was a Member of the Steering Committee (2024), and a Member of the Advisory Board (2025);
- Workshop on Explainable & Interpretable Artificial Intelligence for Biometrics (xAI4Biometrics) – the author was a Publicity and Local Organisation Chair (2021, 2023), a Sponsorship Chair (2021, 2022), a Member of the Organising Committee (2021, 2022, 2023, 2024);
- Doctoral Congress in Engineering of Faculdade de Engenharia da Universidade do Porto – the author was a Member of the Organising Committee of the Symposium on Electrical and Computer Engineering (2023, 2025);

- International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) Hackathon – the author was a Member of the Organising Committee (2022);
- VISion Understanding and Machine intelligence (VISUM) Summer School – the author was a Member of the Organisation (2020), a Member of the Project Committee (2021), and the Leader of the Project Committee (2022).

1.6.3 Scientific Projects

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

- Image.ME: Module for Image Management and Promotion of Aesthetic Assessment of Breast Surgery – the author was a member of the project;
- Meta Breast: Health from Portugal – the author was a member of the 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 – the author was a member of the project;
- XAIBIO: Explainable Artificial Intelligence for Biometrics – the author participated in the writing of the proposal and was a member of the project;
- TAMI: Transparent Artificial Medical Intelligence – the author was a member of the project;
- CLARE: Computer-Aided Cervical Cancer Screening – the author was a member of the project.

1.6.4 Supervision of MSc Dissertations

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

13. Leonor Justo dos Santos Fernandes, “Disentanglement and Assessment of Shortcuts on Ophthalmological Retinal Imaging Exams”, Master in Bioengineering, Faculdade de Engenharia da Universidade do Porto, 2025 (ongoing) – as co-supervisor, alongside Professor Jaime S. Cardoso (supervisor), Luis Filipe Nakayama (co-supervisor) and João Carlos Gonçalves Matos (co-supervisor);
12. Mafalda Luisa Sampaio Oliveira, “Bias detection and mitigation for Biometrics Explainable Artificial Intelligence methods applied to Face Recognition”, Master in Bioengineering, Faculdade de Engenharia da Universidade do Porto, 2024 – as co-supervisor, alongside Ana Filipa Sequeira (supervisor) and Pedro David Carneiro Neto (co-supervisor);

11. Carolina Mendes Catarino Cerveira Albuquerque, “An integrated and user-friendly platform for the deployment of explainable Artificial Intelligence methods applied to Face Recognition”, Master in Bioengineering, Faculdade de Engenharia da Universidade do Porto, 2024 – as co-supervisor, alongside Ana Filipa Sequeira (supervisor) and Pedro David Carneiro Neto (co-supervisor);
10. Diogo de Almeida Campas, “Deep Learning Generative Models for Breast MRI Synthesis: A 3D-GAN Comparative Study”, Master in Electrical and Computer Engineering, Faculdade de Engenharia da Universidade do Porto, 2024 – as co-supervisor, alongside Professor Hélder P. Oliveira (supervisor) and Professor Luís F. Teixeira (co-supervisor);
9. João Afonso Martins Domingues Andrade, “Enhancing Lung CT 3D Image Synthesis with Generative Adversarial Networks”, Master in Informatics and Computing Engineering, Faculdade de Engenharia da Universidade do Porto, 2024 – as co-supervisor, alongside Professor Hélder P. Oliveira (supervisor), Professor Tânia Maria Pereira Lopes (co-supervisor) and Pedro Fernandes Sousa (co-supervisor);
8. Pedro Manuel Santos Pereira, “Deep Learning for Image Super-Resolution: Enhancing Generated CT Scans”, Master in Informatics and Computing Engineering, Faculdade de Engenharia da Universidade do Porto, 2024 – as co-supervisor, alongside Professor Hélder P. Oliveira (supervisor), Professor Luís F. Teixeira (co-supervisor), Professor Tânia Maria Pereira Lopes (co-supervisor) and Pedro Fernandes Sousa (co-supervisor);
7. Carlos Miguel de Sá Veloso, “A Deep Segmentation Approach for the Retrieval of Breast Cancer Treatment Outcomes”, Master in Data Science and Engineering, Faculdade de Engenharia da Universidade do Porto, 2024 – as co-supervisor, alongside Professor Hélder P. Oliveira (supervisor), Mohammad Hossein Zolfagharnasab (co-supervisor) and Nuno Pinheiro da Silva Mota Freitas (co-supervisor);
6. Maria Miguel Santos Beirão, “Classification of Keratitis from Eye Corneal Photographs using Deep Learning: A Comparative Study on Amoeba, Bacteria, and Fungi”, Master in Bioengineering, Faculdade de Engenharia da Universidade do Porto, 2024 – as co-supervisor, alongside Professor Jaime S. Cardoso (supervisor) and João Carlos Ramos Gonçalves Matos (co-supervisor);
5. Inês Alves Martins, “A Data-Driven Framework for the Assessment and Mitigation of Clinical Measurement Disparities”, Master in Bioengineering, Faculdade de Engenharia da Universidade do Porto, 2024 – as co-supervisor, alongside Professor Jaime S. Cardoso (supervisor) and João Carlos Ramos Gonçalves Matos (co-supervisor);
4. Margarida Jorge Cerqueira Vieira, “Classifying and Explaining Pornography Detection: An End-to-End Framework”, Master in Informatics and Computing Engineering, Faculdade de

Engenharia da Universidade do Porto, 2024 – as co-supervisor, alongside Ana Filipa Sequeira (supervisor) and Wilson José dos Santos Silva (co-supervisor);

3. Pedro João Cruz Serrano e Silva, “Human Feedback during Neural Networks Training”, Master in Bioengineering, Faculdade de Engenharia da Universidade do Porto, 2022 – as co-supervisor, alongside Ricardo Cruz (supervisor);
2. José Pedro Moreira de Almeida Baptista, “Website Development for the VCMi Research Group”, Master in Informatics and Computing Engineering, Faculdade de Engenharia da Universidade do Porto, 2022 – as co-supervisor, alongside Professor Hélder P. Oliveira (supervisor) and João Ribeiro Pinto (co-supervisor);
1. João Gonçalves Rodrigues, “WebBCCT.core - A Web Application for the Assessment of Breast Cancer Conservative Treatment Cosmetic Results”, Master in Network and Informatic Systems Engineering, Faculdade de Ciências da Universidade do Porto, 2020 – as co-supervisor, alongside Professor Inês de Castro Dutra (supervisor) and Professor Jaime S. Cardoso (co-supervisor).

Chapter 2

Background

This Chapter provides a (brief) background on deep learning and explainable artificial intelligence. Comprehending the fundamental concepts related to these topics is essential to understanding the work described in the remaining Chapters. Hence, we hope that the following sections help the reader get acquainted with the relevant concepts (and definitions) used or mentioned throughout this thesis.

2.1 Deep Learning

This Section presents the context and fundamentals of deep learning, hoping to provide the reader with a brief (but informative) overview of the research on this field.

2.1.1 Origin and Fundamentals

The study of deep learning (DL) algorithms is originally related to the study of neurobiology and neuroscience. The appearance of the first artificial neural networks (ANNs) dates back to the 1950s, with the works of Hebb [88], McCulloch and Pitts [89], Farley and Clark [90], and Rosenblatt [91]. The latter describes the *Perceptron* model, the precursor of deep neural networks (DNNs), which used an iterative learning process that consisted of adjusting its parameters to minimise an error function in a supervised manner (i.e., using annotated data). Although it was a scientific breakthrough, the simplicity (reduced depth/number of layers) of this model created some learning limitations. The trick to overcome these obstacles was to include additional (hidden) layers and build a *Multilayer Perceptron*, a model that can capture more complex patterns. This model is also known as a universal approximator, as Hornik et al. [92, 93] showed that multilayer networks (if set with proper parameters) can learn any continuous function to an arbitrary small error. A widely used technique to optimise the best parameters is *backpropagation*, proposed by Rumelhart et al. [94]. This algorithm has two stages: 1) the network is initialised with random parameters for the first training iteration, and a forward pass is performed to obtain a prediction; 2) the error between the prediction and the ground truth is given by a *loss* function, the gradient of

the error with respect to the model parameters is computed and then the model's parameters are updated by gradient descent, with more or less intensity (defined by the *learning rate*).

2.1.2 Democratisation and Innovation

While the intuition and (basic) protocols to train DL methods were established in the XX century, the (fast) spread and (large) availability of these algorithms to the scientific community and the public happened at the beginning of the 3rd millennium, when the access to large datasets and computational infrastructure, including powerful graphics processing units (GPUs) [95], became democratised and accessible to (virtually) anyone. The core concept behind DL algorithms is *representation learning* [23]. With representation learning, we can start with raw (or with minimal preprocessing) inputs and let the algorithms learn the best features that maximise performance (or minimise the error) for a given task (e.g., classification). This ability allowed DNNs to permeate diverse domains and applications, such as image recognition and analysis [96], natural language understanding and processing [97], biochemistry [98] or drug discovery [99].

In the image domain (typically associated with computer vision tasks and unstructured data), the scientific breakthrough was achieved with convolutional neural networks (CNNs) [23]. Inspired by Fukushima's *Neocognitron* [100], a typical CNN architecture is composed of convolutional (to extract feature maps) and pooling (to aggregate feature maps) layers. These models operate over patches of the images, reducing the number of parameters of the DNN but allowing the development of larger (deeper) models. Despite these advantages, CNNs became popular due to their outstanding results in the ImageNet competition [101] and the availability of large image datasets (nowadays used as benchmarks for several performances). Allied with regularisation techniques (e.g., dropout [102]) and data augmentation [103], these models achieved (fairly) remarkable results in several different tasks such as classification, segmentation or reconstruction. Naturally, researchers have been devoting several efforts to improve the predictive performance of algorithms, with new additions in the design of the architectures, like the models used in the works described in this thesis, or the changing of the paradigms, such as vision transformers (ViTs) [104] or diffusion models [105].

In the natural language domain (typically associated with sequential data), the precursor of all models is the recurrent neural network (RNN) [106]). RNN-based architectures work as dynamic systems that receive an input sequence and process one element at a time, keeping in their hidden units a *state* vector that (implicitly) contains historical information about the previous elements of that sequence. Although RNNs showed impressive results in neural machine translation and neural language modelling, the community found that these models had difficulties in learning to store information for very-long term dependencies [107]. To fix this limitation, Hochreiter and Schmidhuber [106] proposed long short-term memory (LSTM) networks, a type of RNNs that uses special hidden units to remember the inputs for a long time. Like in the image domain, the research community has been devoting time towards improving the models in terms of design or learning procedures. In this specific case, we point the reader to the importance of attention mechanisms

(see Chapter 4) that gave rise to the transformer models and are the power behind large language models (not discussed in this thesis).

Apart from the advances in these distinct domains, the field of DL also saw the rise of other interesting (and powerful) techniques, evolving at an astonishing pace and being leveraged by different stakeholders to influence (positively or negatively) people's lives. The field of generative AI, for instance, started with generative adversarial networks (GANs) [108] and variational auto-encoders (VAEs) [109], has seen the use of vision language models, large language models or diffusion models to generate artificial data (e.g., tabular, image, text, multiple modalities), enhance or correct data (e.g., super-resolution, denoising), drug discovery [110], or game engines [111]. In a different setting, DNN-based algorithms are also pushing the field of reinforcement learning, where the paradigm relies on AI agents that learn by interacting with their (often simulated) environment through reward and penalisation policies. These methods have challenged human performance in discovering new algorithms [112] or solutions for games [113].

It is undeniable (and, perhaps, inevitable) that these algorithms will permeate the quotidian of global societies, where people will use (and rely on) these algorithms to assist them in several aspects of their lives. While this will prove its value for some (or often most) the cases, we should (at least) question if we are (or will be) able to audit the functioning of these algorithms or understand what or who should be held accountable in the case of an AI-supported decision ends up having (very) bad outcomes.

2.2 Explainable Artificial Intelligence

The main topic of this thesis is explainable artificial intelligence (xAI). To avoid redundancy of concepts across Chapters, we present a brief overview of xAI so that the reader gets acquainted with the main definitions of the field.

2.2.1 Context and Motivation

In the previous Section, we saw that the increase in available computational power and the democratised access to large datasets have leveraged the development of novel AI algorithms and their applications. A key aspect in this revolution was the establishment of a different paradigm (i.e., representation learning) and the seamless integration of these algorithms in the regular routines of distinct aspects of people's lives [23, 114]. However, these algorithms work as black boxes (i.e., the user can not access their internal logic) that receive data and output results without justifying their predictions in a human-understandable way [11]. In domains that depend on high-stakes decisions (e.g., healthcare, law, finance), this lack of transparency might lead to a deficit of trust, which may jeopardise the acceptance of these technologies. Moreover, there is an inherent tension (or trade-off) between machine learning performance (predictive accuracy) and explainability. Usually, the best-performing methods are the least transparent, and the ones providing a clear explanation are less accurate [115].

2.2.2 Definitions

The topic of xAI appeared intending to support the development of a more transparent AI [116, 117]. Although there is no clear distinction between explainability and interpretability [118], one may think of these as a three-stage process [119] (see Figure 2.1): *pre-model*, *in-model*, and *post-model*. These methods aim to clarify the inner logic of machine learning models, thus promoting transparency, trust and accountability and guaranteeing fairness [12].

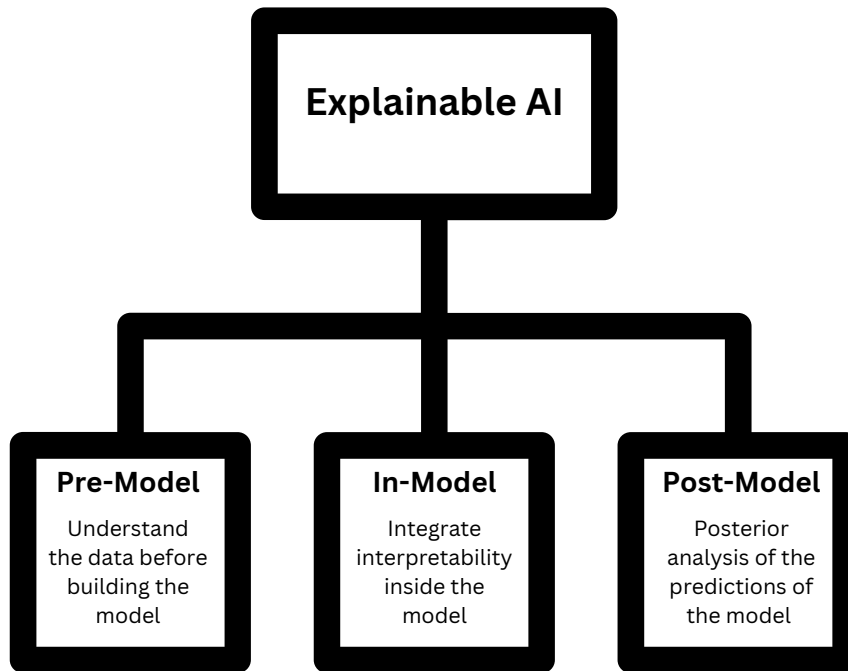


Figure 2.1: Explainable AI as a three-stage process: pre-model, in-model and post-model methods.

2.2.2.1 Pre-model Methods

Pre-model methods focus on understanding the data distribution before building the model through exploratory data analysis [14, 15]. This comprehension of the data may contribute to higher confidence in the posterior decisions that a model can provide.

2.2.2.2 In-model Methods

In-model approaches seek to integrate interpretability inside the model. To do this, one may follow different strategies: models based in rules [18], models based in cases [20], the use of regularisation techniques during training [23, 24, 120], models based in trust scores [121], self-explaining neural networks [122], models that compare distances [123] or models that generate and regularise explanations during training [124, 125].

2.2.2.3 Post-model Methods

Post-model strategies are related to a posterior analysis of the model predictions. For instance, in medical imaging, this can be done using the gradient information to identify the areas of the image that contribute more to the final decision [126, 127, 128], inserting a perturbation and observing the prediction [129], inverting the representations back to the input pixel space [130, 131] or connecting the representations to semantic concepts [132]. In this group, popular methods include Layer-wise Relevance Propagation (LRP) [34], Deep Taylor Decomposition [35], or Gradient Class Activation Mapping (Grad-CAM) [7].

2.2.3 From Post-model to Inherently Interpretable AI

The outstanding predictive performance of DL models and the generalised belief that these models seem to uncover “hidden patterns” actively contributed to the research and development of post-model methods, thus contributing to their popularity among the machine learning community. However, we agree that there are some drawbacks of exclusively investing in a post-model interpretability strategy [11]: explanations are just an approximation to what the model computes, which can be dangerous in the sense that they can be inaccurate representations of the original model in parts of the feature space [133]; explanations may not provide enough detail to understand what the black box model is doing (e.g., the majority of the post-model methods lacks repeatability and reproducibility [134]). Moreover, for applications that depend on high-stakes decisions, it is fundamental to assess the quality of these explanations [11, 135] to ensure transparency, ethics and fairness. Therefore, we should invest more efforts in pre-model methods since understanding first the data distribution that we are dealing with will contribute to an increase of confidence with the posterior decisions and explanations that a given machine learning model is capable of providing [136]; and in in-model interpretability, since inherently interpretable models provide their explanations and are faithful to what the machine learning model computes [11]. Particularly interesting to this thesis are the methods that aim to overcome the transparency and trust issues. Therefore, we tried to focus our research on intrinsically interpretable DNNs, following Rudin’s argument [11] that only inherently interpretable models might guarantee trust and transparency in high-stakes decision fields.

Chapter 3

Datasets

We used several datasets in this thesis, with some across different works. Hence, to avoid the repetition of data description in various parts of this thesis, we devote this Chapter to listing and describing all the datasets used, providing an easy reference for the reader. Table 3.1 presents a summary of all the datasets used in this thesis and might serve as a reference for the reader.

3.1 APTOS2019

The APTOS2019 dataset contains retinography images taken under various imaging conditions and from different camera models and ethnicities. This dataset was generated by Aravind Eye Hospital in India in cooperation with the Asia Pacific Tele-Ophthalmology Society (APTOS). In the version of the dataset used in the works described in this thesis, each retinal image has been graded independently by two ophthalmologists, with any disagreements adjudicated by a third ophthalmologist, and a retinopathy severity score was assigned to each image according to the NHS diabetic eye screening guidelines [137] as R0 (no diabetic retinopathy), R1 (background), R2 (pre-proliferative), R3 (proliferative), M0 (no visible maculopathy), M1 (maculopathy). The current description of the dataset states that the samples were reviewed and labelled by a group of trained doctors following the principle of the International Clinical Diabetic Retinopathy Disease Severity Scale (ICDRSS) and that samples are divided into five categories: no Diabetic Retinopathy (DR), mild DR, moderate DR, severe DR, and proliferative DR [138].

3.2 CINDERELLA Retrieval

Under the scope of the CINDERELLA project [139], the CINDERELLA Retrieval dataset [46] comprises 2,193 instances encompassing clinical attributes (i.e., height, weight, age, bra size, bra-cup size) organised as tabular data, along with their corresponding JPEG images depicting the upper torsos of patients. The images vary in size, ranging from 1869×1117 to 6000×4000 , with an average size of 4810×2903 . To create the ground-truth annotations, the clinicians were asked to rank the 10-15 most similar images to a given query image such that the sample ranked i is more

similar to the query image compared to the one appearing later in the list with rank $i + k$, where $k > 0$. For each query patient, two catalogues are annotated according to the aesthetic classification of the retrieved patients. This classification is done by clinicians according to the Harris scale [140]: one group presented patients with *Excellent/Good* aesthetic outcomes and the other with *Fair/Poor* aesthetic results.

3.3 CUB-200-2011

CUB-200-2011 [141] is a dataset used in bird species classification tasks containing 11,788 images, 200 classes and other additional annotations (i.e., part locations, binary attributes, bounding box). The dataset's images were obtained from Flickr (using its image search functionality) and then validated by multiple users. Each image contains annotations (done by multiple users) for the bounding box, part location (a total of 15 parts annotated by pixel location and visibility in each image), and attribute labels (a vocabulary of 28 attribute groupings and 312 binary attributes selected based on an online tool for bird species identification).

3.4 DTU/Herlev Pap Smear Database

The DTU/Herlev Pap Smear Database [142], built by the Herlev University Hospital (Denmark), contains images of human cells stained by the *Papanicolaou* method: a specimen of cells is smeared onto a glass slide and coloured, making it easier to examine the cells under a microscope for any abnormalities indicating a pre-cancerous stage. This dataset comprises 917 cells, allocated to 7 different classes: (1) Superficial squamous epithelial, (2) Intermediate squamous epithelial, (3) Columnar epithelial, (4) Mild squamous non-keratinising dysplasia, (5) Moderate squamous non-keratinising dysplasia, (6) Severe squamous non-keratinising dysplasia, and (7) Squamous cell carcinoma in situ intermediate. These classes are part of 2 different categories: normal (classes 1–3) and abnormal (classes 4–7).

3.5 International Skin Imaging Collaboration

The International Skin Imaging Collaboration (ISIC), sponsored by the International Society for Digital Imaging of the Skin (ISDIS), aims to improve melanoma diagnosis. The ISIC Archive contains over 13,000 dermoscopic images collected from international clinical centres and acquired using different devices within each centre. All the dermoscopic images in the ISIC Archive are screened for privacy and quality assurance and have, in most cases, associated clinical metadata vetted by melanoma experts. A subset of the images have undergone annotation and markup by recognised skin cancer experts. These markups include dermoscopic features (i.e., global and focal morphologic elements in the image known to discriminate between types of skin lesions).

Regarding the works described in this dissertation, we used two datasets of the ISIC Archive, ISIC2018 and ISIC2020, described below (see Sections 3.5.1 – 3.5.2).

3.5.1 ISIC2018

The ISIC2018 dataset [143, 144, 145], released with the ISIC 2018 Challenge [146], contains data to address three different tasks: lesion segmentation, lesion attribute detection, and disease classification.

3.5.2 ISIC2020

The ISIC2020 dataset [147] is the result of a collaboration between ISIC and the Society for Imaging Informatics in Medicine (SIIM), being released with the SIIM-ISIC Melanoma Classification challenge [148]. This dataset contains data to address melanoma diagnosis prediction.

3.6 MIMIC-CXR

The MIMIC Chest X-ray (MIMIC-CXR) dataset [149] is a large publicly available data set of chest radiographs in the Digital Imaging and Communications in Medicine (DICOM) format with free-text radiology reports. The data set contains 227835 imaging studies for 65379 patients of the Beth Israel Deaconess Medical Centre Emergency Department (Boston, MA) between 2011–2016. Each imaging study can contain one or more images, usually a frontal and a lateral view. A total of 377,110 images are available in the data set. Studies are made available with a semi-structured free-text radiology report that describes the radiological findings of the images written by a practising radiologist during routine clinical care. The data set is de-identified to satisfy the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements. Protected health information (PHI) has been removed.

3.7 PAPILA

The PAPILA dataset [150] was collected at the Department of Ophthalmology of the Hospital General Universitario Reina Sofía (Murcia, Spain) between 2018 and 2020. The patients included in this study belong to two different groups: 1) patients diagnosed with simple chronic glaucoma and 2) patients from primary care who, after an ophthalmological examination, did not show any ocular pathology that could influence the morphology of the optic nerve. This dataset contains records of 244 patients, containing 488 eye fundus images, 3 classes (i.e., healthy, suspect glaucoma, glaucoma), the segmentation masks of the cup and optic disc, and the labelling of the patients based on the evaluation of clinical data.

3.8 PH2

PH2 [151], a public dataset of dermoscopic images, acquired at the Dermatology Service of Hospital Pedro Hispano (Matosinhos, Portugal) under the same conditions. This dataset contains 200 dermoscopic images of melanocytic lesions, including 80 common nevi, 80 atypical nevi, and 40 melanomas. It also includes (image) annotations for medical segmentation of the lesion, clinical and histological diagnosis and assessment of several dermoscopic criteria (i.e., colours, pigment network, dots/globules, streaks, regression areas, blue-whitish veil).

3.9 TCGA-BRCA

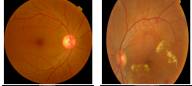
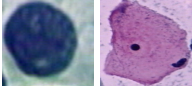
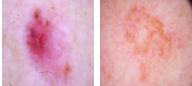
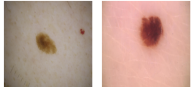
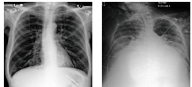
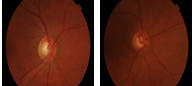
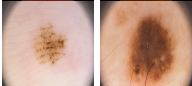
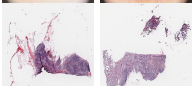
The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) data collection [152] is part of The Cancer Genome Atlas (TCGA), a project focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects, with clinical, genetic, and pathological data from the Genomic Data Commons (GDC) Data Portal [153], and radiological data from The Cancer Imaging Archive (TCIA) [154]. Using the matched identifiers, researchers can explore the TCGA or the TCIA databases to look for correlations between tissue genotype, radiological phenotype, and patient outcomes. Given that the tissues for TCGA were collected from many sites worldwide to reach their accrual targets (usually around 500 specimens per cancer type), image datasets are heterogeneous in terms of scanner modalities, manufacturers and acquisition protocols. Furthermore, in most cases, the images were acquired as part of routine care, not as part of a controlled research study or clinical trial.

3.10 VinDr-CXR

The VinDr-CXR dataset [155] consists of 18,000 postero-anterior (PA) chest X-ray radiographs (CXRs), with information for the localisation of 22 critical findings (i.e., local labels) and classification of 6 common diseases (i.e., global labels), annotated by a group of 17 radiologists with at least 8 years of experience. This dataset results from the collection of more than 100,000 images from Hospital 108 and the Hanoi Medical University Hospital in Vietnam. Regarding data labelling, this dataset contains a total of 28 findings and diagnoses in adult cases (i.e., (1) Aortic enlargement, (2) Atelectasis, (3) Cardiomegaly, (4) Calcification, (5) Clavicle fracture, (6) Consolidation, (7) Edema, (8) Emphysema, (9) Enlarged PA, (10) Interstitial lung disease (ILD), (11) Infiltration, (12) Lung cavity, (13) Lung cyst, (14) Lung opacity, (15) Mediastinal shift, (16) Nodule/Mass, (17) Pulmonary fibrosis, (18) Pneumothorax, (19) Pleural thickening, (20) Pleural effusion, (21) Rib fracture, (22) Other lesion, (23) Lung tumor, (24) Pneumonia, (25) Tuberculosis, (26) Other diseases, (27) Chronic obstructive pulmonary disease (COPD), and (28) No finding), divided into local (1–22) and global (23–28) labels. The local labels reflect the localisation of the findings, while the global labels reflect the diagnosis made by the radiologists. This list of labels was suggested by

a committee of the most experienced radiologists from the two hospitals, taking into account if they were prevalent and if they could be differentiated on CXRs.

Table 3.1: This is a summary of the datasets used in this thesis. The dataset names, references, tasks, and data examples are included in this table. Note: AD - Attribute Detection, C - Classification, R - Retrieval, S - Segmentation, MC - Multi-class, ML - Multi-label.

Dataset	Tasks	Examples
APTOS2019 [138]	C (MC)	
CINDERELLA Retrieval [46]	C, R (MC)	
CUB-200-2011 [141]	AD, C (MC, ML)	
DTU/Herlev Pap Smear Database [142]	C (MC)	
ISIC2018 [143, 144, 145]	AD, C, S (MC)	
ISIC2020 [147]	C (MC)	
MIMIC-CXR [149]	C, R (MC)	
PAPILA [150]	C, S (MC)	
PH2 [151]	C, S (MC)	
TCGA-BRCA [152]	C (MC)	
VinDr-CXR [155]	AD, C (MC, ML)	

Chapter 4

Learning What is Relevant: The Role of Attention Mechanisms in Medical Data

Foreword on Author Contributions

The results of this work have been disseminated in the form of two articles in international journals and one paper at an international conference:

1. T. Gonçalves, I. Rio-Torto, L. F. Teixeira, and J. S. Cardoso, “A Survey on Attention Mechanisms for Medical Applications: Are We Moving Toward Better Algorithms?,” *IEEE Access*, vol. 10, pp. 98909–98935, 2022 [50]
2. 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*, vol. 12, no. 1, p. 20732, 2022 [49]
3. T. Gonçalves, D. Pulido-Arias, J. Willett, K. V. Hoebe, M. Cleveland, S. R. Ahmed, E. Gerstner, J. Kalpathy-Cramer, J. S. Cardoso, C. P. Bridge, and A. E. Kim, “Deep Learning-based Prediction of Breast Cancer Tumor and Immune Phenotypes from Histopathology,” 2024 [48]

4.1 Introduction

The concept of *attention* is not new and has been part of multidisciplinary debates across psychology, physiology, and neuroscience [156]. The inspiration behind the translation of this concept into artificial intelligence (AI) algorithms comes from biological vision systems, which appear to employ a serial computational strategy when inspecting complex scenes [157]. Historically, the pursuit of an *attention mechanism* seems to be deeply related to the problem of *visual search*, where the main objective may be to find a small object in a disorderly environment (e.g., a pen on a desk). A possible solution to this problem is using high-resolution images and a wide field of view. However,

this results in very high dimensional inputs [158]. Therefore, it was essential to devise a strategy capable of selecting the most important parts of the input, deciding where to look next during the processing pipeline, and managing limited computational resources. These aspects motivated the study of *selective visual attention*, which deals with the computational complexity of computer vision tasks and often requires several basic components, such as the selection of a region of interest in the image and the selection of the dimension of features and values of interest [159]. Besides, theoretically, such mechanisms can potentially guide the research into focusing more on sparse local models [160]. Nevertheless, the scientific community proposed several research lines that certainly influenced the modern methodologies of attention-based mechanisms for deep learning. Bandera et al. [161] and Minut et al. [158] proposed *reinforcement learning* policies to regularise the training of computer vision systems in the task of target recognition, Koch and Ullman [156] and Tsotsos et al. [159] proposed a set of rules that promote visual attention via selective tuning and shifts in the processing focus [156] during the analysis of a visual scene, and Itti and Koch [157] addressed this challenge using a *saliency map* based search mechanism. Naturally, the increase of available computational power and the democratised access to *big data* allowed the resurgence of deep learning algorithms and their applications [23, 162]. With this paradigm shift, methods now focus on integrating *learnable* attention mechanisms on an end-to-end basis [163], with interesting impacts in diverse applications [164].

The first published reviews and surveys on the topic gained visibility with the increased availability of literature on attention mechanisms. We introduce the reader to some of these works to give them intuition about the work described in this Chapter and justify the research questions behind it. Borji and Itti [165] reviewed several models of visual attention based on non-deep learning methods and presented a taxonomy indicating their approaches and capabilities. Cho et al. [166] presented an early study on the use of attention mechanisms in convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with a special focus on *soft-attention* for applications such as neural machine translation, image captioning, video description generation or end-to-end neural speech recognition. Wang and Tax [167] published a survey that explored the use of attention mechanisms in RNNs for sequence-to-sequence challenges. Chaudhari et al. [1] presented a survey of attention mechanisms for natural language processing, wherein they proposed a taxonomy and reviewed several neural network architectures in which attention has been incorporated while discussing applications in which modelling attention has shown significant impact. Yang [168] performed a brief overview of attention mechanisms for computer vision and introduced a preliminary taxonomy. Lindsay [169] published a review on attention mechanisms that provides intuition about the definition of attention in the neuroscience and psychology literature and covers several use cases of attention in machine learning, indicating their biological counterparts where they exist.

In 2017, Vaswani et al. [170] proposed the *Transformer*, a deep neural network architecture composed solely of attention operations that discarded recurrence and convolution entirely. In the original paper, the authors applied this architecture to the task of neural machine translation

and explored the use of attention operations as a means of reducing computational complexity when compared to recurrent or convolution operations. The establishment of a similar architecture for computer vision occurred almost naturally when Dosovitskiy et al. [104] proposed the *Vision Transformer* architecture. Han et al. were pioneers in analysing the advantages and disadvantages of using transformer-based architectures in computer vision with a particular interest in the *self-attention* mechanism [171]. Khan et al. [172] also provided a comprehensive overview of transformer-based models from the perspective of their application and architectural design. Guo et al. [173] published a thorough and systematic review of attention mechanisms for computer vision and created a taxonomy that categorises attention mechanisms according to their data domain. Xu et al. [174] revisited the transformer architectures with particular emphasis on low-level vision and generation. Recently, Shamshad et al. [175] explored the use of transformer-based architectures for medical image analysis and published a pioneering review that provides insight into the applications of these algorithms in several dimensions of the clinical daily routine. Almost all of the previous papers focus on the current state-of-the-art, performing a comparative analysis based solely on published results, thus anticipating future and open challenges from a theoretical perspective.

In this Chapter, we present three different case studies where attention mechanisms play a role, seeking to understand their impact on the performance of the algorithms and their potential benefits to interpretability, robustness or transparency. Beyond this Introduction, this Chapter is organised as follows: Section 4.2 presents a historical overview and the fundamental concepts regarding attention mechanisms in machine learning, Section 4.3 provides a comprehensive review of the integration of attention mechanisms into algorithms for medical applications, Sections 4.4 – 4.6 present the experimental work developed to address the research questions related to this topic, and Section 4.7 concludes the Chapter, highlighting the main contributions and anticipating future work directions.

4.2 Background

This Section presents a historical overview and the fundamental concepts regarding attention mechanisms in machine learning. The results of this work have been disseminated in the form of an article in an international journal: T. Gonçalves, I. Rio-Torto, L. F. Teixeira, and J. S. Cardoso, “A Survey on Attention Mechanisms for Medical Applications: Are We Moving Toward Better Algorithms?,” *IEEE Access*, vol. 10, pp. 98909–98935, 2022 [50].

Although there is no clear nor unified concept for *attention*, we can refer to it as the ability to flexibly manage limited computational resources. This research topic spans several knowledge domains and has been studied in conjunction with many other topics in neuroscience and psychology [176], including awareness, vigilance, saliency, executive control, learning, and, more recently, AI. Therefore, we have different ways of perceiving the properties of attention mechanisms at biological, psychological or computational levels [169]. From the perspective of the domains of *arousal*, *alertness* or *vigilance*, attention could be described as a general measure of alertness or the

ability to engage with the surroundings [177]. From the perspective of *sensorial stimuli*, attention is often deployed as an alert subject to specific sensory input, allowing tight control over both the stimuli and the locus of attention (e.g., selective audio attention, visual attention) [178, 179, 180]. From the perspective of *executive control* [181], the assumption is that there are multiple simultaneous competing tasks and that a central controller is needed to decide which to engage in and when; hence, attention can be thought of as the output of executive control, since the executive control system must select the targets of attention and communicate that to the systems responsible for implementing it [169]. From the perspective of *memory*, attention is understood as the ability of the brain to select the subset of information that is well-matched to the needs of the memory system, thus being a choice regarding the deployment of limited resources [169, 182]. From the perspective of AI, the first successful implementations were accomplished with encoder-decoder structures in deep neural network architectures, where the main goal is to learn the best attention weights that connect the encoder to the decoder [166].

In the context of deep learning, RNNs were the pioneering attention-based structures. This is mainly due to their ability to learn and process sequential data (e.g., data with a temporal component), which makes RNNs a class of machines with states (i.e., their state depends on both the input to the system and the current state), such that a signal received at a given moment can alter the behaviour of the network at a much later point in time [183]. These properties allow RNNs to process variable-length structures without further modification, a property that has been extensively explored in several cognitive problems such as speech recognition [184, 185], text generation [186], handwriting generation [187] and machine translation [188]. The work described in [189] is perhaps the most classic example of the design and implementation of computational attention mechanisms with RNNs. In the original paper, the authors focused on the task of neural machine translation and proposed an RNN-based encoder-decoder architecture in which the task of the encoder is to compress the input data (in this case, an English sentence), while the task of the decoder is to receive and decode this data to achieve a meaningful translation (in this case, a French sentence). The principal motivation for introducing an attention mechanism in this learning pipeline is related to the difficulty of achieving a perfect alignment between the input and output sequences, which is an almost impossible goal in most cases. The role of the attention mechanism is to learn weights that make the decoder focus on the relevant parts of the input sequences to produce meaningful outputs. At the time of publication, this methodology improved the state-of-the-art and paved the way for new research questions and the exploration of novel attention-based frameworks.

The study of attention mechanisms comprises two main research lines: *language, text, and speech*; and *computer vision*. Since the clinical context (the subject of this thesis) involves multiple modalities of data, we examine these two main dimensions to help the reader become familiar with the nomenclature, taxonomies, and practices. This background provides the fundamental concepts of attention mechanisms in machine learning, which will aid the reader in understanding the experimental work described in this chapter.

4.2.1 Attention mechanisms for language, text, and speech

An insightful and comprehensive taxonomy for categorising attention mechanisms for language, text, and speech has been proposed by Chaudhari et al. [1]. The authors relate most of the concepts to the encoder-decoder model. Consequently, for better understanding, we need to define the following: *input sequences*, the input vectors of the model (see Figure 4.1’s “Input Sequence”); *output sequences*, the output vectors of the model (see Figure 4.1’s “Output Sequence”); *candidate states*, the hidden states of the encoder (see Figure 4.1’s “Encoder Vector”); *query states*, the hidden states of the decoder (see Figure 4.1’s “Decoder Vector”).

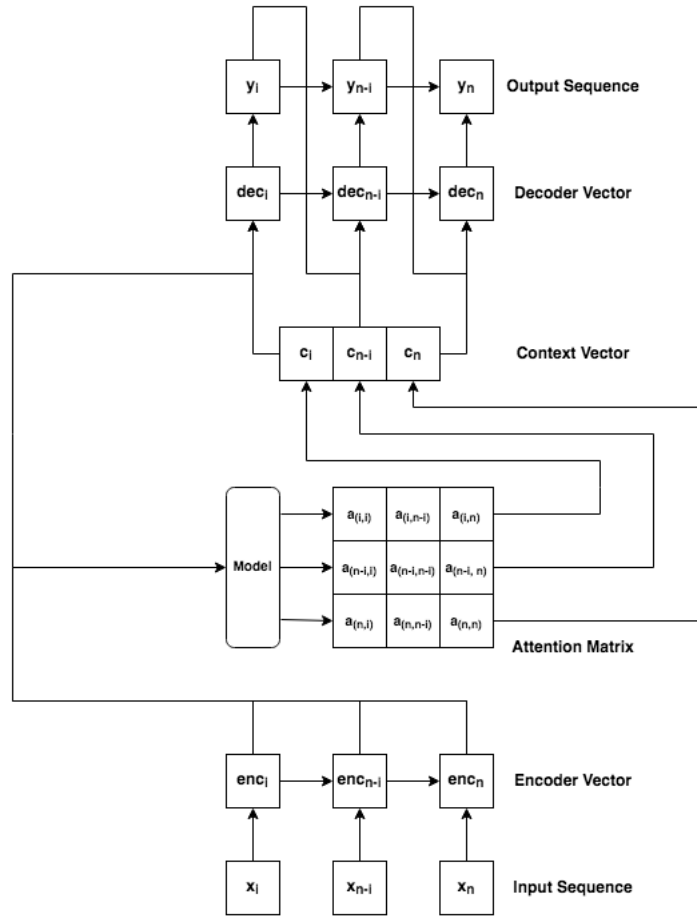


Figure 4.1: Block diagram of the general attention mechanism for language, text and speech, according to [1].

Following this rationale, Chaudhari et al. [1] organise the different types of attention mechanisms into different (non-mutually exclusive) categories:

- **Number of abstraction levels:** This category takes into account the number of representation/feature levels where the model will learn the attention weights. At this level, we may consider the following types of attention: *single-level attention*, when the attention weights are computed only for the original input sequence; and *multi-level attention*, when we apply

the attention mechanism on multiple levels of abstraction of the input sequence, usually in a sequential manner.

- **Number of positions:** This category takes into account the number of positions of the input sequence where the attention weights are learned. At this level, we may consider the following types of attention: *soft attention*, which consists of a weighted average of all the hidden states of the input sequence (i.e., the same as candidate states) to build the context vector (see Figure 4.1’s “Context Vector”); *global attention*, which is the term used in the machine translation field to describe soft attention; *hard attention*, which consists of a stochastic sampling of the hidden states to build the context vector; *local attention*, also part of the machine translation field, consists of the detection of an attention point and on the use of a window around that point to compute local attention weights.
- **Number of representations:** This category considers the number of different feature representations of the input sequence used in the learning task. At this level, we can consider the following types of attention: *single-representational attention*, which is the most common and consists of learning the attention weights using only a single feature representation; *multi-representational attention*, which involves learning the attention weights using multiple representations (of the same input sequence) and creating a context vector that is the result of a weighted combination of these multiple representations and their attention weights; *multi-dimensional attention*, which consists in computing the attention weights along several dimensions of the representation of the input sequence, which will result in the computation of the relevance of each dimension to the learning task.
- **Number of sequences:** This category takes into account the number of input and output sequences. At this level, we may consider the following types of attention: *distinctive attention*, when the candidate and query states belong to two different input and output sequences respectively; *co-attention*, when we consider multiple input sequences at the same time, and the main goal is to jointly learn their attention weights; *self-attention*, when the query and candidate states belong to the same input sequence.

An analysis of the literature shows that, until very recently, most of the works in attention mechanisms for language, text, and speech derived from the framework proposed by Bahdanau et al. [189]. In [190], the authors trained a long short-term memory network (LSTM) on the task of neural machine translation (i.e., English to German) and presented an explicit comparative study of local and global attention mechanisms. In [191], the authors considered speech recognition as a sequence generation task (i.e., speech to transcription) and developed an attention mechanism with convolutional features to refine the quality of the output sequences. On the other hand, oppositely to this generalised use of RNN-based architectures, in [192], the authors proposed a fully-convolutional network (FCN) with an attention block that uses the dot-product operation trained to generate text from speech. Most recently, the successful introduction of the *Transformer* architecture [170], based

solely on attention mechanisms, started a new paradigm for their study. Recent studies invest their efforts in the computational or design improvements to the baseline Transformer architecture [193, 194, 195], or the proposal of novel Transformer-based architectures [196, 197, 198, 199, 200]. On the other hand, opposite to the trend on the increase of model complexity, Merity [201] revisited the fundamentals of LSTMs and proposed a *single-head attention* that could represent an alternative to the use of Transformer-based architectures.

4.2.2 Attention mechanisms for computer vision

Attention mechanisms for computer vision got their inspiration from the human visual system, which can detect whether certain features of an image are relevant or not [202]. In contrast to the literature on attention mechanisms for language, text, and speech, there are already several interesting review articles on this topic [168, 174, 171, 172, 173]. However, only the work of Guo et al. [173] has presented and discussed a complete taxonomy that organises the different types of attention mechanisms for computer vision into different categories:

- **Channel attention:** This category follows the assumption that, in deep CNNs, different channels in different feature maps usually represent different objects [203]. Therefore, the job of channel attention is to calibrate the weight of each channel adaptively. This mechanism acts as an object selector, thus determining *what to pay attention to*.
- **Spatial attention:** This category is analogous to channel attention. Here, the attention mechanism's job is to calibrate each image region's weight adaptively. This mechanism acts as an adaptive spatial region selection process, thus determining *where to pay attention*.
- **Temporal attention:** This category considers that data has a temporal dimension; therefore, in the case of computer vision tasks, this type of attention mechanism is often used for video processing. The job of temporal attention is to calibrate the weight of each time frame adaptively. This mechanism acts as a dynamic time selection process, thus determining *when to pay attention*.
- **Branch attention:** This category considers multi-branched deep learning architectures. The job of branch attention is to calibrate the weight of each branch adaptively. This mechanism acts as a dynamic branch selection process, thus determining *which to pay attention to*.
- **Channel and spatial attention:** This category combines the advantages of both channel and spatial attention. This mechanism acts as a dynamic spatial region and object selection process, thus determining *what and where to pay attention*.
- **Spatial and temporal attention:** This category combines the advantages of both spatial and temporal attention. This mechanism acts as a dynamic spatial region and time-frame selection process, thus determining *where and when to pay attention*.

The application of attention mechanisms in computer vision is almost contemporaneous, with the first proposals for attention mechanisms in the language, text, and speech domains. Below, we present several representative examples from each category.

With respect to *channel attention*, Hu et al. [3] proposed the *squeeze-and-excitation* block. This module adaptively recalibrates channel-wise feature responses by explicitly modelling dependencies between channels. This strengthens the representational power of CNNs by improving the quality of spatial encodings across the feature hierarchy. These properties have been explored as fundamental components of attention mechanisms in [204, 205]. Still, on the scope of channel attention, a different research line related to the *self-attention* mechanism was further explored in [206]. In this work, Bello et al. proposed *attention augmentation* mechanisms combining convolutions and self-attention by concatenating the convolutional feature maps with a set of feature maps generated by self-attention. This approach showed improvements in image classification and object detection while maintaining the number of parameters. In [207], Wang et al. addressed the topic of one-shot learning and proposed a neural network architecture that uses the semantic representation (i.e., the embedding) of the label to obtain attention maps that are used to generate image features. They also developed a *multiple-attention* scheme to extract meaningful information from the input images and used an auxiliary training set to learn the attention weights.

Concerning *spatial attention*, Mnih et al. [208] designed a novel RNN capable of extracting information from an image or video by adaptively selecting a sequence of regions or locations and only processing these regions at a high resolution. Hence, this method can contribute to a high degree of translation invariance and the ability to process variable input sizes. Dai et al. [209] developed the *deformable convolution*, which adds flexibility to the regular (spatial) sampling grid of the standard convolution. The intuition behind this approach is that CNNs with these modules extract better features related to the objects of interest in the images (with a particular focus on non-rigid objects), thus increasing their robustness to affine transformations. Wang et al. [210] revisited the concept of non-local attention and designed the non-local networks, which compute the response at a given spatial location as a weighted sum of the features at all positions in the features. Hu et al. [211] addressed the issue of *context exploitation* in CNNs and proposed a framework consisting of a pair of operators: the *gather* operator, which aggregates contextual information across large neighbourhoods of each feature map, and the *excite* operator (inspired by [3]), which redistributes the pooled information to local features.

Regarding *temporal attention*, Xu et al. [212] have proposed the *attentive spatial-temporal pooling network* architecture, which operates on video sequences by learning a similarity metric over the features extracted from recurrent and convolutional layers and computing the attention weights on spatial (i.e., regions in each frame) and temporal (i.e., frames over sequences) levels. Similarly, Chen et al. [213] developed an attention mechanism to optimise the similarity learning among short-video snippets, and Zhang et al. [214] proposed the *self-and-collaborative* attention network, which also learns a similarity metric between feature representations of video-pairs.

Regarding *branch attention*, Srivastava et al. [215] proposed the concept of *highway networks*,

which consists of deep neural networks with an LSTM-inspired attention gate that allows for computation paths along which information can flow across many layers (i.e., *information highways*). This work can be considered the precursor of this concept of branch attention. Later, Li et al. [216] designed *selective kernel networks*, which contain multiple CNN branches with different kernel sizes that are fused using an attention mechanism guided by the information in these branches. Zhang et al. [217] presented a deep neural architecture that contains several replicated branches (i.e., the *cardinal* branches) that contain more replicated branches (i.e., the *split* branches). In this work, the information fusion is performed with an attention mechanism inspired by [3]. Chen et al. [218] proposed the *dynamic convolution*, which employs multiple parallel convolution kernels dynamically and aggregates this information through a non-linear attention mechanism.

Concerning *channel and spatial attention*, Woo et al. [4] have designed a convolutional block attention module that contains a channel attention module followed by a spatial attention module. According to the authors, this block can be easily integrated into any architecture and enables the regression of finer features for object detection. Following this research line, Zhao et al. [219] have proposed the *point-wise spatial attention network* for the task of image segmentation. This network processes contextual information from all positions in the feature map and uses a *self-adaptive* attention mechanism to connect all of them. Using generative adversarial networks (GANs), Zhang et al. [220] explored a *self-attention* mechanism to enable both the generator and the discriminator to efficiently model relationships between widely separated spatial regions.

In terms of *spatial and temporal attention*, Du et al. [221] introduced the *recurrent spatial-temporal attention network* in which they used an end-to-end LSTM-based network with a spatio-temporal module to process video data, and developed an attention-based mechanism to fuse this information. Song et al. [222] used RNNs and LSTMs together to process video data and added spatial and temporal attention modules, whose information is later fused. Gao et al. [223] addressed the problem of *visual captioning* with hierarchical LSTM models using spatio-temporal attention to select specific regions or frames to predict the associated words and with adaptive attention to model the importance of visual information or the language contextual information. Yan et al. [224] have also addressed the problem of visual captioning with an architecture based on a CNN-encoder and an LSTM-decoder with spatio-temporal attention to extract spatial and temporal features in a video and guide the decoder to automatically select the most significant regions in the most relevant temporal segments for word prediction.

In addition, several works focus on multi-modal data, thus mixing the computer vision and the language, text, and speech domains. In [202], the authors revisited the concepts of *hard* and *soft* attention. They implemented an architecture consisting of a convolutional block to extract image features and an LSTM network with attention to generate the image description (i.e., automatic image captioning).

Although CNN-based architectures had achieved high performance, the transition to Transformer-based architectures occurred almost naturally. Wu et al. [225] discussed three open challenges that motivated the natural implementation of Transformer-based architectures for computer vision tasks:

convolutions uniformly process all image patches regardless of their importance, which can lead to spatial inefficiency (e.g., image classification models should prioritise foreground over background); not all images have all concepts, therefore, applying high-level filters to all images could be computationally inefficient (e.g., features related to an image of a person may not be present in an image of a flower); and, the convolution operation is not good at establishing relationships between spatially-distant concepts, which is an essential property in computer vision. For this reason, Dosovitskiy et al. [104] have proposed the *Vision Transformer* architecture. Analogously to the original Transformer, to work with a Vision Transformer, one needs to split an image into patches and provide the sequence of linear embeddings of these patches as the input. Contemporaneously, Wu et al. [225] proposed a similar architecture that aims to encode semantic concepts in a smaller number of *tokens* and establish a relationship between spatially-distant concepts through an attention mechanism in this token space. Several modifications of the baseline architecture and their application to different tasks have already been reported [226, 227, 228, 229].

4.3 Literature on Attention Mechanisms for Medical Applications

This Section provides a comprehensive review of the integration of attention mechanisms into algorithms for medical applications. For the sake of clarity, we structure this Section into different types of tasks that can integrate different workflows of the medical domain: medical image classification, medical image segmentation, medical report understanding, and other tasks (medical image detection, medical image reconstruction, medical image retrieval, medical signal processing, and physiological and pharmaceutical research). The results of this work have been disseminated in the form of an article in an international journal: T. Gonçalves, I. Rio-Torto, L. F. Teixeira, and J. S. Cardoso, “A Survey on Attention Mechanisms for Medical Applications: Are We Moving Toward Better Algorithms?,” *IEEE Access*, vol. 10, pp. 98909–98935, 2022 [50].

4.3.1 Medical Image Classification

In this Section, we group the works of the most common approaches of attention mechanisms applied to medical image classification.

4.3.1.1 Breast Lesion Classification

In a multi-database and multi-class classification (i.e., benign, malignant, and normal) approach for breast lesions in ultrasound images, Gheflati and Rivaz [230] studied the impact of using data augmentation, and transfer learning on transformer- and CNN-based architectures, and reported similar, and sometimes improved, results. In a multiple instance learning (MIL) approach, Ilse et al. [231] proposed a gated attention mechanism in deep neural networks for the classification of histopathological images (the authors also investigated the application of this methodology to colon cancer).

4.3.1.2 COVID-19 Classification

In the task of Coronavirus disease 2019 (COVID-19) classification, Han et al. [232] proposed a MIL algorithm for the automated screening of COVID-19 from volumetric chest computed tomography (CT) scans that employs an attention mechanism in the pooling operation to select the instances that may be meaningful for the final classification. Perera et al. [233] proposed a lightweight vision transformer that uses ultrasound images to discriminate between COVID-19, bacterial pneumonia and healthy specimens. Jiang and Lin [234] proposed an architecture that combines two transformer-based models to classify chest X-ray images as COVID-19, pneumonia, and healthy specimens. Liu and Yin [235] used X-ray image data and applied a transformer architecture with transfer learning to train a model capable of discriminating between normal and COVID-19 cases. Park et al. [236] explored the potential of vision transformer architectures in a federated learning setting to classify COVID-19 in chest X-ray images and reported that these architectures are suitable for collaborative learning in medical imaging. Hsu et al. [237] approached chest CT scan data with an architecture that uses transformer-based blocks to learn the context of the CT slices, the information at the pixel level, and the spatial-context features with the aid of a deep Wilcoxon signed-rank test to determine the importance of each slice. Zhang and Wen [238] proposed a framework for 3D CT scans that consists of a U-Net [239] module to segment the lungs and a transformer-based module to extract features and perform classification (i.e., COVID-19 diagnosis). Mondal et al. [240] proposed a vision transformer with a multi-stage transfer learning strategy that uses CT and X-ray image modalities to discriminate between COVID-19, pneumonia, and healthy specimens. Gao et al. [241] tested a vision transformer on 2D and 3D medical image data from CT scans and reported improved results against common CNN architectures for COVID-19 classification (i.e., COVID-19 against non-COVID-19). Similarly, Shome et al. [242] also employed a vision transformer on 2D chest CT X-ray images and tested its predictive performance in both binary (i.e., COVID-19 against healthy specimens) and multi-class (i.e., COVID-19 against bacterial pneumonia against healthy specimens) settings, reporting improved results against common CNN architectures. Ambita et al. [243] used a vision transformer to detect COVID-19 in CT scan images and employed a data augmentation strategy that relied on the generation of synthetic images with a self-attention-based GAN. Park et al. [244] explored the use of a vision transformer architecture for COVID-19 classification that uses low-level features generated using a backbone network. The authors also performed several experiments using chest X-ray databases from different institutions.

4.3.1.3 Whole Slide Image Classification

Lu et al. [245] addressed the problem of glioma subtype classification (i.e., a multi-class task) in whole slide image (WSI) data using a contrastive training framework that employs a CNN backbone to learn relevant patch-level feature representations, and a sparse-attention block to aggregate the features of these multiple patches (i.e., multiple instances). Chen et al. [246] employed a transformer-CNN-based architecture to classify gastric histopathology WSIs in a binary setting (i.e.,

normal against abnormal). Jiang et al. [247] addressed the diagnosis of acute lymphocytic leukaemia through the classification of leukaemic B-lymphoblast cells (i.e., cancer cells) and B-lymphoid precursors (i.e., normal cells) with a transformer-CNN ensemble, and a data enhancement method that tackles the problem of class imbalance. Zheng et al. [248] proposed a novel graph-based vision transformer architecture to classify lung WSIs (i.e., adenocarcinoma, squamous cell carcinoma, and normal histology).

4.3.1.4 Retinal Disease Classification

Bodapati et al. [249] proposed an architecture for the automatic diagnosis of diabetic retinopathy that uses multiple pre-trained CNNs with spatial pooling to extract features from colour fundus retinal images and integrates gated attention blocks to guide the model to focus more on lesion portions of the retinal images while paying less attention to the non-lesion regions. Yu et al. [250] addressed the task of retinal disease classification using fundus image data with a vision transformer trained under a MIL setting. Similarly, in diabetic retinopathy classification, Sun et al. [251] proposed a lesion-aware transformer architecture that jointly learns to detect the presence of diabetic retinopathy and the location of lesion discovery using an encoder-decoder structure. Several authors have addressed the task of diabetic retinopathy recognition in a multi-class setting (i.e., no diabetic retinopathy, mild non-proliferative diabetic retinopathy, moderate diabetic retinopathy, severe non-proliferative diabetic retinopathy, and proliferative diabetic retinopathy) with a vision transformer architecture [252, 253]. Yang et al. [254] used a hybrid CNN-transformer architecture to tackle ophthalmic image data in a multi-class setting (i.e., normal, diabetes, glaucoma, cataract, age-related macular degeneration, hypertension, myopia, and other abnormalities) using different data pre-processing strategies.

4.3.1.5 Miscellaneous Classification Tasks

In lung cancer classification in screening CT, Al-Shabi et al. [255] have extended the study of non-local attention and proposed an architecture that is initially trained on simple examples and gradually grows to increase its ability to handle the task at hand, as the classification task becomes more difficult. Moranguinho et al. [256] applied this methodology for the classification of lung biopsy histopathological images with respect to cancer classification and used post-model interpretability algorithms to assess the regions of interest for the predictions of their trained deep neural network. He et al. [257] proposed a deep learning architecture for depression recognition that combines a backbone CNN to extract features, a local attention module that focuses on parts of the input images, a global attention module to learn global patterns from the entire input images, and a weighted spatial pyramid pooling layer to learn the depression patterns after the feature aggregation operation. Dai et al. [258] explored the task of multi-modal and multi-class classification in head, neck, and knee magnetic resonance imaging (MRI) data using a hybrid CNN-transformer model, where the CNN module is used as a low-level feature extractor. Datta et al. [259] performed a comparative study of the effect of *soft-attention* combined with different backbone architectures in skin cancer

classification and reported several advantages against conventional methodologies. Barhoumi et al. [260] proposed a model that aggregates several feature maps extracted using multiple Xception CNNs [261] and uses these features to train a vision transformer for the intracranial haemorrhage classification problem, using CT images. Liang and Gu [262] added an attention mechanism to a backbone network based on the ResNet [263] to aid the diagnosis of Alzheimer's disease in brain MRI data.

4.3.2 Medical Image Segmentation

Using well-known data sets related to different use cases and working on top of [3], Roy et al. [264] verified that the addition of a spatial-channel squeeze and excitation block works as an attention mechanism in FCNs and improves the quality of the segmentation maps. Interestingly, in most recent papers, the authors approach the problem of medical image segmentation using encoder-decoder structures that generally comprise a CNN and a transformer to extract volumetric spatial feature maps, perform global feature modelling and predict refined segmentation maps [265, 266, 267, 268], although there are already some works that replace the CNN-based modules at the encoder or decoder levels and integrate other attention mechanisms to extract features and model long-range dependencies [269]. More recent methodologies on medical image segmentation are taking advantage of a hybrid use of the vision transformer and the U-Net with improved results regarding the quality of segmentation maps [270, 271]. Wu et al. [272] proposed a *feature adaptive* model comprising a dual encoder with CNN and transformer branches to capture local features and global context information simultaneously. The U-Net architecture [239] has high relevance in biomedical image segmentation. The high flexibility of this network's structure allows the integration of several computing blocks, such as attention mechanisms (see Figure 4.2). Hence, as described below, most authors employ a similar structure in their applications.

4.3.2.1 Breast Lesion Segmentation

Using breast ultrasound images, Vakanski et al. [273] integrated an attention block into the well-known architecture U-Net to learn semantic representations that prioritise spatial regions for breast tumour segmentation. In breast tumour segmentation, using ultrasound imaging data, Zhu et al. [274] developed a transformer-based architecture that incorporates prior information of the region of tumours to obtain accurate segmentation. Following the most recent approaches of transformer-based architectures for medical image analysis, Liu et al. [275] proposed a U-Net-like model that integrates an attention module to focus on the tumour region and combines CNN and transformer blocks to extract and process features that may lead to improved segmentation maps.

4.3.2.2 Computational Pathology

Using microscopic images, Prangemeier et al. [276] proposed an architecture based on a CNN-encoder and transformer-decoder for the *classification* and *instance segmentation* of yeast cells.

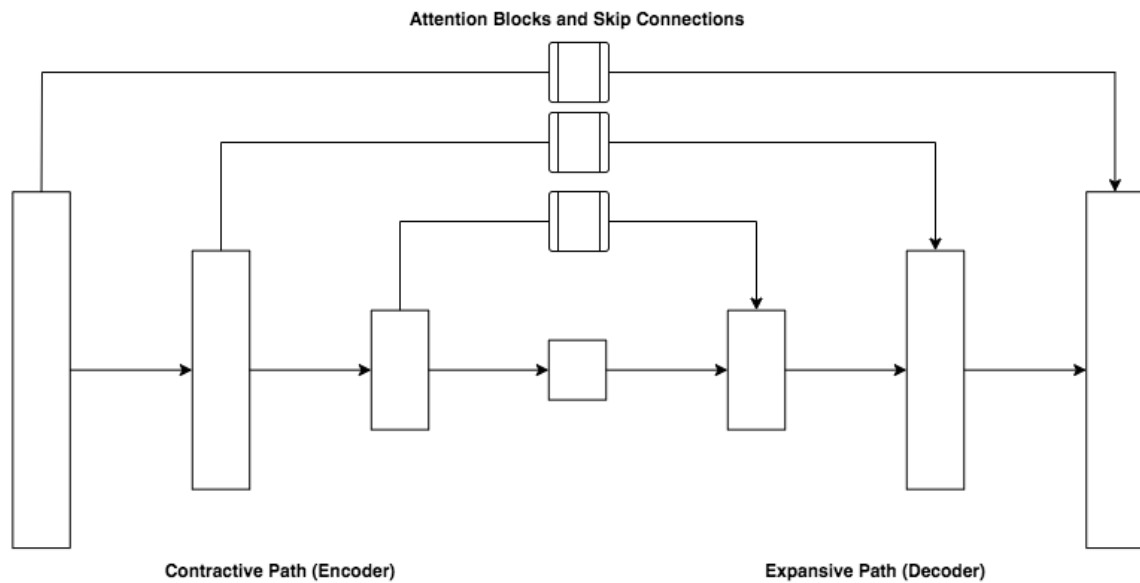


Figure 4.2: U-Net architecture with attention blocks, as proposed in [2]. Please note that rectangles represent the feature maps of the encoder and decoder, horizontal arrows represent the regular flow of information, vertical elbow arrows represent the skip connections, and squares represent the attention blocks.

Using microscopy images of corneal endothelial cells, Zhang et al. [277] revisited the problem of cell segmentation with the proposal of the multi-branch hybrid transformer network, which follows the main structure of its predecessors (i.e., it contains both CNN and transformer modules) to extract and process local and spatial features that may lead to refined segmentation maps. Shao et al. [278] addressed MIL using a transformer-based architecture that explores morphological and spatial information and considers the correlation between instances. The authors tested their framework in different computational pathology problems. Using high-resolution images, Nguyen et al. [279] performed a comparative analysis using architectures based on CNN and transformer modules and reported improved results with the transformer-based approaches. In a similar approach, using hyperspectral image data, Yun et al. [280] proposed an encoder-decoder architecture with CNN and transformer blocks to extract and model spatial and spectral features, achieving competitive results against other methodologies.

4.3.2.3 Cardiac Segmentation

Li et al. [281] addressed the task of cardiac image segmentation using MRI data from different domains with a generative adversarial model with an attention loss, proposed to translate the images from existing source domains, and a stack of data augmentation techniques to simulate real-world transformation to boost the segmentation performance for unseen domains. Concurrently, Kong and Shadden [282] tackled the same problem with a GAN with a cycle consistency loss (i.e., CycleGAN [283]) to generate images in different styles, exchanging the low-frequency features of

images from different domains, and using an architecture based on an attention-gated U-Net that should learn to focus on cardiac structures of varying shapes and sizes while suppressing irrelevant regions. The community has followed these strategies in other contexts as well [284]. In the task of cardiac segmentation, using ultrasound images (i.e., echocardiography), Deng et al. [285] developed an architecture that combines a CNN-based encoder-decoder to extract multi-level features, a transformer structure to fuse these features, and a patch embedding layer created to reduce the number of parameters of the embedding layer and the size of the token sequence. Chen et al. [286] proposed a U-Net-like architecture that studied the potential of several settings (e.g., transformer as encoder, hybrid CNN-transformer as encoder) that reported increased performances for cardiac segmentation (the authors also reported increased performances on multi-organ segmentation). Huang et al. [287] also proposed a transformer-based model that tackles several tasks: the alignment of features, the enhancement of the long-range dependencies and local context, and the modelling of the long-range dependencies and local context of multi-scale features. They experimented with this architecture on cardiac segmentation databases (also on multi-organ segmentation).

4.3.2.4 Multi-task Segmentation

Oktay et al. [2] addressed the task of multi-organ segmentation (i.e., liver, pancreas, spleen) and worked on top of the U-Net architecture by integrating attention modules in the skip connections to promote the learning of relevant features while suppressing feature activations in irrelevant regions. Yao et al. [288] proposed a U-Net-based architecture with deep feature concatenation and an attention mechanism branched into several attention gates for scleral blood vessel segmentation. Following this methodology, Chang et al. [289] added a transformer module to the encoder block of the baseline model and reported improved results on multi-organ segmentation tasks. Using CT images, Xie et al. [290] proposed a framework that connects a CNN and a transformer to extract both feature representations and model the long-range dependency on these feature maps. This model also employs a self-attention mechanism that focuses on specific positions of the image, thus reducing the computational complexity. Similarly, for kidney segmentation using CT images, Shen et al. [291] also proposed an encoder-decoder architecture that leverages the interaction of CNN and transformer modules to learn multi-scale features to achieve improved segmentation results. Following this work, Zhang et al. [292] developed an architecture for multi-organ segmentation that runs a CNN-based encoder and transformer-based segmentation network in parallel and fuses the features from these two branches to make predictions jointly. Tang et al. [293] proposed a U-Net-shaped architecture with a transformer module for multi-organ segmentation with 3D images and employed a specific pre-training strategy with contrastive learning, masked volume inpainting, and 3D rotation prediction. Using images from different modalities (i.e., MRI and CT) and organs (i.e., brain cortical plate, hippocampus, pancreas), Karimi et al. [294] proposed an architecture based entirely on transformers that uses self-attention between neighbouring image patches, without requiring any convolution operations. Cao et al. [295] also proposed a U-Net architecture composed solely of transformer-based modules for multi-organ segmentation that

uses skip-connections for local-global semantic feature learning. On colonoscopy images, Sang et al. [296] proposed an architecture that takes advantage of a ResNeSt *backbone* [217] connected to the coupled U-Nets [297] architecture and several attention gates to combine multi-level features and yield accurate polyp segmentation. In a multi-task segmentation approach, Zhang et al. [298] developed a method that seeks to integrate multi-scale attention and CNN feature extraction using a pyramidal network architecture by working on multi-resolution images. Lin et al. [299] proposed an encoder-decoder architecture comprised of hierarchical transformer-based modules to model the long-range dependencies and the multi-scale context connections during the process of down-sampling and up-sampling. Li et al. [300] proposed a different framework based on transformers that uses a squeeze-and-excitation attention module to both regularise the self-attention mechanism of transformers and learn diversified representations. Ranjbarzadeh et al. [301] addressed brain tumour segmentation in multi-modal MRI data using a cascade CNN that processes both local and global features in two different routes, combined with a distance-wise attention mechanism that considers the effect of the location of the centre of the tumour and the brain. Ribeiro [302] also addressed the task of brain tumour segmentation and proposed an architecture inspired by the U-Net that integrates an attention mechanism with adaptive-scale context to capture multi-scale information by processing parallel branches with different scales. Petit et al. [303] designed a U-shaped architecture for image segmentation with self-attention and cross-attention from transformers (i.e., U-transformer). They performed experiments on two abdominal CT-image databases and obtained superior performance compared to U-Net-based models. Regarding skin lesion segmentation, Wang et al. [304] proposed a *boundary-aware* transformer that aims to take advantage of *boundary-wise priors* through a boundary-wise attention gate that highlights the ambiguous boundaries to generate attention maps to guide the training. For tooth root segmentation, Li et al. [305] proposed a transformer-based architecture with the structure of the U-Net backbone and optimised the training of this model with a Fourier Descriptor loss function that takes advantage of prior knowledge related to shape. Li et al. [306] explored the same task through an anatomy-guided multi-branch transformer with multi-head attention that employs a polynomial curve fitting segmentation strategy (based on keypoint detection) to extract anatomy features. Following the work described in [306], Guo et al. [307] implemented a similar approach for retinal vessel segmentation. Several other approaches were proposed for this task. Wang et al. [308] introduced a network that focuses on refining segmentation errors by using an attention mechanism that considers the differences between the ground truth and the predicted masks as the (source of) supervision for learning the attention maps. Guo et al. [309] proposed a network that builds upon *residual* blocks and spatial attention to promoting the extraction of complex vascular features while alleviating over-fitting. Du et al. [310] proposed a U-Net-based architecture that contains *pyramid pooling* [311] modules and integrates an attention mechanism in its skip connections to promote the ability to obtain global information and efficient learning of semantic features. Gao et al. [312] developed a deep network that combines a feature fusion *residual* module with channel and spatial attention mechanisms to promote the extraction of low and high-dimensional features and effectively explore feature dependencies at spatial and

channel levels. Amer et al. [313] created a U-Net architecture that comprises *residual* blocks and a multi-scale spatial attention module to capture and select multi-scale context information while ensuring an effective fusion of this information. Zhao et al. [314] also proposed a U-Net architecture that focuses on extracting better spatial and channel-level features. After revisiting the work described in [306], Jin et al. [315] employed a 3D U-Net architecture with *residual* blocks and an attention module to extract volumes of interest in liver and brain CT imaging data and perform tumour segmentation. Sobirov et al. [316] applied the transformer-based model proposed by Hatamizadeh et al. [268] to the tasks of head and neck tumour segmentation using multi-modal data and compared their results with CNN-based approaches. Yan et al. [317] also employed a U-Net-based structure for the task of multi-organ segmentation in 3D medical image data; however, with slightly different changes: in their approach, they used a CNN-encoder and CNN-decoder with a transformer model in between to fuse contextual information in the neighbouring image slices.

4.3.3 Medical Report Understanding

In this Section, we group works related to medical report analysis (focused on natural language processing).

4.3.3.1 Medical Report Generation

Using two medical databases containing radiological and pathological images, Jing et al. [318] proposed a methodology consisting of a multi-task deep learning model that jointly performs the prediction of tags and the generation of paragraphs, a *co-attention* mechanism for locating regions containing abnormalities and generating narrations for them, and a hierarchical LSTM model for generating long paragraphs. One of the first mentions of using transformer architectures to generate medical image reports is described in [319]. In this work, Xiong et al. used X-ray data and proposed a hierarchical framework trained with reinforcement learning. This framework consists of an encoder that extracts visual features through a bottom-up attention mechanism aimed at identifying regions of interest and extracting top-down visual features and a transformer-based non-recurrent captioning decoder aimed at generating a coherent paragraph of medical image reports. Lovelace and Mortazavi [320] developed a chest X-ray report generation algorithm that consists of two training stages: the training of the report generation model (based on a transformer architecture) using a standard language generation objective and, after, the fine-tuning of this model on clinical observations extracted from reports sampled from the same model to regularise the degree of coherence between the observations from the generated and ground truth reports. In a similar task, Srinivasan et al. [321] proposed a deep framework that uses a CNN to extract features of the images and generate tag embeddings for each image, and uses transformers to encode the image and tag features with self-attention to get a finer representation. Alternatively, Miura et al. [322] applied a transformer-based architecture trained with reinforcement learning, which comprises two different rewards, namely, one that encourages the system to generate radiology domain entities

consistent with the reference and one that uses natural language inference to encourage these entities to be described in inferentially consistent ways. On the other hand, Chen et al. [323] proposed to generate radiology reports with a memory-driven transformer, where a relational memory is designed to record key information of the generation process and a memory-driven conditional layer normalisation is applied to incorporate the memory into the decoder of the transformer. Also, in automatic chest X-ray report generation, Liu et al. [324] created a *contrastive-attention* methodology that compares a given input image with normal images to distil contrastive information that better represents the visual features of abnormal regions, thus providing more accurate descriptions for an interpretable diagnosis and improving the quality of the generated reports. Najdenkoska et al. [325] tackled the same problem with *variational topic inference* that consists of modelling a set of topics as latent variables to guide sentence generation by aligning image and language modalities in a latent space. Each topic contributes to the generation of a sentence in the report. The entire pipeline is refined with a visual attention module that enables the model to attend to different locations in the image and generate more informative descriptions. Alfarghaly et al. [326] proposed a transformer-based pipeline that involves the use of a fine-tuned CheXNet model [327] to predict specific tags from the image, the computation of weighted semantic features from the predicted tag's pre-trained embeddings, and the conditioning of a pre-trained distilled GPT2 [328] model on the visual and semantic features to generate the full medical reports. Amjoud et al. [329] proposed an encoder-decoder framework that combines the benefits of transfer learning for feature extraction and the transformer architecture for medical report generation. Tipirneni et al. [330] proposed a transformer-based architecture to address multivariate clinical time series with missing values. A different approach was discussed in [331], where the authors propose an encoder-decoder architecture that comprises both CNN and transformer modules and aims to explicitly quantify the inherent visual-textual uncertainties existing in the multi-modal task of radiology report generation both at the report and sentence levels, and explore a novel method to measure the semantic similarity of radiology reports, which seems to capture the characteristics of diagnosis information better. A different research line focuses on imitating the working patterns of radiologists, which typically consist of examining the abnormal regions and assigning the disease topic tags to these regions and then relying on the years of prior medical knowledge and prior working experience to write reports [332]. Hence, in [332], the authors proposed a *posterior-and-prior knowledge exploring-and-distilling* approach that consists of three modules: the *posterior knowledge explorer*, which aims to provide explicit abnormal visual regions to alleviate visual data bias; the *prior knowledge explorer*, which aims to explore the prior knowledge from the prior medical knowledge graph (i.e., medical knowledge) and prior radiology reports (i.e., working experience) to alleviate textual data bias; and the *multi-domain knowledge distiller* that aggregates this processed knowledge and generates the final reports. A very close framework was described by Liu et al. [333], in which the authors proposed an *unsupervised model knowledge graph auto-encoder* which accepts independent sets of images and reports during training, and consists of three modules: the *pre-constructed knowledge graph*, that works as the shared latent space and aims to bridge the visual

and textual domains; the *knowledge-driven encoder* which projects medical images and reports to the corresponding coordinates in that latent space; and the *knowledge-driven decoder* that generates a medical report given a coordinate in that latent space. This modular structure is also employed by Nguyen et al. [334], which added three complementary modules: a CNN-based *classification* module that produces an internal checklist of disease-related topics (i.e., *the enriched disease embedding*); a transformer-based *generator* that generates the medical reports from the enriched disease embedding and produces a *weighted embedding representation*; and an *interpreter* that uses the weighted embedding representation to ensure consistency concerning disease-related topics. Similarly, You et al. [335] proposed a framework which includes two different attention-based modules: the *align hierarchical attention* module that first predicts the disease tags from the input image and then learns the multi-grained visual features by hierarchically aligning the visual regions and disease tags; and the *multi-grained transformer* module that uses the multi-grained features to generate the medical reports. Still, on this line, Hou et al. [336] proposed a CNN-RNN-based medical transformer trained end-to-end, capable of extracting image features and generating text reports that aim to fit seamlessly into clinical workflows.

4.3.3.2 Miscellaneous Tasks

Recently, Zhou et al. [337] proposed a cross-supervised methodology that acquires free supervision signals from original radiology reports accompanying the radiography images and employs a vision transformer designed to learn joint representations from multiple views within every patient study. Regarding automatic surgical instruction generation, Zhang et al. [338] introduced a transformer-based encoder-decoder architecture trained with *self-critical* reinforcement learning to generate instructions from surgical images and reported improvements against the existing baselines over several caption evaluation metrics. Wang et al. [339] proposed a transformer-base training framework designed for learning the representation of both the image and text data, either paired (e.g., images and texts from the same source) or unpaired (e.g., images from one source coupled with texts from another source). They applied their methodology to different chest X-ray databases and performed experiments in three applications (i.e., classification, retrieval, and image regeneration).

4.3.4 Other Tasks

This Subsection includes medical tasks whose interest in the field is currently increasing: medical image detection, medical image reconstruction, medical image retrieval, medical signal processing, and physiology and pharmaceutical research.

4.3.4.1 Medical Image Detection

Regarding colorectal polyp detection, Shen et al. [340] approached several colonoscopy databases and proposed an architecture constituted by a CNN for feature extraction, transformer encoder layers interleaved with convolutional layers for feature encoding and recalibration, transformer

decoder layers for object querying, and a feed-forward network for detection prediction. Similarly, Liu et al. [341] addressed the same task using an encoder-decoder architecture with a ResNet50-based CNN backbone as the encoder and a transformer-based module as the decoder and Mathai et al. [342] addressed the task of lymph node detection using MRI data employing analogous strategies.

4.3.4.2 Medical Image Reconstruction

Gong et al. [343] addressed the reconstruction of linear parametric images from dynamic positron emission tomography (PET) and implemented a 3D U-Net [344] model with an attention layer that focuses on prior anatomical information during training to improve the quality of the reconstruction of dynamic PET images of the human brain. Liang and Gu [262] also explored the reconstruction of brain MRI data as a regularisation strategy to aid the diagnosis of Alzheimer's disease.

4.3.4.3 Medical Image Retrieval

Grundmann et al. [345] and Kim et al. [346] developed new methods to perform an automatic retrieval and analysis of clinical notes. Li et al. [347] and Ma et al. [348] developed new models for the automated evaluation of root canal therapy and significant stenosis detection in coronary CT angiography of coronary arteries, respectively.

4.3.4.4 Medical Signal Processing

Nauta et al. [349] explored the combination of attention mechanisms and causal discovery techniques. Using simulated functional MRI data containing blood-oxygen-level dependent data sets for 28 different underlying brain networks, this work proposed a deep learning framework based on attention that learns a causal graph structure by discovering causal relationships in observational time series data.

4.3.4.5 Physiology and Pharmaceutical Research

Deep learning has also revealed itself to be a popular framework for physiology and pharmaceutical research (e.g., drug combination prediction, protein residue-residue contact prediction). Naturally, as this field also involves high-stakes decisions, the increase in the predictive performance of deep learning architectures in several problems of this topic [350, 351, 352, 353] motivated the need to explain the internal mechanisms of these algorithms and improve their explainability. Interestingly, the strategy employed by machine learning practitioners is based on attention mechanisms. Filipavicius et al. [354] studied the use of transformer-based language models for protein classification using long sequences. The authors also performed several experiments on the compression of the protein sequences and prepared new pre-training and protein-protein binding prediction databases. Chen et al. [355] presented an attention-based (i.e., regional and sequence attention) CNN for protein contact prediction and for identifying interpretable patterns that contain useful insights

into the key fold-determining residues in proteins. Wang et al. [356] recently proposed a deep learning framework based on graph neural networks with an attention mechanism to identify drug combinations that can effectively inhibit the viability of specific cancer cells. Khan et al. [357] used a transformer-based architecture to analyse high-dimensional gene expression data and reported improved results on lung cancer sub-types classification.

4.4 Attention Mechanisms for Medical Image Classification

The results of this work have been disseminated in the form of an article in an international journal: T. Gonçalves, I. Rio-Torto, L. F. Teixeira, and J. S. Cardoso, “A Survey on Attention Mechanisms for Medical Applications: Are We Moving Toward Better Algorithms?,” *IEEE Access*, vol. 10, pp. 98909–98935, 2022 [50]. The code related to the implementation of this work is publicly available in a GitHub repository¹.

4.4.1 Description and Contributions

An attentive look at the literature described in the previous section will show us that the quality assessment of the algorithms powered by attention mechanisms is often based on predictive performance (i.e., accuracy) rather than showing that these algorithms are moving towards increasing transparency or interpretability. However, the authors must consider framing their proposals to fit the application context in a high-risk context like healthcare. In this work, we present an experimental set involving three different use cases (i.e., breast lesion, chest pathology, and skin lesion), wherein the main goal is to provide insight into the impact of using attention mechanisms in deep learning algorithms for medical image applications. We assess this impact from the perspective of predictive performance (i.e., accuracy) and then report post-model attribution maps obtained with different interpretability frameworks to visualise the image regions that contributed most to the final predictions. We acknowledge that the visual assessment of these attribution maps is subject to a high degree of subjectivity. Nonetheless, we argue that this analysis can serve as an indirect evaluation of the degree of interpretability of these models. We approach this topic with a critical view and ask the following research questions:

1. Will attention mechanisms automatically improve the predictive power of deep learning algorithms for medical image applications?
2. What is the impact of integrating attention mechanisms on model complexity?
3. Can we improve the degree of interpretability of deep learning models solely through attention mechanisms?
4. How practical is it to design and build attention mechanisms for deep learning applications?

¹<https://github.com/TiagoFilipeSousaGoncalves/survey-attention-medical-imaging>

To answer these, we develop an experimental protocol with three case studies on medical image classification, using established deep learning architectures with and without state-of-the-art attention blocks and a transformer-based model. We address the open challenges of interpretability and transparency with the help of *post-hoc* explanation methods and perform a visual analysis of the impact of different attention mechanisms in these outputs. The main contributions of this work are:

1. An experimental case study on medical image classification with three different use cases that focus on the integration of attention mechanisms into established deep learning architectures, the analysis of their predictive power, and a visual assessment of their saliency maps generated by *post-hoc* explanation methods;
2. A critical analysis of the claims and potentialities of attention mechanisms presented in the literature;
3. A discussion on the future challenges of medical applications and how they may benefit from attention mechanisms.

4.4.2 Data

In this work, we used APTOS2019, ISIC2020 and MIMIC-CXR datasets (see Chapter 3). To create the binary version of the APTOS2019 dataset, we grouped all the non-R0 classes into the “diabetic retinopathy” class. Table 4.1 presents the number of training, validation and test examples and the meaning of the labels per dataset.

Table 4.1: Data splits and meaning of labels for each data set.

Data set	Splits			Labels	
	Train	Validation	Test	0	1
APTOS2019	2334	778	550	Normal	Diabetic Retinopathy
ISIC2020	19593	6568	6965	Benign	Malign
MIMIC-CXR	61203	534	1072	Normal	Pleural Efusion

4.4.3 Methodology

We conducted all the experiments using the PyTorch library [358] for Python.

4.4.3.1 Backbone Models

In the first phase of this study, we trained two deep learning backbones (i.e., DenseNet-121 [359], ResNet-50 [263]) on each data set. DenseNet-121’s architecture [359] allows connecting all layers (with matching feature-map sizes) directly with each other, thus improving the flow of information and gradients throughout the network and facilitating their training. ResNet50’s architecture

introduced the *deep residual learning framework*, which consists of adding *skip connections* that perform *identity mapping*, and adding their outputs to the outputs of the stacked layers. We use these backbone models in our experiments because they are well-known and widely used in the deep learning community. In this work, we refer to these models as *DenseNet-121* and *ResNet-50*.

4.4.3.2 Squeeze-and-Excitation Attention Block

In the second phase, we adapted the *Squeeze-and-Excitation* (SE) attention block [3] (see Figure 4.3) to each of the backbones and trained these new architectures on the three data sets. According to the authors of the original paper [3], this attention block was designed to adaptively recalibrate channel-wise feature responses by explicitly modelling interdependencies between channels. Since one of the original implementations uses the ResNet-50, we used the author's implementation in our study, which consists of integrating the SE block between the *residual layers* and the activation function (in this case, the *rectified linear unit* (ReLU)). Following this rationale, in the case of DenseNet-121, we integrated the SE block between the *dense* and the *transition* blocks. We decided to use the SE block in our experiments because it was one of the first proposals of channel attention for computer vision, and it is widely used by the community as a comparison reference. Besides, since we are building our case studies in medical image classification using RGB images, our interests also rely on studying the effects of attention on the channel dimension. In this work, we refer to these models as *SEDenseNet-121* and *SEResNet-50*.

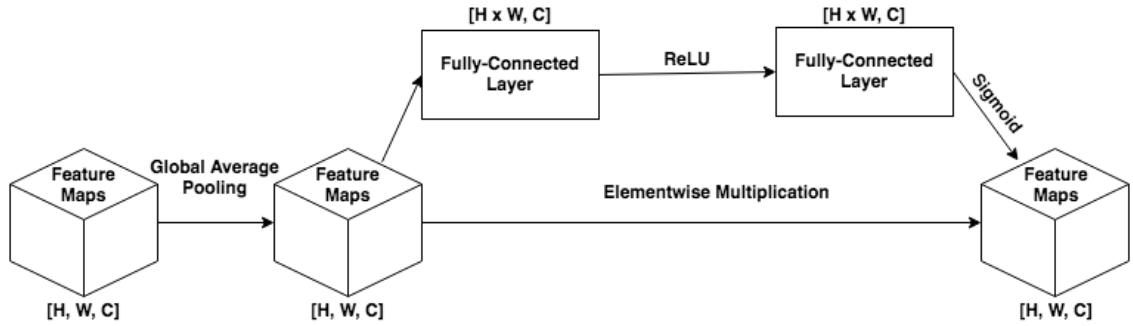


Figure 4.3: Block diagram of the *Squeeze-and-Excitation* (SE) attention block, according to [3].

4.4.3.3 Convolutional Block Attention Module

In the third phase, we adapted the *Convolutional Block Attention Module* (CBAM) [4] (see Figure 4.4) to each of the backbones and trained these new architectures on the three data sets. According to the authors of the original paper, the CBAM mechanism integrates two specific attention blocks: the *channel attention module*, which aims to produce a channel attention map by exploiting the inter-channel relationship of features and is considered as a feature detector (see Figure 4.5); and the *spatial attention module*, which aims to generate a spatial attention map by

utilising the inter-spatial relationship of features, thus being complementary to the channel attention (see Figure 4.6). Like the SE attention block, we integrated the CBAM mechanism at the same topographical location of the architectures of the backbone models. Following the decision on the SE block, we considered it important to study the spatial dimension of images. Therefore, CBAM integrates both the attention and spatial dimensions in its pipeline, so we can perceive this module as an upgrade to the SE block. Furthermore, it is important to note that this module is also well-known in the community, and, therefore, it is validated as well. In this work, we refer to these models as *CBAMDenseNet-121* and *CBAMResNet-50*.

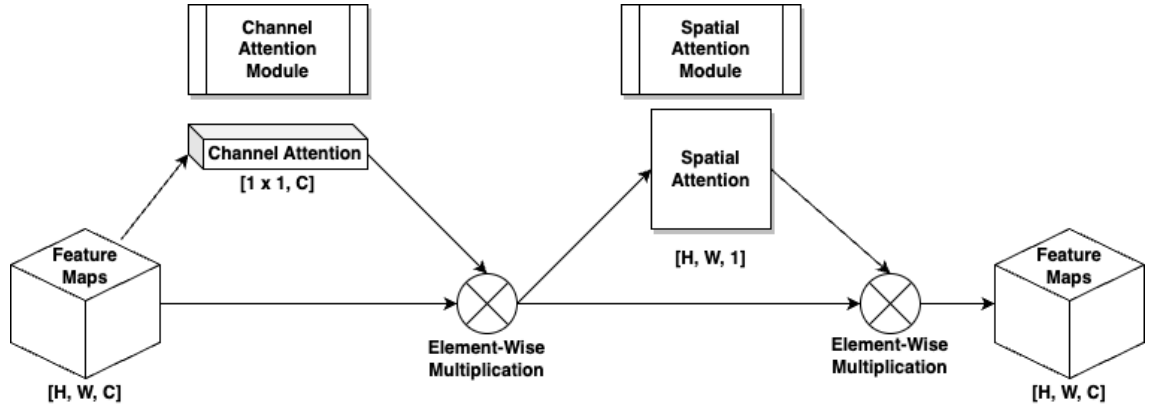


Figure 4.4: Block diagram of the *Convolutional Block Attention Module* (CBAM), according to [4].

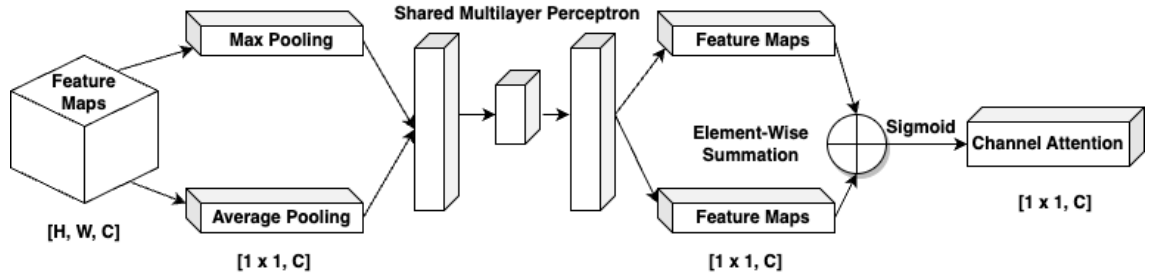


Figure 4.5: Block diagram of the *Channel Attention Module* of the *Convolutional Block Attention Module* (CBAM), according to [4].

4.4.3.4 Data-efficient image Transformer

In the fourth phase, we tested a transformer-based architecture composed solely of attention mechanisms. The Data-efficient image Transformer (DeiT) [229] is an architecture inspired by the vision transformer [104], and trained with fewer parameters. In this case, we used the *DeiT-Ti* variation [229], which has a comparable number of parameters against the chosen CNN backbones. Aside from being validated by the community (as in previous examples), this vision transformer architecture has a proportional number of parameters (see Table 4.4) to the backbone models, thus

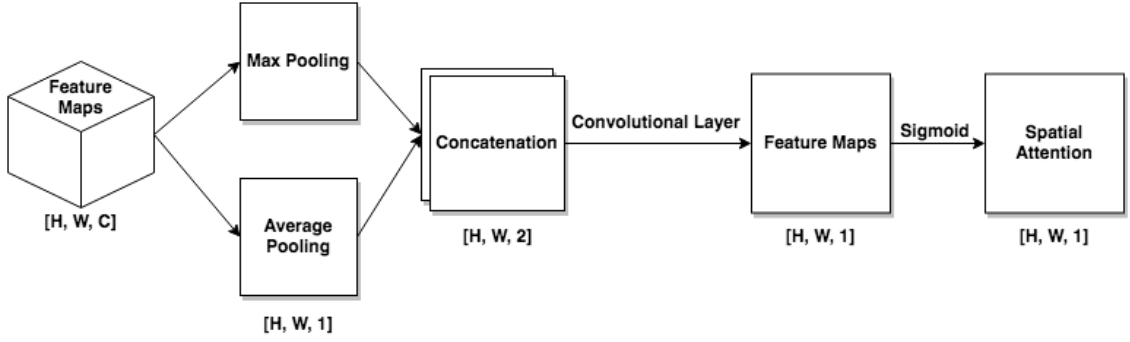


Figure 4.6: Block diagram of the *Spatial Attention Module* of the *Convolutional Block Attention Module* (CBAM), according to [4].

allowing us to mitigate eventual model complexity effects on the predictive performance of this model. In this work, we refer to this model as *DeiT*.

4.4.3.5 Post-model Interpretability

In the fifth phase, we generated visual explanation maps (i.e., saliency maps) using the *Deep Learning Important Features* (DeepLIFT) [30] and the *Layer-wise Relevance Propagation* (LRP) [34] *post-hoc* methods. DeepLIFT [30] compares the activation of each neuron to its related *reference activation*, and assigns contribution scores according to the difference. LRP [34] is a methodology that aims to create visualisations of the contributions of single pixels to predictions. We decided to employ this strategy as a proxy for the degree of interpretability of models, thus allowing our work to be comparable to current research. We used the Captum [360] library for Python to generate these *post-hoc* saliency maps for the backbones and their attention-based versions. Regarding the *DeiT*, we refer the reader to [361], which proposed a framework to generate LRP attributions for transformer-based architectures.

4.4.3.6 Data Processing

Regarding the data processing pipeline, all the images were resized to the dimensions of 224×224 , and a z-normalisation was applied to each RGB channel. The *mean* and *standard deviation* parameters used in this operation depend on the initialisation of the model weights. In the case of DenseNet-121, ResNet-50, SEDenseNet-121, SEResNet-50, CBAMDenseNet-121 and CBAMResNet-50, the weights were initialised from a pre-training on the ImageNet database [362] with $\text{mean} = [0.485, 0.456, 0.406]$, and $\text{std} = [0.229, 0.224, 0.225]$. In the case of *DeiT*, the weights were initialised from a pre-training on the ImageNet database with $\text{mean} = [0.500, 0.500, 0.500]$, and $\text{std} = [0.500, 0.500, 0.500]$.

4.4.3.7 Data Augmentation

Regarding the data augmentation strategy, we created a pipeline composed of several random affine transformations (i.e., random rotations, random translations, random scaling and random horizontal flips). Table 4.2 presents the parameters of data augmentation used in our experiments.

Table 4.2: Data augmentation parameters used during the training step of all the phases.

Parameter	Value
Angle of rotation in degrees	$[-10, 10]$
Horizontal translation shift	0.05
Vertical translation shift	0.1
Scaling factor	$[0.95, 1.05]$
Horizontal flip probability	0.5

4.4.3.8 Training

We trained all the models in the different datasets using the same optimisation hyper-parameters. During training, we saved the best model weights according to a decrease in the validation loss. Table 4.3 presents the optimisation hyper-parameters used in our experiments. We also conducted a *low data regime* experiment; hence, all the models were trained using 1%, 10%, 50% and 100% of the training data. In imbalanced data sets (i.e., ISIC2020, MIMIC-CXR), we applied class-weights in the loss function.

Table 4.3: Set of hyper-parameters used during the training step of all the phases. Please note that we present the batch size as a closed interval since we had to adapt its value during some of the training runs due to computational constraints (e.g., availability of GPU RAM).

Hyper-parameter	Value
Number of epochs	300
Loss function	Cross-entropy
Optimiser	Adam
Learning rate	1×10^{-6}
Batch size	$[2, 32]$

4.4.4 Results

4.4.4.1 Predictive Performance

Figures 4.7, 4.8 and 4.9 respectively present the predictive performance (i.e., accuracy) results of the different models on the APTOS2019, ISIC2020 and MIMIC-CXR data sets, using different percentages of training data.

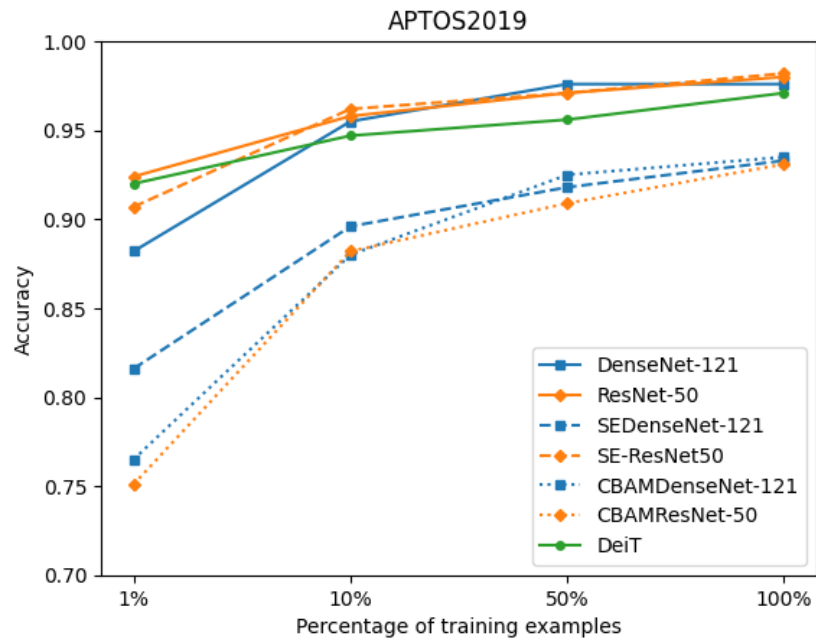


Figure 4.7: Predictive performance (i.e., accuracy) results of the different models on the APTOS2019 data set, using different percentages of training data.

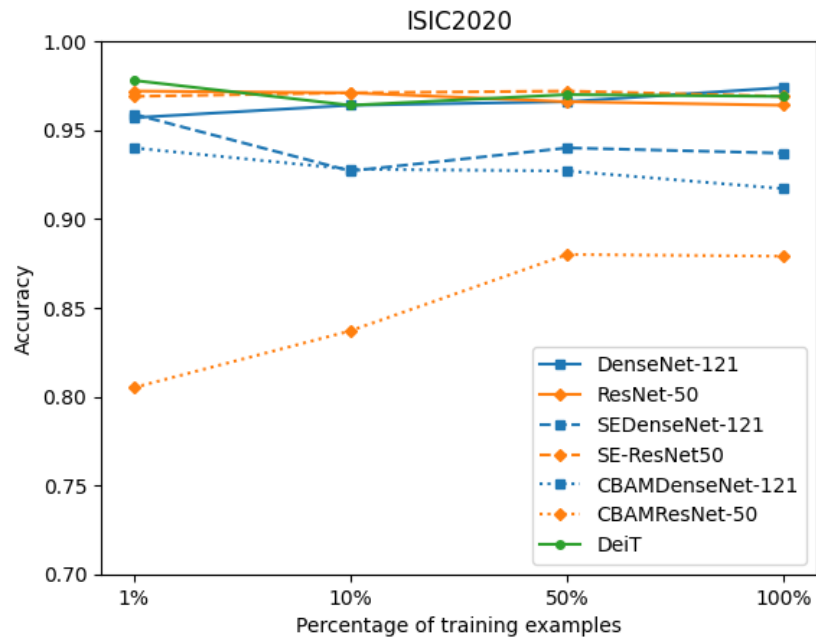


Figure 4.8: Predictive performance (i.e., accuracy) results of the different models on the ISIC2020 data set, using different percentages of training data.

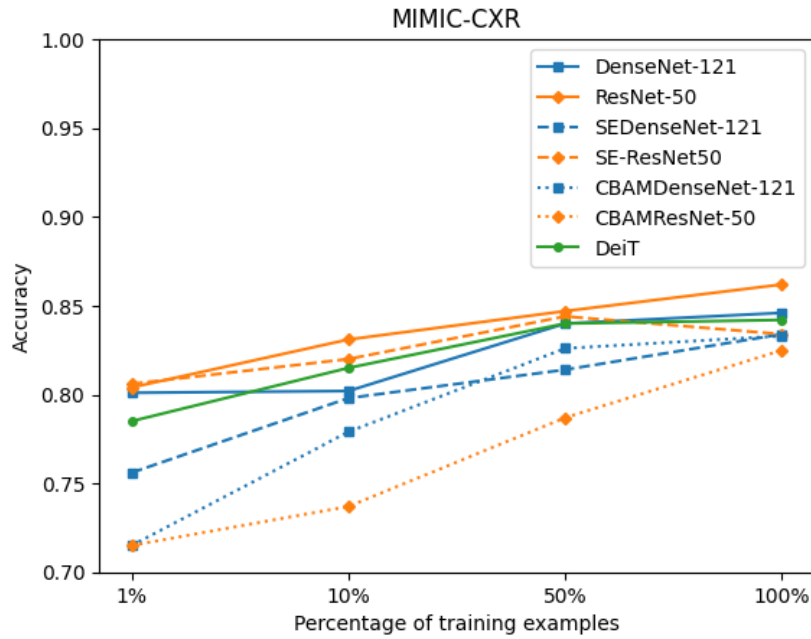


Figure 4.9: Predictive performance (i.e., accuracy) results of the different models on the MIMIC-CXR data set, using different percentages of training data.

4.4.4.2 Model Complexity

Table 4.4 presents the model complexity (i.e., the number of parameters) information for the models used in our experiments.

Table 4.4: Model complexity (i.e., the number of parameters) information for the models used in our experiments.

Model	Number of parameters
DenseNet-121	7,054,210
ResNet-50	23,512,130
SEDenseNet-121	7,357,314
SEResNet-50	26,027,074
CBAMDenseNet-121	7,360,706
CBAMResNet-50	26,044,722
DeiT	5,486,786

4.4.4.3 Post-Model Interpretability

Tables 4.5, 4.6 and 4.7 present examples of LRP and DeepLIFT *post-hoc* saliency maps obtained for images of the APTOS2019 data set with labels 0 and 1 correctly classified, and with label 0 incorrectly classified by all models, respectively. The same structure is followed for examples of the ISIC2020 (Tables 4.8, 4.9, 4.10 and 4.11) and MIMIC-CXR (Tables 4.12, 4.13, 4.14 and 4.15)

datasets, with the addition of examples from the label 1 incorrectly classified by all models. We used the Matplotlib [363] library for Python to generate these saliency maps visualisations, using the “bwr” colour map, which represents negative values in blue, zero values in white and positive values in red.

Table 4.5: Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the APTOS2019 data set with the label 0 (normal) correctly classified as 0 (normal) by all models.

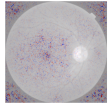
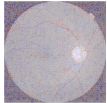
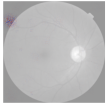
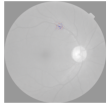
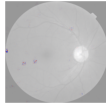
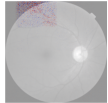
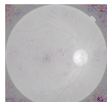
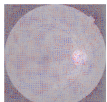
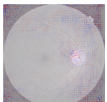
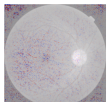
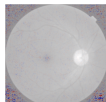
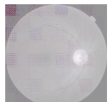
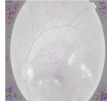
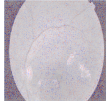
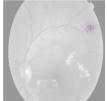
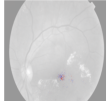
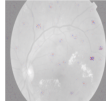
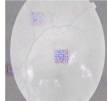
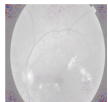
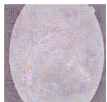
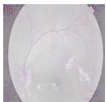
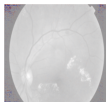
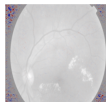
		DENSENET			RESNET		
Original Image		No Attention	+ SE Layer	+ CBAM Layer	No Attention	+ SE Layer	+ CBAM Layer
	LRP						
DeiT (LRP)	DeepLIFT	No Attention	+ SE Layer	+ CBAM Layer	No Attention	+ SE Layer	+ CBAM Layer
							

Table 4.6: Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the APTOS2019 data set with the label 1 (diabetic retinopathy) correctly classified as 1 (diabetic retinopathy) by all models.

		DENSENET			RESNET		
Original Image		No Attention	+ SE Layer	+ CBAM Layer	No Attention	+ SE Layer	+ CBAM Layer
	LRP						
DeiT (LRP)	DeepLIFT	No Attention	+ SE Layer	+ CBAM Layer	No Attention	+ SE Layer	+ CBAM Layer
							

4.4.5 Discussion, Conclusions and Future Work

4.4.5.1 Predictive Performance

All the experiments related to the predictive performance of deep learning models on the different data sets suggest that it is not clear that one should expect improvements in accuracy when using attention mechanisms. Several attention-based architectures tend to lower their predictive performances against their baseline architectures. Besides, the same baseline architectures seem to achieve comparable results even when directly comparing a transformer-based architecture (i.e., DeiT). Given that the intuition behind attention mechanisms is that these end up learning the most

Table 4.7: Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the APTOS2019 data set with the label 0 (normal) incorrectly classified as 1 (diabetic retinopathy) by all models.

		DENSENET			RESNET		
Original Image		No Attention	+ SE Layer	+ CBAM Layer	No Attention	+ SE Layer	+ CBAM Layer
	LRP						

Table 4.8: Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the ISIC2020 data set with the label 0 (benign) correctly classified as 0 (benign) by all models.

		DENSENET			RESNET		
Original Image		No Attention	+ SE Layer	+ CBAM Layer	No Attention	+ SE Layer	+ CBAM Layer
	LRP						

Table 4.9: Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the ISIC2020 data set with the label 1 (malign) correctly classified as 1 (malign) by all models.

		DENSENET			RESNET		
Original Image		No Attention	+ SE Layer	+ CBAM Layer	No Attention	+ SE Layer	+ CBAM Layer
	LRP						

Table 4.10: Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the ISIC2020 data set with the label 0 (benign) incorrectly classified as 1 (malign) by all models.

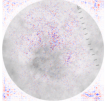
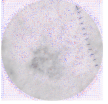
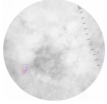
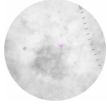
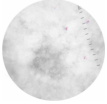
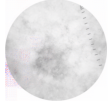
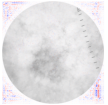
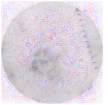
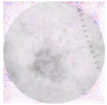
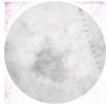
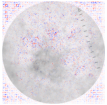
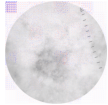
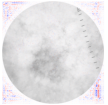
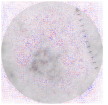
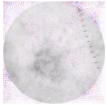
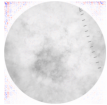
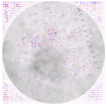
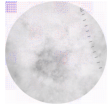
		DENSENET			RESNET		
Original Image		No Attention	+ SE Layer	+ CBAM Layer	No Attention	+ SE Layer	+ CBAM Layer
	LRP						
DeiT (LRP)							
							

Table 4.11: Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the ISIC2020 data set with the label 1 (malign) incorrectly classified as 0 (benign) by all models.

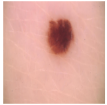
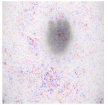
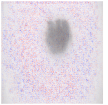
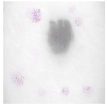
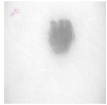
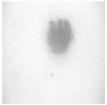
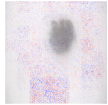
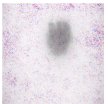
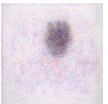
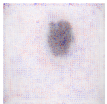
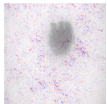
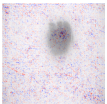
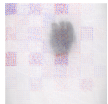
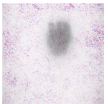
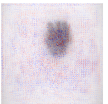
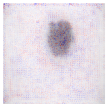
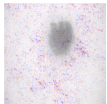
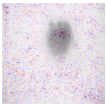
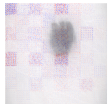
		DENSENET			RESNET		
Original Image		No Attention	+ SE Layer	+ CBAM Layer	No Attention	+ SE Layer	+ CBAM Layer
	LRP						
DeiT (LRP)							
							

Table 4.12: Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the MIMIC-CXR data set with the label 0 (normal) correctly classified as 0 (normal) by all models.

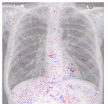
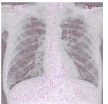
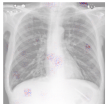
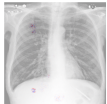
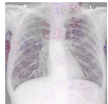
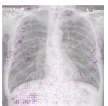
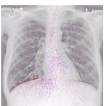
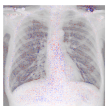
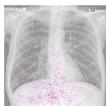
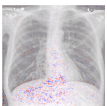
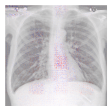
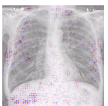
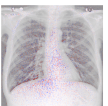
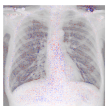
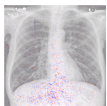
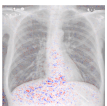
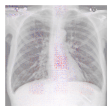
		DENSENET			RESNET		
Original Image		No Attention	+ SE Layer	+ CBAM Layer	No Attention	+ SE Layer	+ CBAM Layer
	LRP						
DeiT (LRP)							
							

Table 4.13: Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the MIMIC-CXR data set with the label 1 (pleural effusion) correctly classified as 1 (pleural effusion) by all models.

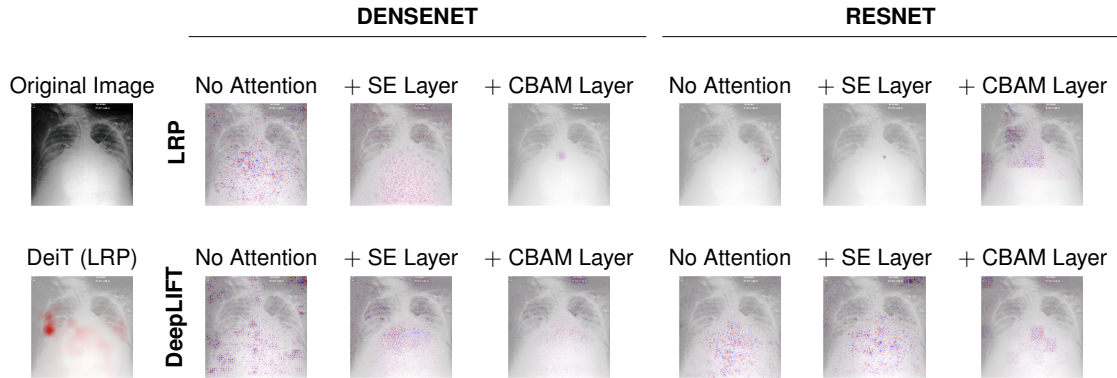


Table 4.14: Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the MIMIC-CXR data set with the label 0 (normal) incorrectly classified as 1 (pleural effusion) by all models.

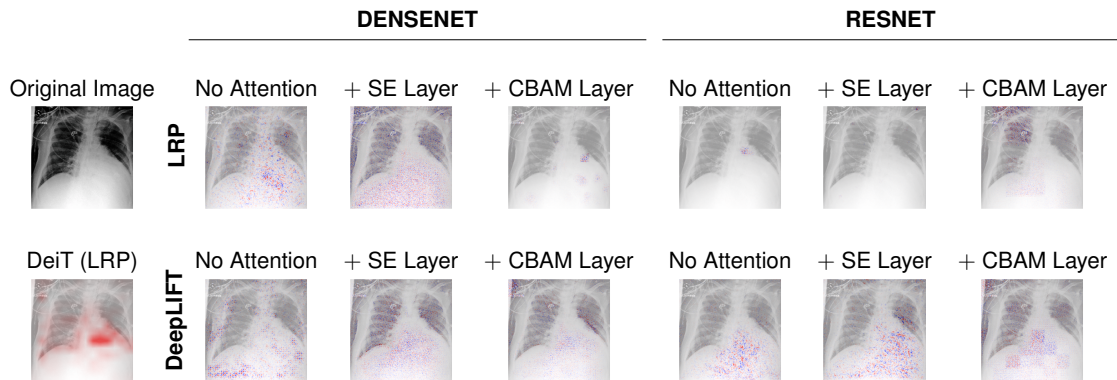
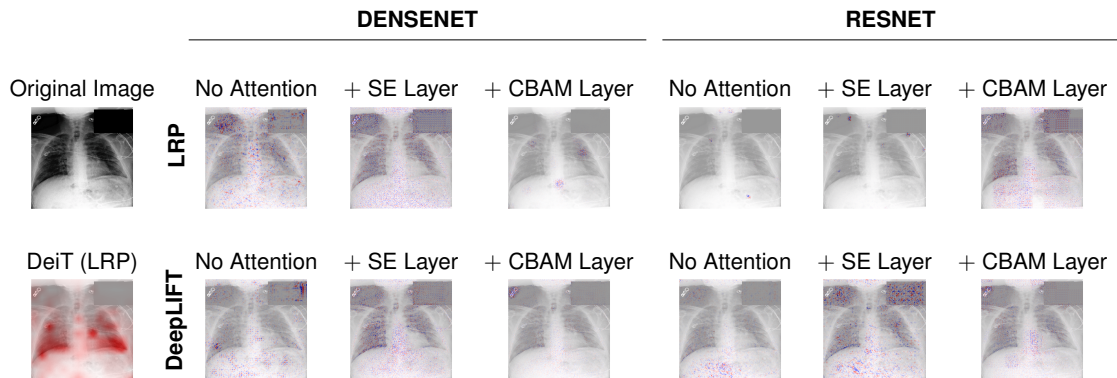


Table 4.15: Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the MIMIC-CXR data set with the label 1 (pleural effusion) incorrectly classified as 0 (benign) by all models.



relevant features, one might expect that attention-based architectures would perform better when trained in low-data regimes. However, results obtained in all data sets suggest that this might not be the case. In fact, we can observe that, apart from the transformer-based architecture, the baseline backbone models often perform better. Hence, it is reasonable to conclude that the use of attention mechanisms will not *always* benefit the training of deep learning algorithms, an argument that contrasts with a trending narrative in most papers. On the other hand, the results reported in the literature often relate to marginal or residual improvements in the state-of-the-art backbone networks. Given that the training of deep learning algorithms is generally a stochastic process, there is a need to assess these reported improvements with a more critical view and with robust statistical tests.

4.4.5.2 Model Complexity

The integration of attention mechanisms increases the number of parameters of the deep learning models, thus increasing their complexity. This information allows us to conclude that, at least for computer vision applications, it is not necessarily true that the use of attention mechanisms contributes to the decrease of model complexity. While this may be true in several natural language processing applications, it is not correct to generalise such a rule to other types of applications. On the other hand, since these attention mechanisms often rely on simple operations (e.g., matrix multiplications), such as convolutions, we acknowledge that their use may reduce the training time of deep learning algorithms (similarly to what happened when the community started using CNNs). This aspect, though, was not objectively assessed in our experiments. However, since they are particularly efficient at modelling long-range dependencies (e.g., transformer-based architectures), we expect these mechanisms to keep increasing their popularity and, hence, become widely used in medical image applications. Moreover, Cordonnier et al. [364] reported that these mechanisms may operate similarly to CNNs. Another question arises from these results: are these attention-based algorithms allegedly performing well because of the inner-functioning of their attention mechanisms themselves or just because we are increasing the number of model parameters? While one may report that this issue is nonsense, we point out that some transformer-based architectures have a considerably high number of parameters [229].

4.4.5.3 Post-model Interpretability

Although there are already several works that criticise the subjectivity of *post-hoc* saliency maps [11, 365, 366, 367, 368, 369], it is still common to use these methods to justify improvements related to the degree of interpretability of models. In this sense, when analysing the results obtained with the DenseNet-121 and ResNet-50 models, we would expect that, with the increase in complexity of the attention mechanism, the distribution of the important pixels around the image would also be more focused (i.e., would have less variance). However, that does not seem to happen in our use cases. Interestingly, besides the inherent properties of the data and the task (i.e., images of the retina are different from skin images and chest X-ray images), the results drastically change when

we use a different backbone. In general, ResNet-based models attain less noisy DeepLIFT and LRP saliency maps, with some exceptions in the ISIC2020 data set (e.g., Table 4.11). On the other hand, the LRP saliency maps obtained for the DeiT seem to highlight somewhat clear regions of the images. Besides, it is also important to remind the reader that these frameworks allow us to generate explanations even for the cases where the model misclassifies. We also stress that one of the limitations of such analysis is that there does not exist an objective ground truth of what a high-quality visual explanation is. Therefore, it is not trivial, in computer vision tasks such as this, to conclude with complete confidence that, even for the cases where the model succeeds, it learned the right correlations. Hence, can we believe the narrative that attention mechanisms are learning the most relevant features of the image? Or is this type of analysis a result of luck in most cases? Truth is, our results still suggest that there is a high degree of subjectivity (i.e., the interpretation of these saliency maps deeply depends on the human that is interpreting them) and that there is no apparent correlation between the use of attention mechanisms and the visual aspect of the *attributions*. Besides, we stress that these visualisations shown in the literature depend on several parameters (e.g., type of colour map, overlay parameter). Moreover, the method we use to generate such visualisations may often change what we expect users to observe. This is also related to another open challenge in *post-hoc* explanation methods that is the sign of the attributions. What should we expect? Positive attributions for the positive class and negative attributions for the negative class, or the other way around? Should we always normalise the sign of the attributions so it is always positive? This motivates our statement on the need for more objective methods to assess the degree of interpretability of models.

4.5 Attention Mechanisms for Medical Image Retrieval

The results of this work have been disseminated in the form of an article in an international journal: 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*, vol. 12, no. 1, p. 20732, 2022 [49]. The code related to the implementation of this work is publicly available in a GitHub repository².

4.5.1 Description and Contributions

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 of the number of images generated per examination [370]. Consequently, this has increased the workload of radiologists, who 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

²<https://github.com/TiagoFilipeSousaGoncalves/ig-cbir-att-ext>

by the radiologists to the patients [371, 372]. Moreover, as the ratio of diagnostic demand to the number of radiologists increases, the diminished effective available time per diagnostic becomes a critical issue [373]. According to the current paradigm, in case of doubt for 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* [374]). After reviewing all possible differential diagnoses (those originally considered and those that came up during the search), the radiologist weighs them 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 their head. This task is time-consuming and often ineffective, requiring 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 low prevalence conditions. Hence, developing disease-targeted content-based image retrieval (CBIR) systems that automatically present disease-matching similar images to the one being analysed is extremely relevant. 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 [375]. In this work, we focus 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 difficult since the clinical analysis is typically constricted to a small image region, discarding the most 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 retrieving the most similar image regarding the disease and disease severity (equal to finding the semantics instead of the literal words). As illustrated in Figure 4.10, those can be quite far apart, as Figure 4.10 (b) is, overall, more similar to Figure 4.10 (a) than Figure 4.10 (c), while in terms of disease and disease severity, it is the opposite, with Figure 4.10 (b) being the least similar image and Figure 4.10 (c) the most similar (from a catalogue of 10 images).

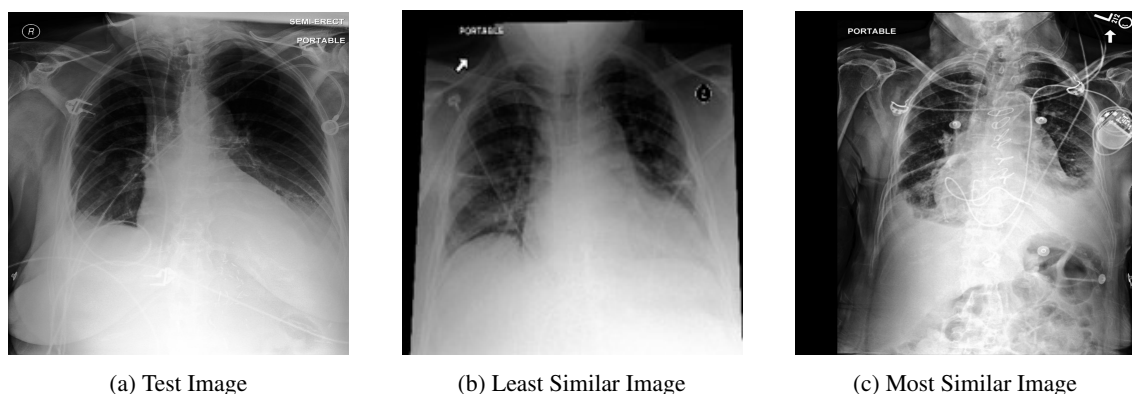


Figure 4.10: Pleural Effusion test image (a) and the least (b) and most (c) similar images of the catalogue according to our board-certified radiologist (in terms of disease and disease severity). The overall most similar image would be (b). However, such matching is not of radiological interest.

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 focus better on 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 the focus of the network in the regions that matter to the decision, and explore two different techniques: one driven by interpretability, and another based on attention mechanisms.

This study builds upon the work proposed in Silva et al. [373] and makes the following contributions:

1. Adding experiments with a second new dataset and a second medical condition (pneumonia);
2. Performing a comparison to a recently proposed network, employing implicit attention mechanisms;
3. Improving our comparisons to expert radiologists by adding two more board-certified radiologists to each study to assess the evaluated methods with respect to the inter-rater variability of these tasks.

4.5.2 Background on Medical Image Retrieval

In the previous Section, we motivated the need for an efficient medical image retrieval system to help clinicians in their daily routines. In this Section, we briefly review the literature on medical CBIR to contextualise the reader.

The main difficulties in the development of CBIR systems are related to 1) the development of algorithms that generate useful semantic representations of medical images to retrieve the most similar examples [376] effectively; and 2) the integration of these algorithms in end-user applications [377, 378].

In this work, we focus on the first difficulty and aim to give the reader an overview of the works presented in the literature to find the most suitable representation to perform the retrieval: Tizhoosh [379] explored the use of bar code annotations as an auxiliary method for feature-based image retrieval; Srinivas et al. [380] 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 [381] 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 [382] 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. [383] created a method that consists of a weighted graph whose nodes represent the images

and edges measure their pairwise similarities; Nowaková et al. [384] presented a novel method for fuzzy medical image retrieval using vector quantisation with fuzzy signatures in conjunction with fuzzy S-trees; Qayyum et al. [385], Ayyachamy et al. [386] and Owais et al. [387] trained CNNs on multi-modal and multi-class datasets, and used the learned features and the classification results to retrieve medical images; Cai et al. [388] 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. [389] used a CNN-based auto-encoder method in the feature extraction process to improve the results of the retrieval process; Mbilinyi et al. [390] 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, one can perform feature representation using one of the following strategies: 1) statistical measures; 2) hand-crafted features; 3) learned features; or 4) a combination of the previously mentioned strategies.

However, to 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 utilise AI interpretability methods to guide the retrieval process, focusing on the disease without necessitating any additional related label information.

4.5.3 Materials and Methods

4.5.3.1 Data

In these experiments, we used the MIMIC-CXR dataset (see Chapter 3). From all the available labels, we focused our attention on the following conditions: pleural effusion and (potential) pneumonia. Even though the provided label is pneumonia, our main board-certified radiologist considers that a chest X-ray does not allow a conclusive diagnosis of pneumonia; thus, we perform our analysis by referring to the cases as having potential pneumonia. To train and evaluate our models, the original splits selected by the data providers [149] were used, i.e., we considered the training fold for training, the validation fold to select the final model, and the test set to evaluate the performance. From these splits, we only considered frontal view images, either acquired in an anterior-posterior (AP) or posterior-anterior (PA) setting. For completeness, one of the analyses was done with AP view images (pleural effusion) and the other with PA view images (potential pneumonia). For each condition, there are four possible annotations: 1, the label was positively mentioned in the associated study; 0, the label was negatively mentioned in the associated study; -1, the label was either mentioned with uncertainty or with ambiguous language in the report; and “missing”, no mention of the label was made in the report. Regarding our experiments, we only considered images associated with the labels 1 and 0, thus not introducing uncertainty in the training of the models. Given that selection, we ended up with 61203 training images, 534 validation images, and 1072 test images for pleural effusion; and 18226 training images, 133 validation images, and 258 test images for pneumonia.

To evaluate the performance of the different types of methods in the retrieval task, we split the test data into query and catalogue images. For pleural effusion, we considered ten query images, each associated with ten catalogue images. For pneumonia, the test set was considerably smaller, and we only considered five query images, each associated with ten catalogue images. For both conditions, query and catalogue images were randomly chosen. All images considered for the evaluation were labelled by our main board-certified radiologist. Besides labelling all images, our main radiologist also ranked the associated catalogue images in relation to each of the test images, considering the similarity in terms of disease severity (Figure 4.11 shows an example of the ranking annotations provided for pleural effusion and Figure 4.12 shows an example of the ranking annotations provided for pneumonia). In addition to our main board-certified radiologist, four other experienced radiologists also analysed and ranked the cases to assess the inter-rater variability in such a task.

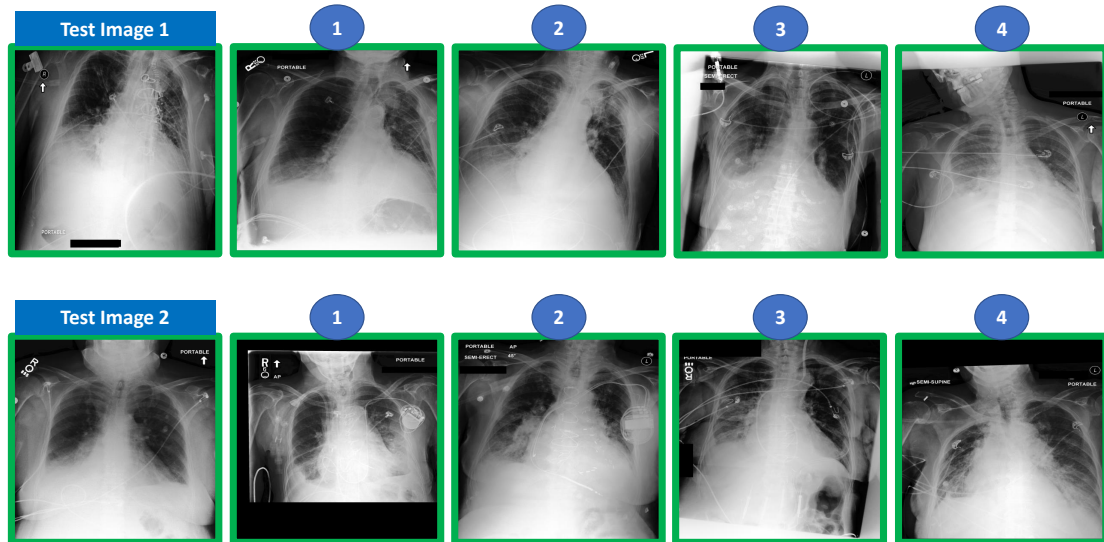


Figure 4.11: Example of test cases and ranking annotation for pleural effusion condition (the top 4 is shown) performed by the radiologist. The numbers on top of the image represent the ranking position of the image in the catalogue (when compared to the query/test image). The green box means pleural effusion case (according to the dataset label).

4.5.3.2 Methods

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) [391]. As in Silva et al. [373], 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.

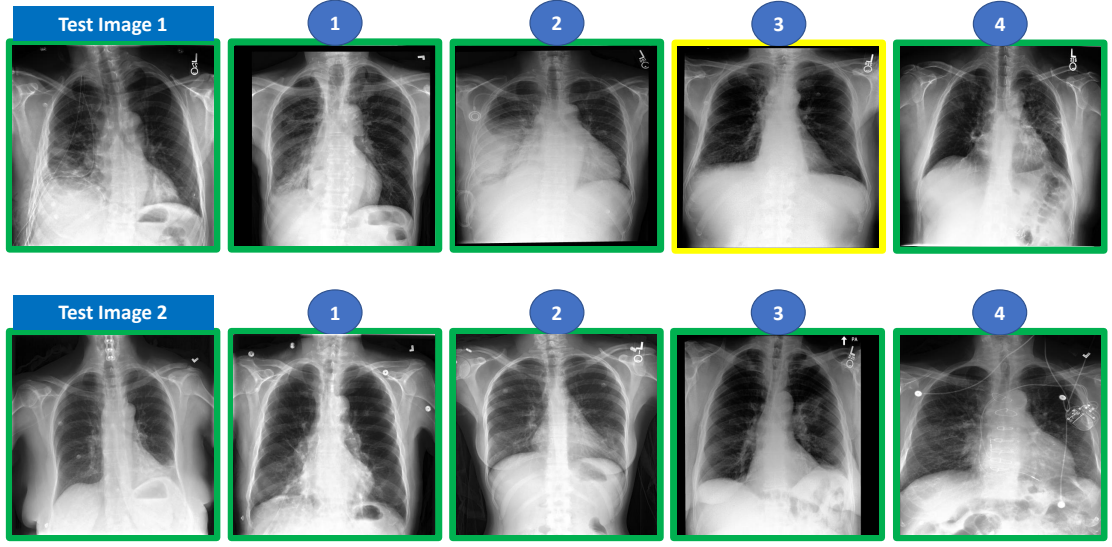


Figure 4.12: Example of test cases and ranking annotation for potential pneumonia condition (the top 4 are shown) performed by the radiologist. The numbers on top of the image represent the ranking position of the image in the catalogue (when compared to the query/test image). The green box means pneumonia case (according to the dataset label), the yellow box means radiologist considers the case as potential pneumonia, and the dataset label is no pneumonia.

Convolutional Neural Network (CNN) The second method to be considered is a deep learning based method, where the relevant features are automatically learned [375, 381, 392, 393]. As in Silva et al. [373], we use the DenseNet-121 [359] as our CNN architecture. However, in this work, we do not initialise its weights with the ImageNet pre-training, but instead with the pleural effusion CheXpert CNN model from Silva et al. [373] for the pleural effusion condition, and with the pleural effusion model from this work for the pneumonia condition, as pre-training using data more similar to the final domain is more effective than using ImageNet pre-training [394, 395]. 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 formalised in Equation 4.1, 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).

$$d_{\text{CNN}}(I_t, I_c) = \|F(\theta_{\text{CNN}}, I_t) - F(\theta_{\text{CNN}}, I_c)\|_2 \quad (4.1)$$

Interpretability-guided Network (IG) The third method being considered is the method proposed in Silva et al. [373]. It uses the exact same architecture as the CNN model but has as input

the saliency maps instead of the original images (Figure 4.13) to focus the network on the disease-related characteristics. Those saliency maps are computed using the Deep Taylor interpretability method [34] and are based on the previously presented CNN model. This time, the deep learning network was initialised with the CheXpert IG model from Silva et al. [373] for the pleural effusion condition and with the IG pleural effusion model from this work for the pneumonia condition. As with the CNN approach, the similarity is computed based on the layer previous to the last layer of the model. The distance between two images is formalised in Equation 4.2, 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.

$$d_{IG}(I_t, I_c) = \|F(\theta_{\text{IG}}, S(\theta_{\text{CNN}}, I_t)) - F(\theta_{\text{IG}}, S(\theta_{\text{CNN}}, I_c))\|_2 \quad (4.2)$$

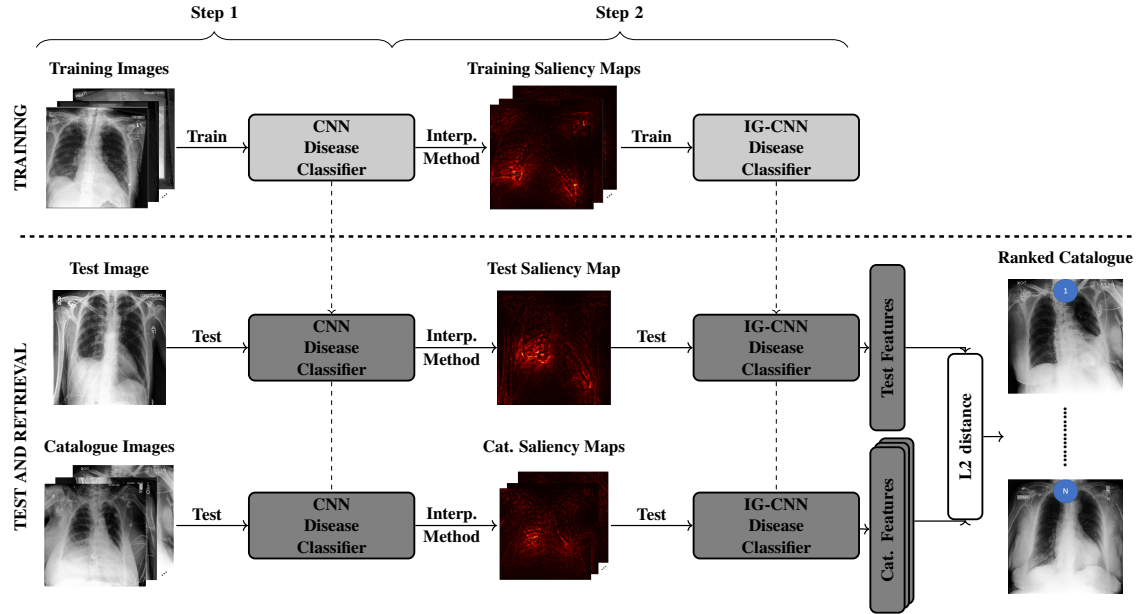


Figure 4.13: Overview of the proposed Interpretability-guided approach. Blocks in light gray (■) mean neural networks are being trained (i.e., weights are being updated), whereas blocks in dark gray (■) represent trained neural networks (i.e., weights are fixed). In the saliency maps, brighter colours 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. The L2 distance is computed between the test image’s latent features and each catalogue image’s latent features.

Attention Network (ATT) Here, we add a fourth method to the comparison, driven by attention mechanisms. Recently, a CNN with a multi-level dual-attention mechanism (MLDAM) has been proposed for macular optical coherence tomography classification [396]. 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, Mishra et al. [396] state that this technique enables the network to utilise the information of coarser features, preventing the 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 [397] 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) [7] and concluded that the use of attention modules had enabled the network to shift the focus 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 [359]. Following the notation in Mishra et al. [396], 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 layer previous to the last layer. The distance between two images is formalised in Equation 4.3, 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.

$$d_{ATT}(I_t, I_c) = \|A(\theta_{ATT}, I_t) - A(\theta_{ATT}, I_c)\|_2 \quad (4.3)$$

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 (see Equation 4.4, where y is the binary indicator, $\ln$ the natural logarithm, p the predicted probability, and θ the model parameters).

$$\mathcal{L}(\theta) = -(y \ln(p(\theta)) + (1 - y) \ln(1 - p(\theta))) \quad (4.4)$$

For the pleural effusion condition, the deep learning models were trained for 10 epochs, with a batch size of 32, using the *Adadelta* optimiser [398]. 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 [399].

Regarding the pneumonia condition, the deep learning models were trained for 15 epochs, with a batch size of 32, and using the *Adam* optimiser [400] 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. Hyperparameter values were empirically optimised for the CNN models and replicated for all the others. Final models were selected based on the F1 score (see Equation 4.5) in the validation set.

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4.5)$$

We note that the training process was agnostic to the ranking task at hand. No information on the 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 [401] 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 [402] implementation of the Deep Taylor method, as in Silva et al. [373].

4.5.3.3 Evaluation and Comparison

The quality of the retrieval is evaluated by computing the normalised Discounted Cumulative Gain (nDCG) (see Equation 4.6), which is the normalised version of the Discounted Cumulative Gain (DCG) (see Equation 4.7), being it a common metric in learning to rank tasks [403]. 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 consider for the evaluation (e.g., when we evaluate the entire set of retrieved images, $p = 10$). In Equation 4.7, rel_i represents the relevance value assigned to the ranking position i , with the least similar image having relevance of 1 and the most similar image having relevance of 5.5 (i.e., the relevance of two contiguous positions differs by 0.5). Thus, the first positions of the catalogue ranking are more important than the last ones, with the importance gradually being reduced as we go from the first to the last ranked image.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (4.6)$$

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i + 1)} \quad (4.7)$$

To contextualise 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 can check the inter-rater variability in ranking tasks, which also helps us to have a more complete evaluation of the quality of our methods.

4.5.4 Results and Discussion

4.5.4.1 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 were used to select the final classification model, and the test images were used 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 to compare inter-rater variability with our models' performance.

In Figure 4.14 (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 align with the other radiologists, demonstrating the high quality of both methods. Furthermore, as was expected, the CNN approach leads to better results than the SSIM method. The same can be observed in Figure 4.14 (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 penalising 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, the IG and ATT approaches remain the best methods and are still in line with the performance of the two radiologists.

In Figure 4.15, we show the top-4 retrieved results obtained by each of the methods and provided by the radiologists compared to the ground-truth defined by our main radiologist for one split, corresponding to a specific test case and catalogue. All machine learning methods (i.e., CNN, IG, and ATT) attained extremely high nDCG results in this split. 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 those 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.

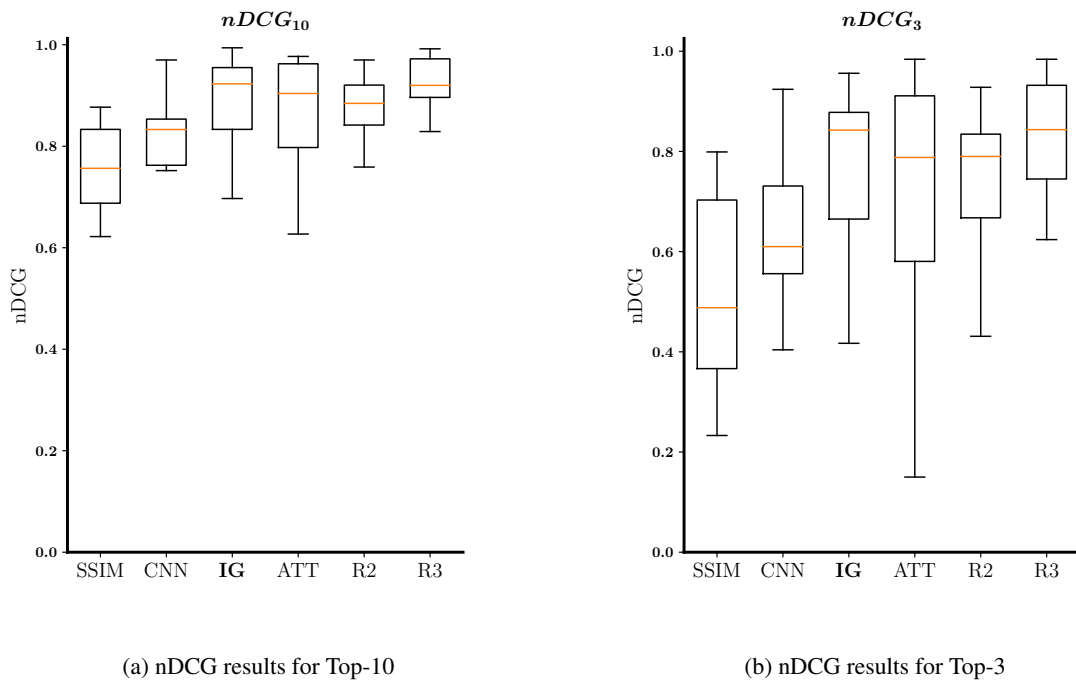


Figure 4.14: Box-and-whisker plots regarding the nDCG results for the pleural effusion top-10 (a) and top-3 (b) retrieved images. SSIM is the statistically-based baseline, CNN is the CNN-based baseline, IG is the proposed interpretability-guided approach, ATT is the attention method, R2 is the ranking provided by the second board-certified radiologist, and R3 is the ranking provided by the third board-certified radiologist.

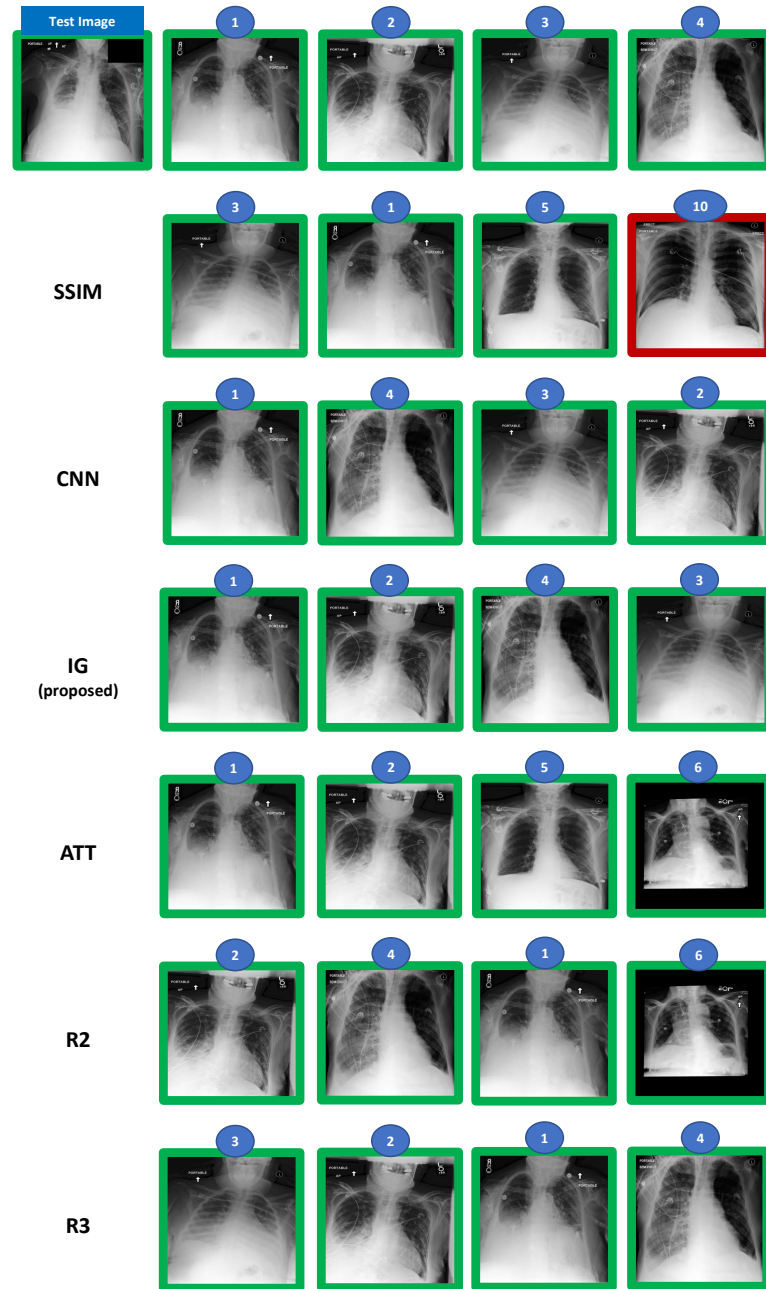


Figure 4.15: Example of test case and the top-4 retrieved images given by each of the radiologists (R1 = ground-truth, R2, and R3) and each of the machine learning methods. In this split, both the CNN, IG, and ATT obtained nDCGs (top-10) > 0.9 . The green box means pleural effusion case, and the red box means no pleural effusion (according to the dataset label).

In terms of classification performance in the entire test set (the 1072 test images), the interpretability-driven model was the one leading to the best results (F1-score=0.862), followed by the attention-based model (F1-score=0.809), and the standard CNN model (F1-score=0.790).

4.5.4.2 (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, we considered 18226 training images, 133 validation images, and 258 test images for the experiments. 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 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 only indicative of potential pneumonia and not of a definitive diagnosis. Thus, catalogue images were ranked, keeping in mind their potential as pneumonia cases. In Figure 4.16 (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 that were 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 Figure 4.16 (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 generally considerably worse. 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 MIMIC-CXR label, and the diagnosis provided by our main board-certified radiologist). In Figure 4.17, we present an example query case and the respective top-4 retrieved images obtained by each of the methods and provided by the radiologists compared to the ground-truth 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 the most similar image in terms of pneumonia to the test image in the first position. 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 potential pneumonia and the MIMIC-CXR label being

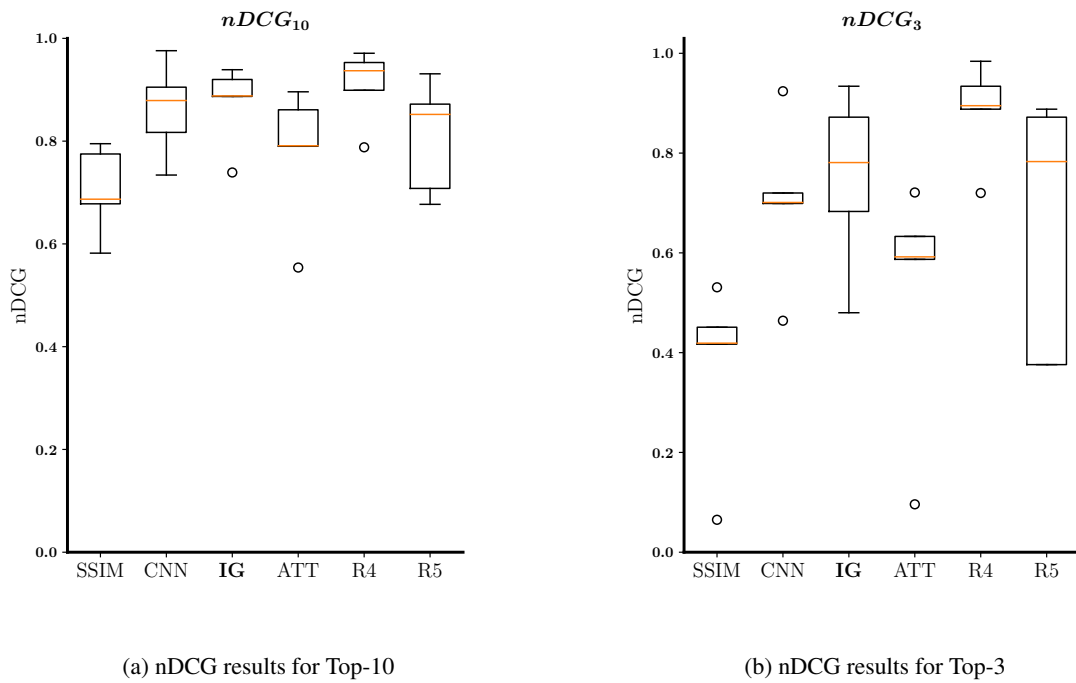


Figure 4.16: Box-and-whisker plots regarding the nDCG results for (potential) pneumonia top-10 (a) and top-3 (b) retrieved images. SSIM is the statistically-based baseline, CNN is the CNN-based baseline, IG is the proposed interpretability-guided approach, ATT is the attention method, R4 is the ranking provided by the fourth board-certified radiologist, and R5 is the ranking provided by the fifth board-certified radiologist.

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.

In terms of classification performance in the entire test set (the 258 test images), this time, the attention model was the one leading to the best results (F1-score=0.655), followed by the interpretability-driven model (F1-score=0.645), and the standard CNN model (F1-score=0.620).

4.5.4.3 Ablation study

We also studied the relevance of training with the saliency maps, not only using them to compute the features in the semantic space. In Figure 4.18, 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 - IG. By observing Figure 4.18, 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.

4.5.5 Conclusions and Future Work

We have investigated the use of different CBIR 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 can 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 textual descriptions, motivating 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 (i.e., the 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.

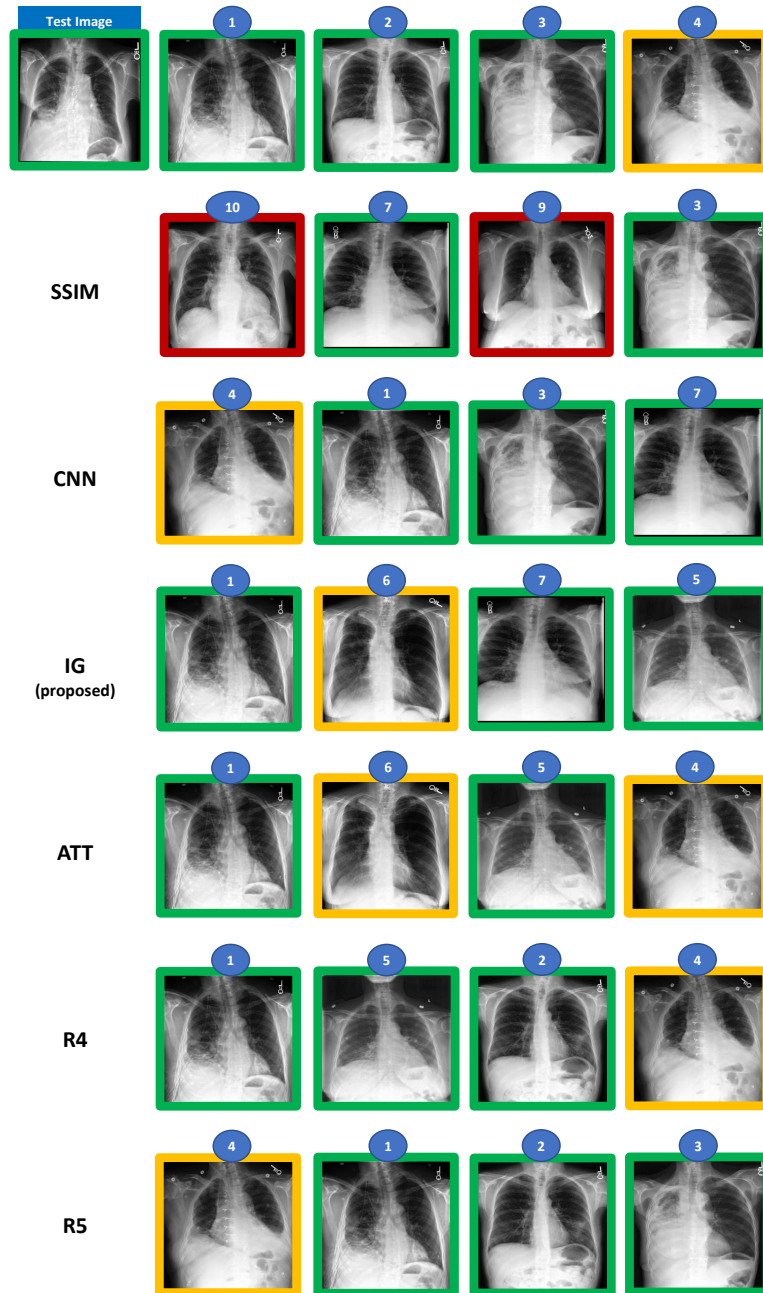


Figure 4.17: Example of test case and the top-4 retrieved images given by each of the radiologists (R1 = ground-truth, R4, and R5) and each of the machine learning methods. In this split, both the CNN, IG, and ATT obtained nDCGs (top-10) > 0.8. The green box represents a potential pneumonia case, the red box represents no potential pneumonia, and the orange box represents a disagreement between R1 and the label, with R1 considering the case as potential pneumonia.

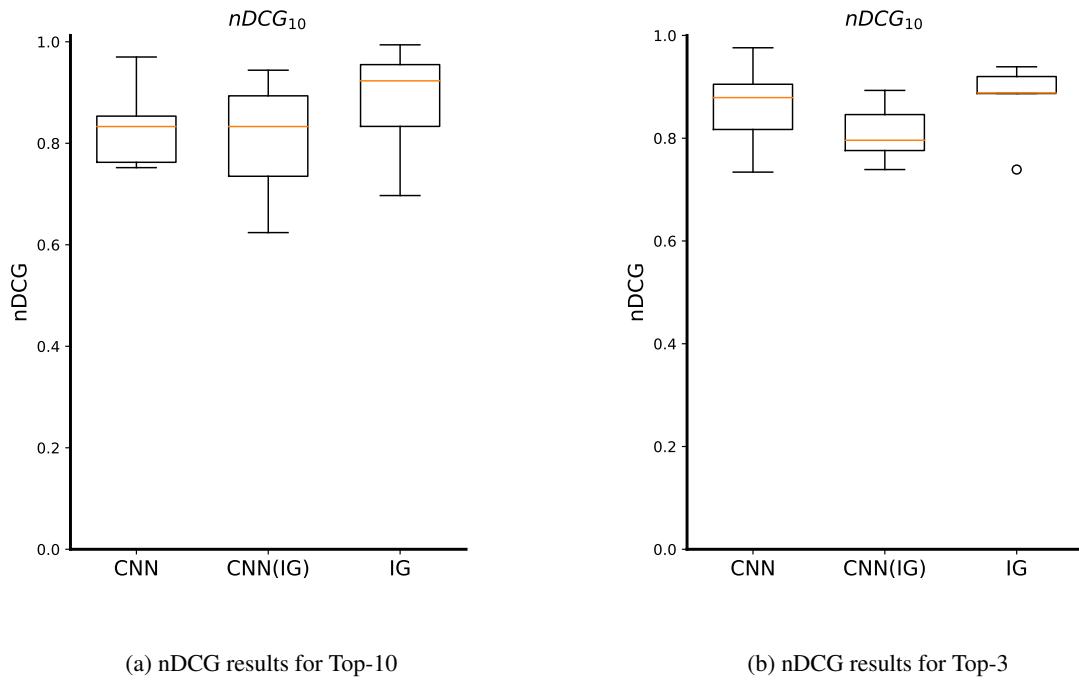


Figure 4.18: Box-and-whisker plots regarding the nDCG results for top-10 retrieved images. CNN is the CNN baseline model, CNN(IG) is the CNN model having as inputs the Deep Taylor saliency maps, and IG is the proposed interpretability-based approach (i.e., it was trained (fine-tuned) with saliency maps and has as inputs also saliency maps).

It is important to emphasise that all methods were only trained to solve binary classification tasks, not using any ranking information. This means the annotation effort 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. Test and catalogue images, apart from being ranked, were 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, our proposed method obtained an excellent ranking performance even for the pneumonia condition.

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 investigation is required to translate these algorithms into the clinics. Considering the evaluation aspect, it is crucial to have more extensive studies, datasets, and 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 multi-modal data, the introduction of causal knowledge [404], privacy-preserving image retrieval [405], and also exploring federated learning settings [406]. By the use of multi-modal 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 could 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 [407]. To prevent this from happening, integrating a causal structure is essential, or via effective interpretability-guided inductive biases as reported recently in Mahapatra et al. [408, 409]. For clinical applications where personal characteristics are exposed, primarily if these systems are used for educational purposes, where the images are shown to unauthorised personnel, it is extremely important to anonymise the retrieved cases before showing them. Even though Montenegro et al. [405, 410] already explored the use of privacy-preserving methods to anonymise medical images, further research is required to improve the 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.

4.6 Attention Mechanisms for Computational Pathology

The results of this work have been disseminated in the form of a paper at an international conference: T. Gonçalves, D. Pulido-Arias, J. Willett, K. V. Hoebel, M. Cleveland, S. R. Ahmed, E. Gerstner, J. Kalpathy-Cramer, J. S. Cardoso, C. P. Bridge, and A. E. Kim, "Deep Learning-based Prediction

of Breast Cancer Tumor and Immune Phenotypes from Histopathology,” 2024 [48]. The code related to the implementation of this work is publicly available in a GitHub repository³.

4.6.1 Description and Contributions

A cardinal facet of precision oncology uses complementary biological facets from a patient’s tumour to optimise clinical decision-making. While genomic sequencing of tumour tissue has transformed the therapeutic landscape of modern oncology, it has become increasingly apparent that genomic biomarkers (DNA sequencing analysis) are insufficient as stand-alone tools, given the wide variability in clinical outcomes for patients with similar biomarker profiles. To improve this limitation, some have proposed incorporating transcriptional profiling (RNA sequencing analysis) into the precision oncology framework to define improved biomarkers of efficacy [411]. Recent studies have linked immune activity, angiogenesis, and tumour mitoses pathways in the tumour microenvironment (TME) to risk of metastasis and therapeutic efficacy [412], thus highlighting a need to monitor therapeutically relevant facets of the TME.

To date, there is no commercially available method in clinical practice to evaluate a patient’s TME in clinical oncology. Efforts to develop a widely available gene expression panel quantifying facets of the TME have been limited due to the need for RNA extraction from fresh tissue instead of the more widely available paraffin-embedded tissue. Furthermore, pathology workflows mainly evaluate tumour cell characteristics (e.g., grading, sub-typing) within H&E and cannot reproducibly assess the TME. As spatial profiling efforts suggest that the physical co-localisation of non-malignant cells to tumour cells modulate intracellular communication, tumour phenotypes, and clinical outcomes [413], there is likely biologically relevant information from H&E that is not used in clinical care. Therefore, given the importance of the TME and the wide availability of H&E, we sought to build upon recent advances in computer vision to estimate tumour and immune phenotypes from H&E WSIs.

Here, we present an exploratory application of MIL algorithms to characterise tumour and immune phenotypes from the H&E slide of primary breast tumours. We hypothesised that deep learning (DL) could predict these facets of the TME by recognising biologically relevant features (e.g., tissue architecture, tumour-immune interface) within an individual H&E image. Using a DL classification pipeline, we employed different feature extraction techniques and model architectures to optimise the prediction of tumour and immune phenotypes (see Figure 4.19). To the best of our knowledge, our effort is one of the first to comprehensively monitor the TME from H&E using an end-to-end DL approach. This work can inspire collaboration for computational pathology biomarker development to facilitate the development of a widely available tool that quantifies facets of the TME for each patient.

³<https://github.com/QTIM-Lab/dl-prediction-brca-tip>

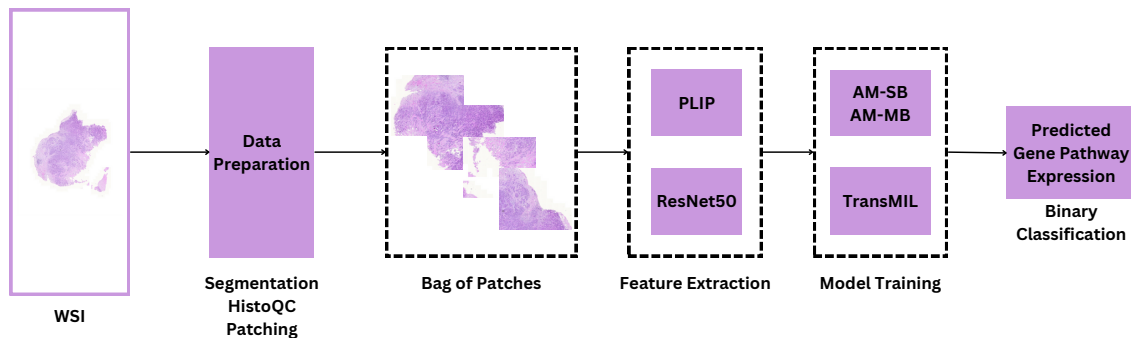


Figure 4.19: Overview of the proposed pipeline.

4.6.2 Background on Attention Mechanisms for Computational Pathology

With the increasing availability of large biobanks of patient-derived H&E samples, multi-Omics, and high-performance computing capabilities, there is great promise to combine these tools towards driving new insights in oncology [414, 415, 416, 417]. However, analysis of H&E WSIs is computationally difficult due to their multi-pyramidal gigapixel resolution. Down-sampling these images is undesirable, as this would result in a loss of biologically-relevant information that may harm a DL model’s performance. One popular strategy for WSI-based DL analysis is weakly-supervised MIL, as this strategy avoids the need for manual data annotations. After a WSI is cropped into patches, feature vector embeddings are extracted from each patch individually. The collection of patch-level feature vectors is then inputted into a MIL framework to yield the corresponding WSI-level representation and score [418].

One example of this strategy is the Clustering-constrained Attention Multiple instance learning approach (CLAM) [419]. This method uses attention-based pooling to aggregate patch-level representations, weighted by the attention score, into slide-level representations. This slide-level representation, instead of the entire WSI, is the input to the final classification module, so CLAM requires fewer computational resources to train and optimise. However, an important limitation of MIL is that each patch is assumed to be independent of surrounding patches [420]. This assumption may not be valid, given the importance of spatial localisation of different cell populations in dictating tumour phenotype. To this end, recently developed vision transformer models account for spatial correlation and context between patches. The Transformer based MIL (TransMIL) [278] and the Hierarchical Image Pyramid Transformer (HIPT) [421, 422] leverage morphological and spatial information to learn representations that encode the natural hierarchical structure inherent in WSIs. While there is early evidence that these architectures may boost performance for tasks in computational pathology, it is not immediately clear whether this boost in performance may justify the increased computational resources that vision transformers require. Therefore, we evaluated the impact of the architecture on model performance in predicting tumour and immune phenotypes.

Another timely question for H&E-based model optimisation is the best strategy for feature extraction from the WSI. While non-medical DL models pre-trained on ImageNet have achieved tremendous performance, it is unclear whether this strategy is optimal for H&E-based tasks. A

prevailing hypothesis is that models pre-trained on H&E data would have improved performance on H&E-based tasks, compared to a model pre-trained on real-world images [423, 424]. To this end, recent efforts have combined large language models and image analysis pipelines with self-supervised learning to create H&E-based foundation models. The MI-Zero framework [423] and pathology language–image pre-training (PLIP) foundation model [424] use visual-language encoders to extract image and text features, compute the similarity scores between text and image embeddings, and employs an aggregation function to generate a slide-level prediction. Here, we evaluated the optimal method for feature extraction by applying both an ImageNet- and an H&E-based feature extractor to predict TME phenotypes.

4.6.3 Methodology

4.6.3.1 Data

WSI Data Preparation In these experiments, we used the TCGA-BRCA dataset (see Chapter 3). For quality assessment, we used the default parameters of HistoQC [425, 426]. We obtained a curated segmentation mask (i.e., tissue *versus* background) for each WSI. Next, using this mask, we split each WSI ($20\times$, $40\times$ magnification) into patches of size 256×256 using the pipeline described in [419]. Finally, we extracted feature vectors from each patch using two different approaches: a ResNet50 [263, 419] backbone pre-trained on the ImageNet dataset [101] and the PLIP [424] image encoder pre-trained on the OpenPath dataset [424].

Quantification of Gene Pathway Expression We obtained the normalised expression of individual genes (bulk RNA-sequencing data) for the TCGA-BRCA cohort from the cBioPortal for Cancer Genomics [427]. Each biological pathway was defined as per Gene Set Enrichment Analysis (GSEA) protocols. We then employed single-sample GSEA (ssGSEA) to calculate gene set enrichment scores for each sample and pathway pairing. ssGSEA transforms a single sample's gene expression profile to a gene set enrichment profile. This enrichment score represents the activity level of the biological pathway in which the gene set's members are coordinately up (i.e., positive values) or down-regulated (i.e., negative values) [428, 429, 430].

4.6.3.2 Models and Training Details

We approached the problem as a binary classification task. To this end, the ssGSEA scores underwent binarisation, with negative scores receiving the label 0 and the rest label 1. A random data split allocated 70% to training, 15% to validation, and 15% to testing, ensuring exclusive assignment of a given patient's data to one split. We trained the models using CLAM [419] and TransMIL [278].

CLAM There are two variations of this architecture: single-branch CLAM (CLAM-SB), which employs a fully connected layer for slide-level classification, and multi-branch CLAM (CLAM-MB), which uses a fully connected layer per class, using the class with the maximum score for slide-level classification. CLAM also promotes instance-level clustering under a mutual exclusivity assumption (e.g., cancer subtypes are mutually exclusive: when one subtype is present in the WSI, we may assume that there is no morphology of the other subtypes concurrently present). As this assumption does not apply to our case, we only used the attention mechanism (along with the classifier), giving us two model architectures: AM-SB and AM-MB. We trained these models for 300 epochs, using cross-entropy as the loss function and the adaptive moment estimation (Adam) optimiser [400] with a learning rate of 2×10^{-4} and weight decay of 1×10^{-5} . We saved the best model according to the loss on the validation set.

TransMIL We trained TransMIL for 200 epochs, using cross-entropy as the loss function and the Adam optimiser with a learning rate of 2×10^{-4} and weight decay of 1×10^{-5} . We saved the best model according to the loss on the validation set.

4.6.4 Results and Discussion

Consistent with our hypothesis, our models predict TME phenotypes from H&E WSIs to reasonable accuracy. Table 4.16 presents the area under the receiver operating characteristic curve (AUROC) results on the test for both model architectures, using either ResNet50-derived or PLIP-derived features. First, we observe that models predicting immune phenotypes (e.g., B-cell proliferation, T-cell mediated cytotoxicity, immunosuppression, antigen presentation) achieve satisfactory performance (AUROC 0.75-0.8). Given prior work linking the spatial distribution of tumour-infiltrating lymphocytes and myeloid cells to immune activity [413], our findings illustrate DL's ability to identify and learn spatial patterns of cell populations predictive of specific facets of immune function. These findings are noteworthy, as the type (e.g., T-, B-, and dendritic cell) or function (e.g., cytotoxic or exhausted) of lymphocytes is indistinguishable from manual pathology review on H&E. Future work is needed to identify specific spatial patterns on H&E indicative of immune activity or exhaustion.

Next, we observed that our models predicting facets of tumour aggressiveness (e.g., cell cycling, angiogenesis, and epithelial-mesenchymal transition (EMT)) achieve reasonable performance and can be estimated from H&E. These findings are biologically plausible, as cell cycling measures mitotic activity and aggressiveness of tumour cells, angiogenesis represents the growth of new blood vessels (a hallmark of cancer growth), and EMT measures a tumour's ability to invade surrounding tissue. Biological activity associated with these pathways can be observed on H&E but is often difficult for pathologists to quantify. Conversely, we observe fatty acid metabolism and oxidative phosphorylation achieving relatively poor performance. These tasks are metabolic pathways that are often not apparent in cell morphology and tissue architecture on H&E and likely require more specific stains to measure.

From a technical perspective, we observe that AM-SB and AM-MB architectures marginally outperform TransMIL-based prediction of all gene expression tasks, suggesting that the type of architecture likely does not play a crucial role in model performance. Notwithstanding, models trained with PLIP-based features generally outperform those trained with ResNet50-based features (see Table 4.17), suggesting that using a feature extractor pre-trained on H&E-related data allows the model to better learn essential features or relationships specific to histopathology. Importantly, PLIP-based features (512-dimensional feature vectors) also possess a computational advantage since they are half the size of ResNet50 features (1024-dimensional feature vectors), thus being more manageable to store and process.

Finally, we sought to identify explainable morphological descriptors that may govern tumour and immune function from our trained models. We generated an attention map on a H&E WSI that was predicted to have a high level of T-cell mediated cytotoxicity (see Figure 4.20). We note that a high-attention patch corresponded with a high degree of infiltrating lymphocytes, and a low-attention patch highlighted areas of necrotic tumours. These attention maps provide additional evidence, albeit preliminary, that our trained models are learning therapeutically-relevant patterns. Further work is needed to evaluate and optimise this approach rigorously.

Table 4.16: **AUROC results on the test set, obtained for all model architectures, using ResNet50- and PLIP-derived features.** The best results per type of feature-extraction strategy are highlighted in bold. While the AM-SB and AM-MB model architectures outperform TransMIL prediction of all gene expression tasks, the increase in performance was marginal, suggesting that model architecture may not play a critical role in the classification of gene expression.

Task	Class Proportion (0 / 1)	Architecture & Features		
		AM-SB	AM-MB	TransMIL
		ResNet50 / PLIP	ResNet50 / PLIP	ResNet50 / PLIP
B-cell proliferation	60.45% / 39.55%	0.6974 / 0.7755	0.7322 / 0.7735	0.6960 / 0.7673
T-cell mediated cytotoxicity	53.72% / 46.28%	0.7149 / 0.7703	0.7564 / 0.7770	0.7223 / 0.7196
Angiogenesis	47.92% / 52.08%	0.7042 / 0.7435	0.7053 / 0.7213	0.7019 / 0.7214
Epithelial-mesenchymal transition	44.08% / 55.92%	0.8110 / 0.7848	0.8023 / 0.8082	0.7545 / 0.7934
Fatty acid metabolism	52.83% / 47.17%	0.6323 / 0.6030	0.6294 / 0.5920	0.5258 / 0.5634
Glycolysis	49.31% / 50.69%	0.7996 / 0.8118	0.7954 / 0.8330	0.7834 / 0.8045
Oxidative phosphorylation	49.50% / 50.50%	0.6894 / 0.7145	0.6926 / 0.7332	0.6699 / 0.6826
Immunosuppression	59.51% / 40.49%	0.7996 / 0.8458	0.8133 / 0.8542	0.7572 / 0.8113
Antigen processing and presentation	49.50% / 50.50%	0.7450 / 0.7806	0.7599 / 0.7924	0.7503 / 0.7342
Cell cycling	51.26% / 48.74%	0.7768 / 0.7939	0.7809 / 0.7852	0.7121 / 0.7229

4.6.5 Conclusions and Future Work

Our study represents one of the first efforts to predict therapeutically relevant facets of the TME from an H&E WSI. Here, our work suggests that tumour and immune phenotypes can be inferred with reasonable accuracy using state-of-the-art DL methods. Importantly, we note that the choice of feature extraction strategy may impact performance. A model pre-trained on H&E-based features had improved performance for most tasks compared to a model pre-trained on ImageNet-based

Table 4.17: **AUROC results on the test set, obtained for the best model of each feature set.** The best-performing feature-extraction strategy is highlighted in bold. Models trained with PLIP-derived features generally perform better than ResNet-derived features, suggesting that using a feature extractor pre-trained on H&E-related data may improve performance.

Task	Features	
	ResNet50	PLIP
B-cell proliferation	0.7322	0.7755
T-cell mediated cytotoxicity	0.7564	0.7770
Angiogenesis	0.7053	0.7435
Epithelial-mesenchymal transition	0.8110	0.8082
Fatty acid metabolism	0.6323	0.6030
Glycolysis	0.7996	0.8330
Oxidative phosphorylation	0.6926	0.7332
Immunosuppression	0.8133	0.8542
Antigen processing and presentation	0.7599	0.7924
Cell cycling	0.7809	0.7939

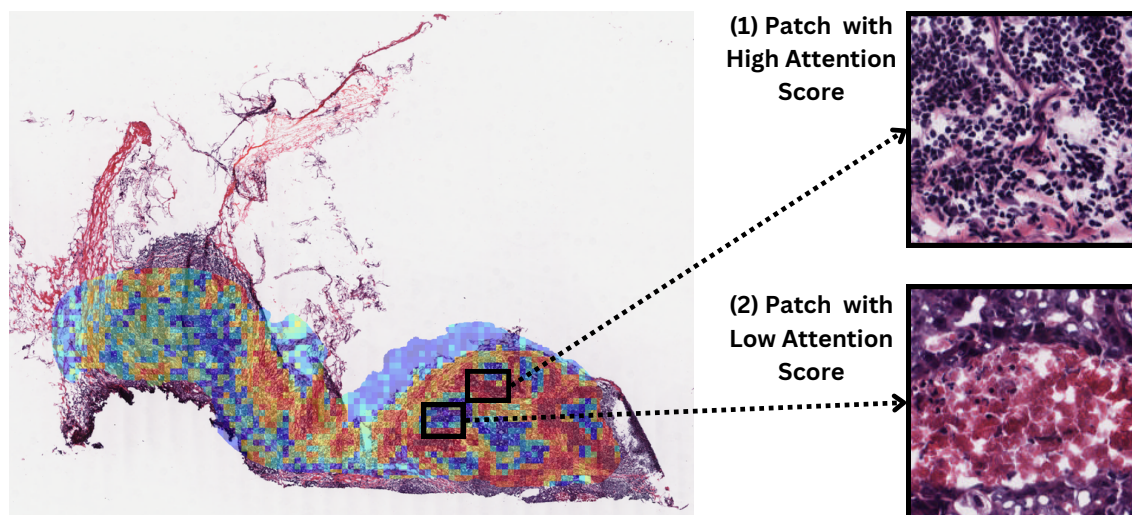


Figure 4.20: **Attention map obtained using the AM-SB architecture for a WSI predicted to have a high-degree of T-cell mediated cytotoxicity.** Red zones and blue zones indicate high and low attention scores, respectively. On the right, we provide two exemplar patches: 1) the high-attention patch illustrates abundant tumour-infiltrating lymphocytes without tumour cells, which are suggestive of high immune activity, and 2) the low-attention region demonstrates areas of tumour necrosis and minimal lymphocytes, consistent with low immune activity.

features. While additional work is needed to build upon and validate this work, we believe our results may be of interest. Firstly, for several other medical tasks, feature extraction strategies more native to the modality of interest do not always outperform natural image-based feature extraction [431, 432, 433, 434]. Secondly, as prediction of tumour and immune phenotypes can be inferred from H&E, these computational H&E biomarkers would have major implications for cancer patients and precision oncology as many hospitals only possess the capability for H&E and lack the resources for DNA/RNA sequencing.

In future work, we will experiment with other feature extraction frameworks (e.g., HIPT, MI-Zero) and feature fusion approaches (e.g., combining two different feature sets in a single model). Additionally, we intend to explore ordinal and multi-task classification and measures to boost model interpretability. Finally, we will trial multi-modal architectures, integrating relevant clinical metadata during training to assess the impact on performance.

4.7 Summary and Conclusions

The development of algorithms that learn what the bits of information that are relevant for the optimisation of a given task are is important to machine learning interpretability given that, intuitively, they may represent a means to achieve sparser and more robust models. These two properties also foster model transparency since using less information on the inputs or fewer parameters to output a prediction (i.e., sparsity) makes it easier for an external stakeholder (e.g., governmental institution) to debug and audit the model, and having a model that works consistently well on new examples (i.e., robustness) will increase users' confidence or trust in the model outputs.

In this Chapter, we delve into the realm of attention-based machine learning for healthcare applications and perform a thorough literature review of the topic, complemented by a systematic comparative study using three medical use cases (i.e., breast lesion, chest pathology, and skin lesion) and different model architectures with and without attention. Our analysis allowed us to look at attention-based algorithms with criticism and conclude that these methods often increase model complexity (i.e., the number of parameters), will not always benefit training or performance, and the quality of their post-hoc explanations is highly subjective. Nevertheless, these algorithms can still be interesting for some applications. We showed that in medical image retrieval (with chest X-ray), the use of an attention-based mechanism guides the models into learning an embedding space that contains more information about the pathology, thus allowing to retrieve cases that are more similar to the query, achieving, in some cases, comparable results to an interpretability-guided method (which requires more steps in its pipeline). Furthermore, in computational pathology, attention mechanisms seem to play a crucial role in the feature extraction methods (the features that lead to better performance originate in a transformer-based model) and MIL aggregation operator (state-of-the-art methods use an attention-based operator to aggregate the classification scores of the multiple instances). Additionally, some results also suggest that the attention maps produced by these methods have clinical significance.

Despite their limitations, attention mechanisms have proved worthy in several applications (including healthcare). In this thesis, we focus on the image dimension, evaluating several properties of these methods from the perspective of interpretability and trying to demystify some of the claims present in several works in literature while providing the reader with specific applications or situations where these methods might contribute to the improvement of their models.

Chapter 5

Learning from Disparate Sources of Data: Explaining Multi-modal Algorithms

Foreword on Author Contributions

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

1. Mohammad Hossein Zolfagharnasab and Nuno Freitas and Tiago Goncalves and Eduard Bonci and Carlos Mavioso and Maria J. Cardoso and Helder P. Oliveira and Jaime S. Cardoso, “Predicting Aesthetic Outcomes in Breast Cancer Surgery: A Multimodal Retrieval Approach,” in *Proceedings of the Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care (in MICCAI 2024)*, 2024 [46]

We also include experimental work (see Sections 5.5 – 5.6) that was not published when we submitted this thesis. Nevertheless, we hope that these results or improved versions of these works will be published in the foreseeable future.

5.1 Introduction

In medicine, acquiring multi-modal data on the same target has become a growing field as more advanced techniques and devices have become available, particularly in the biomedical imaging field [435]. A multi-modal database usually has several scans of the same subject acquired with different methods, thus allowing the storage of different tissue characteristics [436]. Popular methods include computed tomography (CT), functional magnetic resonance imaging (fMRI), magnetic resonance imaging (MRI), positron emission tomography (PET), single-photon emission computed tomography (SPECT) and ultrasound [437]. Besides, the simultaneous acquisition of several modalities is already encouraged during clinical practice [435], and there is already the

possibility of integrating modalities into the hardware of a single scanner [436]. This trend follows the intuition that various imaging modalities encompass abundant information which is different and complementary to each other [438]. Similarly to ensemble learning [439], the challenge resides in how this information should be aggregated to produce meaningful outputs that could support the decisions of the practitioners [435].

However, there are several open challenges (e.g., process of data gathering, the strategy of data fusion, assessment of the degree of interpretability of the algorithms) that require further investigation [440, 441, 442, 443, 444]. Using only medical images to develop models and explanations is not a realistic approach concerning the fact that clinicians use several clinical data sources to make decisions. Undoubtedly, this created the need to work on multi-modal methodologies and contributed to the increasing trend in the development of multi-modal deep learning strategies for medical purposes.

In this Chapter, we present three different case studies where multi-modal strategies (potentially) benefit the predictive power of algorithms, indirectly contributing to their robustness and transparency. Beyond this Introduction, this Chapter is organised as follows: Section 5.2 presents a brief background on data fusion techniques, motivating the need for multi-modal algorithms, Section 5.3 provides a general literature review of multi-modal machine learning for medical applications to contextualise the reader into state-of-the-art (and open challenges) of the field, Sections 5.4–5.6 present the experimental work developed to address the research questions related to this topic, and Section 5.7 concludes the Chapter, highlighting the main contributions and anticipating future work directions.

5.2 Background

Information infrastructures continuously gather vast quantities of data through multiple sensors and applications at an increasing rate. Therefore, these databases often contain structured, semi-structured, and unstructured data that belong to multiple modalities (i.e., multi-modal data). Naturally, each source, type, and distribution modality has specific and complementary information [438]. Reasonable fusion strategies may help to understand the tasks we are addressing while contributing to the development of creative solutions [445]. Moreover, this increase in the availability and accessibility of multi-modal data motivates the research on the fusion of information and opens new opportunities to deepen our knowledge of cross-modality and shared-modality information [440].

Data fusion deals with the association, correlation, and combination of data and information from single and multiple sources and is defined as the study of efficient processes for the automatic or semi-automatic transformation of these data into a representation that supports human or automated decision making [446, 447]. This process has three important elements [440]: **data sources** (i.e., single or multiple data sources from different positions and at different points of time); **operation** (i.e., the operation of a combination of data and refinement of information); **purpose** (i.e., the main outcome of our application). Concerning the operation element, we consider it important

to mention some of the available hierarchies since it will be essential to understand the pipelines applied in the medical context. Luo and Kay [448] consider low-to-high levels to represent data in different fusion processes. These different levels relate to the diversity of input data patterns, the context of application or purpose of these data, and the degrees of promotion of information quality:

1. **Signal level:** also known as *low-level fusion* or *raw data fusion*, this strategy consists of directly combining raw data captured from sensors into fusion models. Raw data can be fused at this level if they are commensurate or in the same pattern. In the case of image processing, we may consider *pixel level fusion* as a special case of signal level fusion.
2. **Feature level:** also known as *medium level fusion* or *characteristic level fusion*, this strategy consists of converting raw data into features that will be used as input into fusion models.
3. **Symbol level:** also known as *high level fusion* or *decision level data fusion*. This strategy combines intermediate decisions to obtain a global and accurate decision.

A different proposal is described in [449]. In this hierarchy, the fusion processes take into account an input-output perspective that considers the nature of input data and output data, thus reducing uncertainty in the previous architecture:

1. **Data-In-Data-Out fusion:** in this type of fusion, a processing operation is applied to input data to make them more accurate or polished (e.g., signal processing and image processing).
2. **Data-In-Feature-Out fusion:** in this type of fusion, a feature engineering operation is applied to the input data to extract meaningful features.
3. **Feature-In-Feature-Out fusion:** in this type of fusion, a feature refinement operation is applied to the preliminary extracted features.
4. **Feature-In-Decision-Out fusion:** in this type of fusion, a decision algorithm is applied to the feature inputs to obtain a sequence of decisions.
5. **Decision-In-Decision-Out fusion:** in this type of fusion, a decision algorithm is applied to low-level decisions to generate a high-level decision.

These two hierarchies present a clear overview of the current approaches to data fusion strategies. Nevertheless, an extensive comparison of several data fusion strategies is available in [450, 451].

Multi-modal learning has been deeply explored in a variety of tasks. Neural image captioning [452] combines image and text data to generate image captions automatically, visual question answering [453] and visual reasoning [454] combine an image and question to get a coherent and correct answer, text and image for image retrieval [455, 456] uses image-text pairs to retrieve similar images, multi-modal classification takes advantage of different data modalities to achieve multi-modal representations [457] or multi-modal classifiers [458, 459] to solve a classification task.

5.3 Literature Review

An attentive look at the previous examples allows us to infer that most works described in literature explore combinations of image and natural language modalities; however, this is not necessarily a paradigm in the medical context. Lorenzi et al. [460] performed a multi-modal analysis of Alzheimer's disease using MRI and PET brain images through the description of the relationship between atrophy and hypo-metabolism employing a computationally tractable formulation of partial least squares regression. Suk et al. [461] use a Deep Boltzmann Machine (DBM) to extract hierarchical feature representations from CT and MRI images to diagnose Alzheimer's disease and mild cognitive impairment. Correa et al. [462] employed multi-set canonical correlation analysis [463] to perform data fusion of concurrently acquired fMRI and electroencephalography (EEG) data from multiple subjects. Guo et al. [435] conducted a comprehensive study on data fusion strategies for image classification and segmentation with CT, PET and MRI images from soft-tissue sarcoma and reported that the network trained with multi-modal images showed superior performance compared to networks trained with single-modal images, and, regarding tumour segmentation, performing image fusion within the network (e.g., fusing at convolutional or fully connected layers) is generally better than fusing images at the network output (e.g., voting). Dolz et al. [464] proposed a novel CNN architecture (HyperDense-Net) that integrates several levels of data fusion on brain image segmentation. In the domain of cervical dysplasia diagnosis, Xu et al. [465] and Chen et al. [466] explored different strategies, architectures and aggregation methods to leverage the informative power of cervical examination data (e.g., HPV tests, Pap tests, digital cervicography, patient information). Carneiro et al. [467] addressed the classification of mammograms using craniocaudal and mediolateral oblique views and their respective mass and micro-calcification segmentation (of the same breast), trained a separate CNN for each view and segmentation map, and, finally, used the features learned from each segmentation map and unregistered (i.e., not aligned) views to train a final CNN classifier that estimates the patient's risk of developing breast cancer. Liang et al. [468] implemented a Deep Belief Network (DBN) [469] framework to capture intra-modality and cross-modality relationships and to identify cancer subtypes from multi-platform genomics (e.g., gene expression, miRNA expression, DNA methylation) and clinical data. The latent features of each input modality are encoded with a Restricted Boltzmann Machine (RBM) [470], and then a joint representation of hidden variables is used to capture the common features resulting from multi-platform input data. The final states of these hidden variables are learned unsupervised, using contrastive divergence [471]. The final states of the joint representation of latent features are used to define the cancer subtypes based on these multi-modal input data. Multi-modal learning has also been used in other medical tasks besides image classification and segmentation. For instance, Cheng et al. [472] addressed deep similarity learning for multi-modal medical images with a multi-modal stacked denoising autoencoder trained on CT and MRI images, and Cao et al. [473] explored multi-modal approaches for medical image retrieval using medical images and their associated text descriptions. On the other hand, several approaches explore solely natural language processing (NLP) approaches [474]. Qin et al. [475] proposed a multi-modal model

to estimate the probability of early sepsis that incorporates clinical notes (about the patient) and structured data (i.e., patient measurements) and uses NLP models in the pipeline to obtain semantic representations of the text. Similarly, Jin et al. [476] developed an NLP service that performs named entity extraction and negation detection on clinical notes and composes the selected entities into a new text corpus to train document representations that are fused with time-series signals to train a multi-modal neural network that predicts the hospital mortality risk for patients.

5.4 Multi-modal Retrieval for the Aesthetic Assessment of Breast Cancer Surgery Outcomes

The results of this work have been disseminated in the form of a paper at an international conference: Mohammad Hossein Zolfagharnasab and Nuno Freitas and Tiago Goncalves and Eduard Bonci and Carlos Mavioso and Maria J. Cardoso and Helder P. Oliveira and Jaime S. Cardoso, “Predicting Aesthetic Outcomes in Breast Cancer Surgery: A Multimodal Retrieval Approach,” in *Proceedings of the Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care (in MICCAI 2024)*, 2024 [46]. The code related to the implementation of this work is publicly available in a GitHub repository¹.

5.4.1 Description and Contributions

Breast cancer remains a prevalent global health concern, though advancements in screening and treatment have significantly improved survival rates, exceeding 90% at 5 years in many high-income countries [477]. Consequently, attention within the clinical community has shifted towards enhancing post-surgery patient quality of life [478]. Locoregional treatments, such as radiation therapy and surgery, commonly employed in patients, often impact aesthetic outcomes, influencing body image significantly. Ongoing research aims to establish a standard for evaluating these outcomes [479]. Recent studies on the quality of life of breast cancer patients after treatment show that a patient’s body image has a high correlation to this dimension [480]. Despite that, the current methods for objective assessment using panel scores or software do not adequately correlate with the patient’s self-image [481]. Currently, clinicians use visual examples of similar cases to discuss potential outcomes of breast cancer surgery with patients [479]. However, as databases grow, selecting cases becomes time-consuming and subjective, highlighting the need for objective methods to efficiently retrieve clinically and visually relevant examples that meet both criteria. AI algorithms have emerged as crucial tools in this context, enabling accurate analysis of complex medical images [482, 483].

In Section 4.5, we introduced the topic of medical image retrieval and explained the fundamental concepts of these systems and their (potential) importance in the daily routine of several healthcare professionals. In this Section, we extend the topic to contextualise the reader regarding the rationale

¹<https://github.com/MsainZn/bcs-aesth-mm-modal-retrieval>

and intuition behind our hypotheses and experiments described here. Image retrieval systems are designed to retrieve samples from extensive databases [484] efficiently. They fall into two main categories: content-based image retrieval (CBIR), which analyses the visual content of a query image to find similar images; and collaborative filtering image retrieval (CFIR), which employs scoring mechanisms to identify similarities among available entities. In medical applications, CBIR systems are common due to operating without requiring detailed clinician annotation [485]. Examples include chest X-rays [486, 487] and brain MRIs [488]. While CFIR systems are less explored, some studies with CFIR systems [489, 490] report improved performances, likely because they incorporate more clinician input, which translates to learning more from the data. Thus, a retrieval system can be effective by allowing clinicians to show patients previous cases and allow them to understand the possible outcomes of treatment fully. These systems enhance efficiency and objectivity in retrieving relevant samples, thus improving diagnostic accuracy and patient outcomes. Despite these advancements, challenges such as the *semantic gap* between system analysis and user expectations, variability in image quality, and limited data availability persist, needing ongoing attention and innovation [491, 492]. Recent literature on image retrieval systems focuses extensively on enhancing feature extraction algorithms [493]. This includes convolutional neural network (CNN) and vision transformer (ViT) architectures, which follow a process of feature extraction to encode image features, pattern recognition for high-level representations, and embedding these features into a space where similar images are positioned close together for efficient retrieval by comparing feature vector distances [494].

With that in mind, we point the reader to the works of Wang et al. [495], Bhandi et al. [496], and Maji et al. [497], who evaluated the performance of CNN-based image retrieval architectures against traditional computer vision methods, and concluded that the first performed better than the latter. Naturally, with the recent advancement in transformers, similar investigations are carried out using ViTs [498]. For instance, Song et al. [499] and Dubey et al. [500] published pioneer studies on the comparison between CNN-based retrieval and ViT-based retrieval, showcasing the superior performance of ViTs across multiple datasets. Following these studies and using the same pre-trained models, Zhou et al. [501] reported viable accuracy improvements with a semi-supervised approach in chest X-ray application. For medical applications, variations of the CLIP model, especially MedCLIP, have also shown promising results in both zero-shot use cases and image-retrieval systems [502]. Using MedCLIP and a contrastive learning strategy with expert annotation, Kumar et al. [503] were able to improve the retrieval accuracy by almost 5% in all of their test cases. Similar improvements were reported from researchers working with CLIP in X-ray [504], and MRI images [505]. Studies such as Gao et al. [506] also reported improvements in multi-view mammograms.

Considering these challenges, this work introduces a deep learning (DL)-based multi-modal retrieval pipeline for predicting aesthetic outcomes in breast cancer patients post-locoregional treatments. Given a query image from the upper torso, our proposed algorithm retrieves the top 10 most analogous patient cases from the image database. While our focus is on pre-surgery and

post-surgery comparisons in breast cancer, this pipeline is adaptable to various domains. In this regard, our model can be categorised as a weak CFIR since it is trained to imitate the clinicians' rankings and take advantage of visual features. The main contributions of this work are:

1. A novel multi-modal retrieval framework for predicting the aesthetic outcomes of locoregional treatments of breast cancer patients that combines tabular clinical data and image features trained and evaluated on an annotated dataset.
2. A comparative study with ViTs and CNNs regarding the impacts of feature extraction on retrieval quality.

5.4.2 Methodology

This Section presents the details of the implementation of this work.

5.4.2.1 Data

Mathematical Modelling: Triplet Loss The triplet loss is a robust technique used in several tasks (e.g., image retrieval, face recognition, and metric learning) to learn embeddings that capture semantic similarity among data points [507]. It involves three entities: a query, a positive sample (similar to the query), and a negative sample (dissimilar to the query). Equation 5.1 shows the triplet loss formulation, where A , P , and N are embeddings of the query, positive, and negative samples, respectively, $dist(A, P)$ and $dist(A, N)$ represent distances (e.g., Euclidean or cosine) between these samples, and α is a margin hyper-parameter enforcing a minimum difference between distances for the loss to be zero. This approach ensures that semantically similar samples are closer to each other while dissimilar samples are farther apart in the embedding space. This approach effectively mimics the exact ranking determined by clinicians during training and optimises the weights to enhance the clustering of similar samples.

$$L(A, P, N) = \max(0, dist(A, P) - dist(A, N) + \alpha) \quad (5.1)$$

Dataset: Triplets of Catalogues For training with triplet-loss, the catalogues were organised into triplets, considering all possible combinations of query and annotated samples: each triplet includes a query, a positive sample, and a negative sample, where the negative sample is ranked further from the query than the positive one. Regarding pre-processing, we consider that several details deserve the reader's attention. Concerning the tabular data, we normalise each element in the tabular feature vector according to the training set (e.g., we apply min-max normalisation to the patient's height using the training set elements). Regarding image data, we resize all images to 224×224 (with zero padding), and, for each model (i.e., CNN or ViT), we use its specific image processing module (available on Hugging Face [508]) to process the images according to the model's requirements. With this pipeline, we can assess various models using the same framework.

5.4.2.2 Proposal

This study presents four different approaches for patient retrieval. The first approach focuses on clinical (tabular) data. The second uses pre-trained DL image models to extract features from the images. The third is similar but performs fine-tuning of the models on the training set. The fourth combines both the image and clinical data. Except for the third approach, all models were trained on the data. Using clinicians' ranking, we organise the data into triplets and utilise the triplet loss to update model weights. Figure 5.1 briefly summarises the retrieval pipeline proposed in this work.

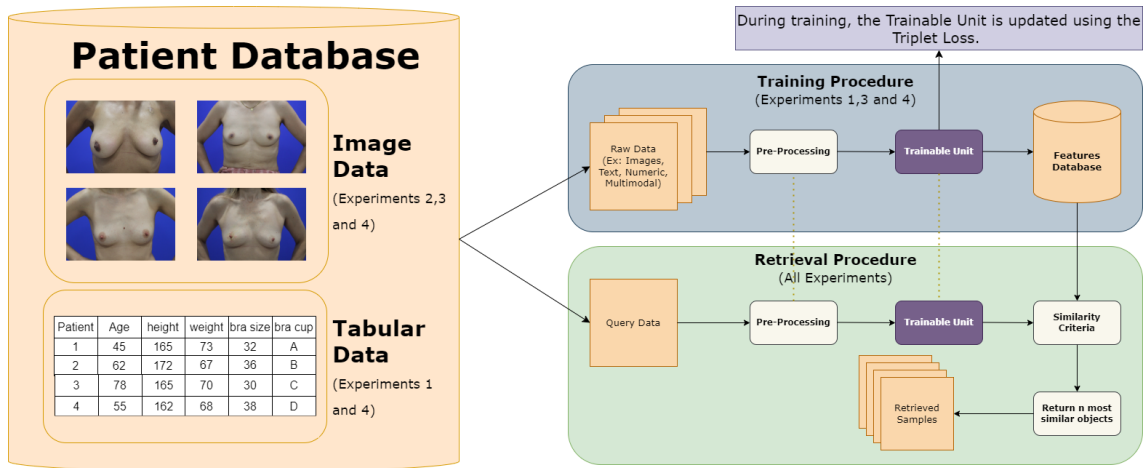


Figure 5.1: Illustration of the training and retrieval procedure. The trainable unit is optimised using the triplet loss to promote retrieval according to clinicians' ranking during training. It is worth noting that for images, the trainable unit must incorporate feature extraction, while for tabular data, it can be as simple as Weighted Euclidean Distance (WED). During retrieval, we use the trainable unit to extract the features corresponding to a given query. Afterwards, we use the Euclidean distance (similarity criterion) to calculate the distance between the query and the samples to retrieve the most similar samples. If we aim to improve efficiency, we can store the extracted feature vector of a particular image in a feature database.

Approach #1: Tabular Data Models The first approach only uses clinical (tabular) data. We started by using a Weighted Euclidean Distance (WED) to optimise the retrieval process because of the simplicity of the tabular data (i.e., a vector with five features). Nevertheless, the same triplet-loss formulation was employed to update the weights according to the clinicians' rankings. Equation 5.2 formalises the WED, where u is the feature vector for sample A, v for B, and W is the weight matrix. By setting the weight matrix as identity ($W = I$), we recover the standard (baseline) Euclidean distance. Unlike WED, baseline Euclidean contains no learnable parameter; thus, it requires no training. Besides WED, we employed a more flexible approach using a shallow 4-layer artificial neural network (ANN) for tabular data, consisting of layers with 16 neurons activated by a hyperbolic tangent ($\tanh$). Like WED, we use the triplet loss to update the model's weights, but the ANN provides more learnable parameters, which promotes capturing complex interactions between

features.

$$WED(u, v, W) = |u - v|_{1 \times N} W_{N \times N} |u - v|_{N \times 1}^T \quad (5.2)$$

Approach #2: Pre-trained Models For the second experiment, we shift our attention to the images, and we leveraged competitive DL models to perform feature extraction. It is important to note that this experiment did not include any additional fine-tuning with respect to clinicians' ranking. We did not update the model's weight based on triplet loss in this experiment, and we essentially captured a zero-shot feature extraction from the image models. For this purpose, we utilised a selection of pre-trained DL models, including well-known CNNs and ViTs. Below, we briefly describe the models we used in this work:

- **BEiT (ViT)**: Bao et al. [509] developed the BEiT model, extending the BERT methodology [510] and using a masked image modelling task to pre-train ViTs. Their method uses two image views in the pre-training procedure (i.e., patches and tokens), and the objective is to recover tokens based on corrupted patches.
- **DeiT (ViT)**: Touvron et al. [229] worked on top of the ViT baseline architecture [104], and proposed an attention-driven knowledge distillation method based on distillation token (of the student network) optimised to reproduce the label estimated by the teacher network (similarly to a classification token).
- **DINOv2 (ViT)**: Oquab et al. [511] trained a large ViT model using a novel self-supervised trained approach and an automatic pipeline to build a curated image dataset and distilled it into a series of smaller models.
- **Google-ViT (ViT)**: Wu et al. [512] introduced the Visual Transformer (VT), using spatial attention to convert the feature map into a compact set of semantic tokens and feeding these tokens to a self-attention module to capture token interactions.
- **ResNet50 (CNN)**: He et al. [263] proposed a deep residual learning framework, where residual connections across the network facilitate the learning of deeper networks by guiding the stacks of layers to fit a residual mapping, instead of directly fitting an underlying mapping.
- **VGG16 (CNN)**: Simonyan and Zisserman [513] were pioneers in the study of the depth (16 to 19 layers) of deep neural networks with tiny (3×3) convolutional kernels.

Approach #3: Fine-tuned Models The third approach mirrors the second. In this case, we utilised triplet loss to fine-tune the extracted feature vectors from the images to align with clinicians' ranking criteria. This approach serves two purposes: ensuring semantic alignment with clinicians' objectives and emphasising features that aid in replicating clinicians' rankings. As a result, this approach falls under weak CFIRs, where clinicians' predefined scores enhance sample retrieval for new cases.

Approach #4: Multi-modal Models The fourth approach enhances the trainable DL model by incorporating multi-modal inputs. In this case, we concatenate the feature vectors from fine-tuned image models with the tabular feature vectors. This combined vector is later used as the input of another shallow MLP with the same properties as the first experiment, and the model weights are optimised with the triplet loss.

Evaluation Metrics Equations 5.3–5.4 present the Adjusted Discounted Cumulative Gain (ADCG) and Contrastive Accuracy (CAcc). We used these metrics to ensure a more accurate assessment of our models' performance, taking advantage of some of their benefits for the evaluation of the performance of retrieval systems: ADCG accounts for prioritising the relevant instances with respect to the query, which makes it particularly interesting for our application where results ordering is crucial; and, CAcc offers insights into the model's ability to distinguish between similar and dissimilar items in relation to a query image, a fundamental detail for our dataset as it ensures a detailed comparison of all catalogue elements against the query, enabling a comprehensive assessment of all samples. We remind the reader that Section 4.5 contains more information about retrieval assessment metrics.

$$DCG_p = \sum_{i=1}^p \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)}, \quad ADCG_p = \text{Norm}_{MinMax}(DCG_p) \quad (5.3)$$

$$\text{CAcc} = \sum_{\text{cat}=1}^m \frac{\text{Number of correct predictions per triplet}}{\text{Total number of triplets}} \quad (5.4)$$

5.4.3 Results and Discussion

This study aimed to develop and evaluate a DL-based multi-modal retrieval system to predict aesthetic outcomes in breast cancer patients after locoregional treatments. This system combines tabular clinical data and image features to provide a robust retrieval framework. Table 5.1 compares different methods and models regarding their performance. Furthermore, we developed distinct retrieval models based on varying aesthetic assessments (i.e., *Excellent/Good* and *Fair/Poor*). Clinicians typically require retrieval of positive and negative treatment outcomes to illustrate diverse scenarios to patients. This separation enables our algorithm to address these preferences and improve its ability to accurately reflect clinicians' needs while ensuring each scenario is appropriately represented.

5.4.3.1 Tabular Data Retrieval

In this approach, we used clinical (tabular) data to perform the retrieval using WED and an MLP. The WED method provided a baseline for comparison, revealing moderate performance on both datasets. The MLP approach outperformed the WED, achieving higher CAcc and ADCG values. However, no model achieves performance superior to 66% on ADCG and 65% on CAcc, both in training and testing. These results suggest that clinical data alone may not provide sufficient information to

achieve high retrieval accuracy, indicating its limitations in capturing the full complexity required for precise aesthetic outcome prediction.

5.4.3.2 Image Retrieval

We report the results of pre-trained and fine-tuned models for image retrieval (see Table 5.1). Generally, pre-trained models exhibit lower performance when compared to most tabular models across both datasets, potentially due to their limited alignment with specific aesthetic criteria used by clinicians. On the other hand, we note that ViTs (especially BEiT) appear to be the standout performers. Intuitively, by fine-tuning pre-trained DL models to better align their feature vectors with the clinicians' ranking criteria, we observe that the performance of all image models was notably improved. For instance, the fine-tuned VGG16 shows an improvement of CAcc from 52.10% to 69.26% and ADCG from 46.71% to 76.87% in the test set. These improvements underscore the importance of tailoring pre-trained models for retrieval tasks, as it allowed them to better capture the semantic and aesthetic considerations critical to clinicians. When comparing architectures (i.e., CNNs and ViTs), ViTs seem to have a promising potential (e.g., Google-ViT stood out as the top performer, achieving superior results in 5 out of 8 performance metrics evaluated).

5.4.3.3 Multi-modal Retrieval

The final method integrated tabular and image features, resulting in a more comprehensive input for the retrieval system. As expected, the multi-modal approach achieved the highest performance, with an average 1-2% improvement in the retrieval accuracy metrics across all evaluation criteria. Besides improving accuracy, utilising multi-modal data is valuable as it promotes robustness of the overall prediction in case the DL models under-perform or encounter out-of-distribution samples. For such scenarios, the clinical data may provide structured, interpretable insights that help maintain consistency and reliability in the predictions, ensuring more informed decision-making in clinical settings.

5.4.4 Conclusions and Future Work

This study evaluated various configurations of clinical and image-based retrieval methods using triplet loss optimisation for predicting aesthetic outcomes in breast cancer patients following loco-regional treatments. Results show that clinical (tabular) data alone attains moderate performance, indicating limitations in capturing complexities for precise aesthetic outcome prediction. On the other hand, ViTs (e.g., BEiT) showed superior performance compared to traditional CNNs (i.e., ResNet50 and VGG16), especially after fine-tuning. Finally, we found that integrating clinical (tabular) data with image features resulted in the highest overall performance, possibly due to combining structured clinical insights with nuanced image features. Multi-modal BEiT showed

Table 5.1: Retrieval performance metrics for various models categorised by aesthetic quality. Each entry is of the form (Excellent/Good, Fair/Poor). Best results, per model, highlighted in bold.

Type	Model	Train		Test	
		CAcc (%) $\uparrow$	ADCG (%) $\uparrow$	CAcc (%) $\uparrow$	ADCG (%) $\uparrow$
Tabular	Baseline	(61.37, 61.77)	(65.21, 63.52)	(55.68, 61.87)	(57.75, 62.05)
	Matrix	(60.41, 60.40)	(65.29, 62.02)	(57.58, 61.58)	(58.54, 65.31)
	MLP	(62.65, 64.89)	(66.64, 64.89)	(57.09, 64.90)	(57.70, 66.56)
Pre-trained	VGG16	(52.02, 46.68)	(54.55, 46.42)	(52.10, 44.72)	(56.69, 46.71)
	ResNet50	(54.96, 47.81)	(58.75, 50.56)	(53.38, 50.56)	(55.20, 53.74)
	DeiT	(52.08, 45.85)	(52.87, 46.24)	(53.25, 44.47)	(54.87, 53.13)
	Google-ViT	(53.24, 47.29)	(56.21, 49.99)	(52.10, 49.09)	(54.87, 53.13)
	DINOv2	(56.00, 46.31)	(58.88, 47.82)	(54.88, 46.15)	(60.231, 44.95)
	BEiT	(56.03, 49.25)	(61.06, 49.39)	(54.79, 50.98)	(57.58, 54.35)
Fine-tuned	VGG16	(90.51, 90.55)	(95.18, 95.42)	(69.26, 71.16)	(71.96, 76.87)
	ResNet50	(89.47, 87.11)	(94.22, 93.52)	(68.33, 72.17)	(70.30, 78.29)
	DeiT	(76.40, 73.58)	(62.75, 79.02)	(60.06, 63.09)	(64.61, 64.58)
	Google-ViT	(92.65, 91.95)	(96.45, 96.28)	(69.30, 73.18)	(73.87, 79.86)
	DINOv2	(89.39, 87.77)	(94.38, 93.82)	(69.92, 70.78)	(77.89, 76.97)
	BEiT	(89.78, 95.33)	(93.95, 81.88)	(71.95, 72.76)	(73.85, 77.92)
Multi-modal	Google-ViT	(98.69, 98.91)	(99.39, 99.56)	(68.55, 73.85)	(71.14, 80.62)
	DINOv2	(97.89, 98.25)	(99.10, 99.33)	(70.14, 73.56)	(77.40, 78.96)
	BEiT	(99.10, 98.87)	(99.65, 99.65)	(72.06, 73.05)	(73.63, 78.45)

superior retrieval accuracy on the test set. Apart from exploring the potential of ViTs (see Section 5.5), future work should examine the potential for image-segmentation models instead of image classifiers to achieve higher performance and evaluate out-of-distribution instances, given that the current dataset has limited variability.

5.5 Attention-Driven Multi-modal Retrieval for the Aesthetic Assessment of Breast Cancer Surgery Outcomes

The results of this work have not been disseminated in any other venue besides this thesis. The code related to the implementation of this work is publicly available in a GitHub repository².

5.5.1 Description and Contributions

Section 5.4 presented a pioneering approach to multi-modal retrieval for the aesthetic assessment of breast cancer surgery outcomes. From that work, we concluded that using triplet loss optimisation, multi-modal ViTs attained the best performances. Those results align well with the scope of this thesis because they revisit the potential of attention mechanisms in deep learning architectures

²<https://github.com/TiagoFilipeSousaGoncalves/bcs-aesth-mm-attention-mir>

(through transformers) and the integration of relevant clinical information (through tabular data) to the models with the hope of bringing more transparency and achieving results with more quality. In this work, we build upon the framework proposed in [46] and extend the experiments on the transformer-based architectures, comparing ten different models and two different optimisation settings (i.e., fine-tuned and multi-modal). We validate the results published in the original paper and report new ones with the introduced architectures.

5.5.1.1 Brief Overview of the Transformer Models

In this Section, to contextualise the reader, we present a brief review of the transformer-based models used in this study (please note that we do not repeat the presentation of the models used in the previous work).

ConViT D’Ascoli et al. [514] designed the ConViT architecture, introducing the gated positional self-attention (GPSA), a self-attention layer that mimics the properties of convolutional layers.

CrossViT Chen et al. [515] developed the CrossViT architecture, a dual-branch transformer that processes small-patch and large-patch tokens with two separate branches of different computational complexity and fuses these tokens purely by attention to multiple times to complement each other and produce stronger image features.

DaViT Ming et al. [516] created the DaViT architecture, a dual attention ViT that exploits self-attention mechanisms with spatial tokens (related to the feature dimension) and channel tokens (related to the feature representation).

LeViT Graham et al. [517] proposed the LeViT architecture, aiming for a paradigm of strategies to shrink ViTs, such as using attention as a downsampling mechanism, replacing the ViT positional embedding with a translation-invariant attention bias or optimising the computational efficiency of certain layers.

MaxViT Tu et al. [518] released the MaxViT architecture, which relies on a multi-axis attention module composed of blocked local and dilated global attention to capture both local and global spatial interactions throughout every stage of the network and achieve global perception in linear complexity.

MViTv2 Li et al. [519] extended the multi-scale vision transformers framework, resulting in the MViTv2 architecture, that introduced improvements in attention pooling along two axes, with shift-invariant positional embeddings using decomposed location distances to inject position information, and a residual pooling connection to compensate the effect of pooling strides in attention computation.

5.5.2 Methodology

We follow the methodology described in Section 5.4.2.

5.5.3 Results and Discussion

Table 5.2 presents the retrieval performance metrics (CAcc and ADCG) for the models tested in this work, categorised by aesthetic quality (i.e., related to catalogues with Excellent and Good or Fair and Poor aesthetic outcomes). Unlike Table 5.1, we organise these results by model instead of type. Looking at the training set, we can observe that the multi-modal approaches often attain the best results, showing that adding patient clinical data brings meaningful insights to the learning process during training. Contrastively, the results on the test set are less consistent with the image (fine-tuned) approaches being equally competitive with multi-modal strategies, suggesting that the latter might be more prone to over-fitting. Moreover, this tendency to over-fit is clearer when we look at the differences in results between the train and test sets, where we notice that the performance for both metrics decreases systematically from one subset to the other.

Focusing on the main conclusions presented in Section 5.4, we observe that these results align well with the latter, with attention-based models attaining higher performances. On the other hand, when extending this framework only to transformer architectures, we observe that, at this point, it is not clear whether it is preferable to use the image (fine-tuned) or the multi-modal. If we look at the performance metrics, results suggest that using only the image information is sufficient for the model to perform a good retrieval. On the other hand, one can argue that the multi-modal model has more information about the patients, and even if it suffers from a marginal performance decrease, it is still more interpretable. Besides, we also highlight that the size of the current dataset (see Chapter 3) might not be enough to draw additional conclusions regarding the path that leads to the best results.

5.5.4 Conclusions and Future Work

The extension of the framework presented in Section 5.4 to transformer-based architectures confirms the performance advantages of using these models for retrieving cases with similar aesthetic outcomes of breast cancer surgery. On the other hand, this setting suggests that only using image features might be equivalent to using multi-modal data, undermining the intuition of using more information about the patient to achieve refined results. Although the differences in results are marginal in most cases, we believe there is still space to improve both the image and the multi-modal approaches, contributing towards an end-to-end retrieval framework. To achieve this, we should devote future work to gathering more data and building new catalogues (preferably with more diversity regarding the aesthetic outcomes, conditions of acquisition, patients' ethnicity or type of treatment), training a segmentation or detection algorithm to extract the region of the image that contains the breasts (and discard the rest), or training the multi-modal framework in a fully end-to-end strategy.

Table 5.2: Retrieval performance metrics for attention-driven models, categorised by aesthetic quality. Each entry is of the form (Excellent/Good, Fair/Poor). Best results, per model, highlighted in bold.

Model	Type	Train		Test	
		CAcc (%) $\uparrow$	ADCG (%) $\uparrow$	CAcc (%) $\uparrow$	ADCG (%) $\uparrow$
Beit	Image (Fine-tuned)	(93.05, 96.02)	(97.00, 98.46)	(69.95, 73.28)	(74.48, 76.66)
	Multi-modal	(97.88, 98.40)	(99.25, 99.39)	(68.75, 71.71)	(71.31, 74.21)
ConViT	Image (Fine-tuned)	(97.11, 96.02)	(98.93, 98.58)	(70.41, 75.19)	(73.29, 79.87)
	Multi-modal	(99.39, 98.07)	(99.65, 99.24)	(65.79, 71.33)	(67.16, 75.05)
CrossViT	Image (Fine-tuned)	(96.72, 96.49)	(98.76, 98.51)	(67.87, 71.75)	(69.43, 77.10)
	Multi-modal	(98.67, 98.44)	(99.45, 99.35)	(66.17, 70.22)	(66.77, 73.50)
DaViT	Image (Fine-tuned)	(97.43, 96.93)	(98.91, 98.86)	(69.12, 74.72)	(69.35, 79.96)
	Multi-modal	(98.92, 98.07)	(99.39, 99.18)	(69.87, 74.00)	(73.19, 80.32)
DeiT	Image (Fine-tuned)	(92.13, 93.30)	(96.58, 97.37)	(64.42, 72.60)	(65.48, 77.20)
	Multi-modal	(98.20, 97.76)	(99.30, 99.19)	(64.96, 70.77)	(66.27, 75.79)
DinoV2	Image (Fine-tuned)	(91.92, 91.01)	(96.41, 95.87)	(68.04, 74.47)	(68.03, 78.51)
	Multimodal	(97.61, 98.01)	(99.15, 99.16)	(68.87, 75.49)	(70.60, 80.06)
Google	Image (Fine-tuned)	(96.51, 96.07)	(98.60, 98.67)	(69.62, 74.38)	(69.91, 79.21)
	Multi-modal	(96.17, 96.71)	(98.41, 98.63)	(70.25, 74.30)	(70.89, 76.98)
LeViT	Image (Fine-tuned)	(69.79, 93.98)	(75.23, 97.64)	(61.55, 67.88)	(64.21, 71.54)
	Multi-modal	(75.63, 91.85)	(82.53, 96.46)	(62.96, 70.01)	(65.99, 72.79)
MaxViT	Image (Fine-tuned)	(96.70, 96.03)	(98.73, 98.58)	(69.25, 74.13)	(70.28, 77.78)
	Multi-modal	(98.39, 93.93)	(99.39, 97.40)	(70.12, 71.84)	(72.99, 75.03)
MViTv2	Image (Fine-tuned)	(96.57, 96.79)	(98.75, 98.71)	(70.45, 73.58)	(72.46, 79.10)
	Multi-modal	(98.51, 97.96)	(99.28, 99.21)	(67.50, 71.71)	(71.13, 75.30)

5.6 Multi-modal Approaches for Computational Pathology

The results of this work have not been disseminated in any other venue besides this thesis. The code related to the implementation of this work will be available in a GitHub repository³.

5.6.1 Description and Contributions

Section 4.6 presented a study on applying multiple instance learning (MIL) with attention mechanisms to characterise tumour and immune phenotypes from the H&E slide of primary breast tumours. While we obtained interesting results using only the whole slide images (WSIs) (i.e., image modality), we hypothesised that the models could benefit from including patient clinical data in the pipeline. Additionally, following the intuition presented in Sections 5.4 – 5.5, we also argue that the integration of clinically meaningful data (i.e., data that healthcare professionals use

³<https://github.com/QTIM-Lab/xai-mmml-prediction-brca-tip>

to interpret or make a specific clinical decision) into these pipelines contributes to the development of more informed, transparent and trustworthy algorithms. In this Section, we present our findings on integrating clinical data into DL algorithms to predict these (therapeutically relevant) facets of the tumour microenvironment (TME). Figure 5.2 shows an overview of the proposed pipeline.

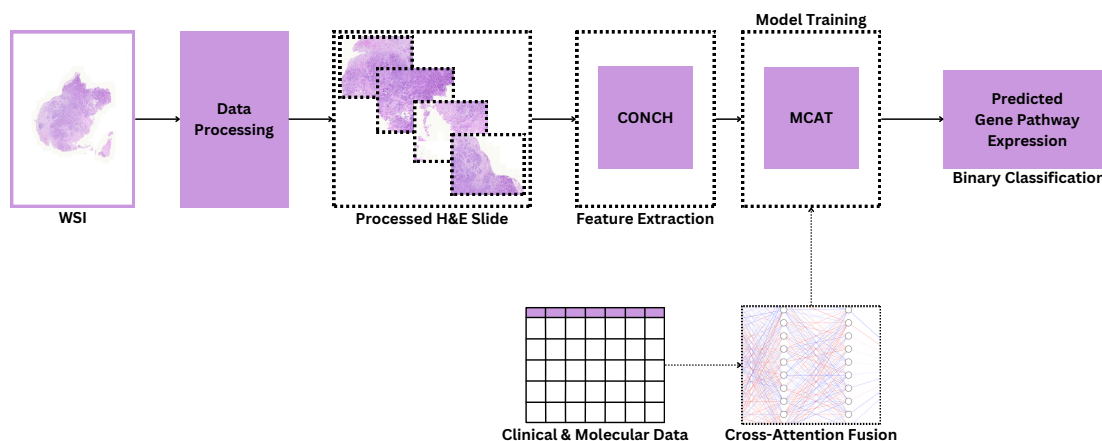


Figure 5.2: Overview of the proposed pipeline.

5.6.1.1 Brief Update on the Rise of Foundational Models for Computational Pathology

Almost immediately after the publication of the work described in Section 4.6 [48], the research on foundational models for pathology saw the release of (at least) three new foundational models, publicly available to the scientific community to train (or fine-tune) in their in-house tasks or to use the feature extractor to obtain embeddings for use in downstream tasks. More interestingly, these foundational models were published (almost) simultaneously by the group led by Faisal Mahmood [520].

Lu et al. [521] proposed the contrastive learning from captions for histopathology (CONCH), a visual-language foundation model developed using diverse sources of histopathology images, biomedical text and image–caption pairs through task-agnostic pre-training. Chen et al. [522] developed UNI, a general-purpose self-supervised model for pathology (pre-trained on a large dataset of H&E WSIs across (major) tissue types and evaluated on representative computational pathology tasks) that showed relevant modelling capabilities (e.g., resolution-agnostic tissue classification, slide classification using few-shot class prototypes, disease sub-typing generalisation). Recently, Ding et al. [523] introduced TITAN, a multi-modal foundation model (pre-trained via visual self-supervised learning and vision-language alignment with corresponding pathology reports and captions generated from a multi-modal generative artificial intelligence (AI) copilot for pathology [524]) that can extract general-purpose slide representations and generate pathology reports that generalise to resource-limited clinical scenarios.

5.6.2 Brief Review of Multi-modal Machine Learning for Computational Pathology

The intuition that motivates multi-modal machine learning in computational pathology is that alternate modalities (e.g., clinical data, transcriptomics) can complement the phenotypic information from histology (cells types and organisation) with signals that potentially linked to prognosis [525]. The computational pathology community has been leveraging these techniques to predict outcomes and discover prognostic features correlated with poor and favourable outcomes (including cancer survival prediction) [526, 527]. Following the trend in other research areas, multi-modal machine learning has also invested in transformer-based architectures [528]. Chen et al. [529] proposed the multi-modal co-attention transformer (MCAT) architecture to learn an interpretable and dense co-attention mapping between WSIs and genomic features encoded in an embedding space. On top of MCAT, Jaume et al. [530] designed the SurvPath model to learn biological pathway tokens from transcriptomics that can encode specific cellular functions, combined with histology patch tokens that encode the slide morphology, using a multi-modal transformer to model interactions between pathway and histology patch tokens.

5.6.3 Methodology

5.6.3.1 Data

For data preparation (and pre-processing), we followed Section 4.6.3.

5.6.3.2 Models and Training Details

Comparison of Feature Extractors Before experimenting with multi-modal architectures, we repeated the experiments reported in [48] with a new feature extractor. In this case, we ran the AM-SB architecture with CONCH features to understand if there was any gain in changing the feature extractor for the following experiments.

Models using Clinical-only Data The decision on the set of clinical features to use in the multi-modal approaches happened iteratively after several discussions with clinicians. The final set of clinical features (reported in this thesis) with which we did the experiments is: PAM50 molecular subtypes (PAM50), test result for estrogen receptor (ER), test result for progesterone receptor (PR), test result for human epidermal growth factor receptor 2 (HER2), gender/sex (G), race (R), and histological type (HT). Before testing with multi-modal DL architectures, we sought to analyse the predictive power of the clinical features (without information from the WSI). We performed an intermediate experiment with classic machine learning models (i.e., decision tree (DT), Gaussian naive Bayes (GNB), logistic regression (LR), support vector machine (SVM)), using specific combinations of clinical features (after discussions with clinicians). Apart from different combinations of features, we also tried different feature encodings: categorical encoding and one-hot encoding.

Models using Multi-modal Data Finally, we adapted the MCAT architecture for our use case, performing modifications in the workflow of the transcriptomics biological pathway tokens to accommodate clinical data, keeping the cross-attention module to combine the last with WSI features. Given the difference of dimensions between CONCH features (i.e., 512) and the clinical features (i.e., between 1–10), we use a learnable multilayer perceptron (MLP) to increase the size dimension of these features (by a factor of 100) so that they are comparable to the WSI features.

5.6.4 Results and Discussion

Table 5.3 presents the area under the receiver operating characteristic curve (AUROC) results on the test set, obtained for the best model of each feature set (ResNet and PLIP results transcribed from [48]). These results suggest that models trained with CONCH-derived features generally perform better than ResNet-derived features. When comparing with PLIP-derived features, we also observe that models with CONCH-derived features perform better in most tasks (6 out of 10), suggesting that replacing PLIP with CONCH is a good option regarding the feature extractor.

Table 5.4 presents the AUROC results on the test set obtained for the different clinical-only data machine learning models. We observe that the best set of clinical features is {PAM50, ER, PR, HER2, G, R, HT} (21 out of 40 runs) with one-hot encoding. With the proper encoding, results suggest that using more clinical information might contribute towards increasing the predictive power of the models. Focusing on the models, we can see that the winning approaches are the LR (5 out of 10 tasks) and the SVM (4 out of 10 tasks). This intermediate (but relevant) step provides us with more intuition about the potential gain of using clinical data together with imaging information from the WSI to (hopefully) increase the performance approaches presented in Section 4.6.

Table 5.5 presents the AUROC results on the test set obtained for the different multi-modal MCAT models. From these results, we can infer that multi-modal approaches (using the MCAT framework) can be competitive when compared (5 out of 10 tasks) with the baseline model that solely uses imaging (WSI) features (AM-SB with CONCH). On the other hand, in this case, the main challenge is selecting the set of features that maximises the predictive power of the models. From the different combinations presented, the {PAM50, ER, PR, HER2} (together with WSI) seems appropriate, achieving the best results with the group of multi-modal experiments. However, predictive performance suffers in half of the tasks that combine clinical data.

5.6.5 Conclusions and Future Work

This Section presented a straightforward follow-up to the work published in [48]. After several discussions with clinicians, we tried to understand if performing data fusion of clinical (tabular) data and WSI (image) information would contribute to increasing the predictive performance of our algorithms. Along the way, we updated our feature extractor (we are using CONCH instead of PLIP), analysed the predictive signal of the clinical data (only), using classical machine learning approaches and different combinations of clinical features (with different encodings), and adapted

Table 5.3: **AUROC results on the test set, obtained for the best model of each feature set.** The best-performing feature-extraction strategy is highlighted in bold. Models trained with CONCH-derived features generally perform better than ResNet-derived features. When comparing with PLIP-derived features, we also observe that models with CONCH-derived features perform better in most tasks (6 out of 10), suggesting that replacing PLIP with CONCH is a good option regarding the feature extractor.

Task	Features		
	ResNet50	PLIP	CONCH
Angiogenesis	0.7053	0.7435	0.7605
Antigen Processing and Presentation	0.7599	0.7924	0.8260
B-cell Proliferation	0.7322	0.7755	0.8039
Cell Cycling	0.7809	0.7939	0.8100
Epithelial-Mesenchymal Transition	0.8110	0.8082	0.8120
Fatty Acid Metabolism	0.6323	0.6030	0.6649
Glycolysis	0.7996	0.8330	0.8276
Immunosuppression	0.8133	0.8542	0.8469
Oxidative Phosphorylation	0.6926	0.7332	0.7194
T-cell Mediated Cytotoxicity	0.7564	0.7770	0.7934

the MCAT framework [529] to our tasks and experimented with diverse settings (i.e., combinations of WSI and clinical features).

The AUROC results obtained with the clinical-only models showed that the clinical information could be an interesting predictor of the expression of the gene pathways reported in this thesis. Naturally, we should devote future efforts to running a more systematic study with all the (possible) combinations of clinical features (and encodings). Nevertheless, these results show that we might be in the right direction regarding the potential use of clinical data in our DL algorithms. Preliminary results obtained with the multi-modal approaches (using MCAT) partially support this conclusion. Although it is not completely clear that multi-modal approaches perform better than the unimodal (i.e., with image modality), it is possible to observe that the predictive power of the algorithms does not suffer too much (and sometimes improves). That said, we believe this line of research is worth our attention in the future. Besides MCAT, we think the SurvPath architecture with our data, with different feature combinations, might be a good starting point. However, there is still a margin to improve the image modality inputs with data augmentation [531, 532].

5.7 Summary and Conclusions

Developing algorithms that benefit from multiple data modalities during their training, inference or both phases can contribute to the conversion of black-box into white-box algorithms, given that using different sources of (hopefully meaningful) information might mitigate the use of spurious correlations by the models during their logical processing. Naturally, dealing with disparate data sources (as healthcare practitioners often do) is not a trivial task. Besides, it is of utmost importance that we understand first the reasoning process of healthcare professionals and how they use each

Table 5.4: **AUROC results on the test set, obtained for the different clinical-only data machine learning models.** The best results are highlighted in bold. Each column presents a specific set of clinical features (selected based on discussions with clinicians). Note: We use (C) for categorical and (1H) for one-hot encoding.

Model	Task	AUROC			
		ER,PR (C)	ER,PR,HER2 (C)	ER,PR,HER2,PAM50 (C)	PAM50,ER,PR,HER2,G,R,HT (1H)
LR	Angiogenesis	0.5513	0.5462	0.6045	0.5954
	Antigen Processing and Presentation	0.5619	0.5889	0.6510	0.6140
	B-cell Proliferation	0.5712	0.5715	0.6662	0.6026
	Cell Cycling	0.6779	0.7185	0.7715	0.8258
	Epithelial Mesenchymal Transition	0.5422	0.5371	0.6102	0.6300
	Fatty Acid Metabolism	0.6099	0.6326	0.6323	0.6299
	Glycolysis	0.6363	0.6609	0.7049	0.7186
	Immunosuppression	0.6257	0.6334	0.6315	0.7224
	Oxidative Phosphorylation	0.5021	0.5721	0.6658	0.5309
	T-Cell Mediated Cytotoxicity	0.6094	0.6345	0.7034	0.7025
DT	Angiogenesis	0.5846	0.5151	0.5901	0.5594
	Antigen Processing and Presentation	0.5904	0.5525	0.5458	0.5964
	B-cell Proliferation	0.6133	0.5726	0.5372	0.5758
	Cell Cycling	0.6315	0.7228	0.8415	0.8031
	Epithelial Mesenchymal Transition	0.5404	0.5025	0.6235	0.6259
	Fatty Acid Metabolism	0.6155	0.6413	0.5732	0.5866
	Glycolysis	0.6367	0.6775	0.6932	0.7026
	Immunosuppression	0.6398	0.6425	0.6588	0.6952
	Oxidative Phosphorylation	0.4835	0.5449	0.5737	0.4655
	T-Cell Mediated Cytotoxicity	0.5486	0.6474	0.6724	0.6435
GNB	Angiogenesis	0.5481	0.5553	0.5646	0.5765
	Antigen Processing and Presentation	0.4785	0.5195	0.6256	0.6165
	B-cell Proliferation	0.5744	0.5470	0.5844	0.5738
	Cell Cycling	0.6678	0.7047	0.7393	0.8159
	Epithelial Mesenchymal Transition	0.5491	0.5463	0.5615	0.6300
	Fatty Acid Metabolism	0.6028	0.6333	0.6321	0.6694
	Glycolysis	0.6362	0.6640	0.6986	0.7104
	Immunosuppression	0.6295	0.6563	0.6653	0.6976
	Oxidative Phosphorylation	0.5402	0.5449	0.6268	0.5346
	T-Cell Mediated Cytotoxicity	0.5922	0.6474	0.6893	0.6659
SVM	Angiogenesis	0.5492	0.5054	0.5946	0.6066
	Antigen Processing and Presentation	0.5111	0.5357	0.6037	0.6089
	B-cell Proliferation	0.5869	0.5477	0.6369	0.6383
	Cell Cycling	0.6241	0.7246	0.7930	0.8455
	Epithelial Mesenchymal Transition	0.5167	0.5049	0.6152	0.6413
	Fatty Acid Metabolism	0.6122	0.6466	0.6637	0.6101
	Glycolysis	0.6259	0.6353	0.6909	0.7436
	Immunosuppression	0.6553	0.6697	0.6797	0.6592
	Oxidative Phosphorylation	0.5017	0.5419	0.6212	0.5306
	T-Cell Mediated Cytotoxicity	0.6195	0.6243	0.6630	0.6942

Table 5.5: **AUROC results on the test set, obtained for the different multi-modal MCAT models.** The best results are highlighted in bold. Each column presents the results of the WSI features (CONCH) combined with a specific set of clinical features (selected based on discussions with clinicians). The baseline results correspond to the first column.

Task	AUROC				
	WSI	ER,PR	ER,PR,HER2	PAM50,ER,PR,HER2,G,R,HT	PAM50,ER,PR,HER2
Angiogenesis	0.7605	0.7681	0.7530	0.7322	0.7783
Antigen Processing and Presentation	0.8260	0.7116	0.7631	0.7757	0.7269
B-cell Proliferation	0.8039	0.7732	0.7629	0.7419	0.7604
Cell Cycling	0.8100	0.7707	0.7664	0.8316	0.8554
Epithelial Mesenchymal Transition	0.8120	0.8134	0.8209	0.8198	0.8078
Fatty Acid Metabolism	0.6649	0.6289	0.5935	0.6602	0.6831
Glycolysis	0.8276	0.8142	0.8137	0.8186	0.8061
Immunosuppression	0.8469	0.7996	0.8227	0.7940	0.8014
Oxidative Phosphorylation	0.7194	0.7105	0.6372	0.6903	0.7267
T-cell Mediated Cytotoxicity	0.7934	0.7345	0.7535	0.7643	0.7500

source of information so that we input or encode these data into the models correctly, ensuring that the models are, indeed, more informed and using that information effectively. In this Chapter, we tackled the topic of multi-modal machine learning from the perspective of medical image retrieval and computational pathology.

The main idea in medical image retrieval using DL is to achieve an embedding space that encodes the relevant information required to retrieve the most similar cases to a given query. Using the aesthetic outcomes of breast conservative surgery as the use case, we proposed a multi-modal framework that allows practitioners to optimise the retrieval process independently of the feature extractor backbone they use. In that sense, to establish the bridge with interpretability and transparency and connect this work to the topics presented in Chapter 4, we also did a quasi-exhaustive study with attention-based mechanisms (in this case, transformers) to assess their potential. We concluded that transformers are the models that learn the best latent spaces for this retrieval task and that integrating multi-modal data (often) contributes to increasing their predictive performance. While the size and diversity of data and annotations of our dataset present (perhaps) the main limitations of our study, we sought to explore this framework with other approaches or algorithms (e.g., segmentation). Additionally, the multi-modal step of this pipeline has a margin of improvement. In the studies presented in this Chapter, we optimise the multi-modal part after training the (image) feature extractor. While this process saves us time (it is faster and easier to optimise one module at a time), we think it would be more interesting and elegant to achieve a fully end-to-end framework that only requires one training procedure.

In computational pathology, the intuition is that the use of patient clinical data in the DL algorithms, combined with the WSI (image) information, might lead us to improved algorithms while increasing their robustness (given that we have different sources of data that contribute with more or less weight to the prediction) and transparency (given that the users will know what is the data that the models are using to compute the prediction). As we did in Chapter 4, the follow-up of this line of work should begin with interpretable post-hoc machine learning through a detailed

analysis of the attention maps produced by these multi-modal frameworks that use transformers and cross-attention and then move into in-model approaches. Besides, a different line of work should focus on multi-task (multi-label) models to achieve a fully end-to-end model that can predict a result for all the tasks (our current approaches require training 10 different models, i.e., one per task).

Chapter 6

Learning in a Human Understandable Way: Inherently Interpretable Models

Foreword on Author Contributions

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

1. I. Rio-Torto, T. Gonçalves, J. S. Cardoso, and L. F. Teixeira, “On the Suitability of B-cos Networks for the Medical Domain,” in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2024 [47]

We also include experimental work (see Sections 6.5 – 6.6) that was not published when we submitted this thesis. Nevertheless, we hope that these results or improved versions of these works will be published in the foreseeable future.

6.1 Introduction

The high success of machine learning algorithms in several tasks has motivated their implementation in critical domains (e.g., justice, healthcare) to support the work of decision-makers (e.g., judges, medical doctors). In the medical domain, there are already proposals of deep learning algorithms that claim to achieve or surpass human-like performance [9]. However, the black box behaviour of these models does not help decision-makers to have a clear understanding of their inner-functioning, thus preventing them from diagnosing errors and potential biases or deciding when and how much to rely on these models [533]. This lack of transparency represents an obstacle to the deployment and the adoption of these algorithms in the medical context since it jeopardises the trust of their potential users [33]. There has been a huge effort into the development of post-model strategies to explain the behaviour of black box models; however, the outputs of these algorithms are prone to subjective evaluation (i.e., different users will have different interpretations) [133] and may be misleading [118]. Moreover, recent works have shown that it is possible to fool these post-model methods in the sense that an arbitrary attacker can manipulate these *explanations* for their benefit

(e.g., hide discriminatory biases of a black box model) [533, 534]. Hence, we agree that just being able to obtain explanations is not enough and that we rather need to take into account at the development stage that these methods must respect specific constraints that give them the capability of generating human-understandable explanations and make decisions based on such premises [11].

To facilitate trust (and increase transparency) in artificial intelligence (AI) algorithms, it is important to ensure *a priori* that these models are interpretable. This can only be achieved if we understand first how decisions are made in the clinical context. For instance, to produce a decision, clinicians may use examples that have similarities with the sample under analysis. Contrarily, using examples that represent the opposite to understand the decision (i.e., counterfactual explanation) [535] is also common. This type of prior knowledge can be integrated during the training of the models (e.g., imposing model sparsity through regularisation) or before, by building models based on cases [119]. On the other hand, it is important to understand what the algorithms are already learning and evaluate the quality of such explanations (e.g., understand if the algorithms extract relevant features for the clinical context). A different dimension of the application of AI in sensitive domains such as healthcare is the development of ethical and fair algorithms [536]. Therefore, it is fundamental that we take into account the intrinsic variability of the data sets to produce models that can be agnostic to intrinsic biases [537]. This strategy is supported by the new European Union's General Data Protection Regulation (EU-GDPR) [538], which advises that these algorithms should be able to explain their decisions for the sake of transparency. Additionally, transparent algorithms might appropriately enhance the trust of clinical professionals in future AI systems.

In this Chapter, we present three different case studies that depict how inherently interpretable algorithms work and show their applications in medical tasks. We show experimental work with B-cos networks for medical image classification and retrieval and dissect the potential of prototypical learning for counterfactual approaches. Beyond the Introduction, this Chapter is organised as follows: Section 6.2 presents a historical overview and the fundamental concepts regarding inherently interpretable machine learning, Section 6.3 provides a comprehensive review of the prototypical machine learning and counterfactual approaches, Sections 6.4 – 6.6 present the experimental work developed to address the research questions related to this topic, and Section 4.7 concludes the Chapter, highlighting the main contributions and anticipating future work directions.

6.2 Background

The development of intrinsically interpretable machine learning algorithms has gained attention in recent years mainly due to the concerns of the stakeholders of the fields where high-stakes decisions take part. Moghaddass et al. [539] proposed a model based on the self-controlled case series method [540, 541] to estimate the effects of transient drug exposures on a collection of health outcomes (insulin treatments for diabetes), where the effects of all drugs and all outcomes were estimated simultaneously. The authors also studied the use of latent factors [542] in their approach

as a strategy to reduce parameters and increase interpretability by illuminating latent classes of drugs and outcomes. Lakkaraju and Rudin [543] used a Markov decision process approach to learn a decision list to generate cost-effective, actionable and interpretable treatment recommendations for asthma patients. A different line of work focuses on learning and optimising scoring systems. Scoring systems are linear classification models that only require users to add, subtract and multiply a few small numbers to make a prediction; however, they are challenging to learn from data because they need to be accurate and sparse, have co-prime integer coefficients, and satisfy multiple operational constraints [19]. Following this path, Ustun and Rudin [544] presented a method for creating data-driven scoring systems called Supersparse Linear Integer Model (SLIM) that can seamlessly incorporate a wide range of operational constraints related to accuracy and sparsity and can produce highly tailored models without parameter tuning. The authors implemented this method in a medical context (collaboration with the Massachusetts General Hospital Sleep Laboratory), where SLIM was used to create a scoring system for screening sleep apnea [545]. In an analogous setting, Zhou et al. [546] developed and evaluated a risk score based on the *Lasso-Cox* model [547] for critical care patients with viral or unspecified pneumonia and who are eligible for extracorporeal membranous oxygenation. Under the scope of case-based explanations, Barnett et al. [22] proposed a deep learning architecture based on [21] that highlights parts of the image, explains that it considers these parts of images similar to prototypical cases it has seen before, and provides a score for the probability of the specific diagnosis for this image; this framework was applied to the classification of mass lesions in digital mammography. Koh et al. [548] revisited the *concept bottleneck* definition and proposed a deep learning architecture that learns to predict the concepts that are given at training time and then uses these concepts to predict a label; hence, we can intervene on these concept bottleneck models by editing their predicted concept values and propagating these changes to the final prediction. The authors tested this methodology on a database that contained radiological and clinical data on patients at risk of knee osteoarthritis. Newer approaches take into account causality, ethics and fairness. Wang and Rudin [27] introduced causal rule sets that learn a small set of short decision rules to capture a subgroup where the average treatment effect is elevated, and Gan et al. [28] conducted a study on the benefits of incorporating a large confounded observational data set alongside a small deconfounded observational data set when estimating the average treatment effect. In [549], Afnan et al. perform an essay on what would be an ethical implementation of AI to select embryos in *in vitro fertilisation* and point to interpretability as a requirement in this process. Lipton et al. [550] studied the relation between *treatment disparity* (if the algorithm treats members of protected subgroups differently) and *impact disparity* (if the outcomes differ across subgroups, even unintentionally) in machine learning algorithms.

6.3 Literature Review

This Section introduces the basics of *counterfactual reasoning* and *prototype-based deep learning*, which are fundamental to understanding the motivation, methodologies, and results presented in Sections 6.4–6.6. We describe and summarise the papers in the literature in an almost exhaustive chronological order to give the reader a sense of historical context and evolution.

6.3.1 Counterfactual Reasoning

Several methods attempt to generate human-understandable explanations. In this work, we followed the line that assumes that human explanations are contrastive in the sense that humans often try to explain the phenomena that would *not* lead to an event, rather than trying to deconstruct all the causes behind it [551]. A particular type of such explanation includes *counterfactual explanations*. Counterfactual explanations have been a subject of study in several fields (e.g., philosophy, psychology, social sciences) and are usually a type of explanation that establishes a bridge between what could have happened if the input to a model had changed in a particular way [535]. Verma et al. [535] provide an exhaustive review of counterfactual reasoning in machine learning. However, they only focused on tabular data. On the other hand, they also point the readers to suggested references on other modalities (i.e., image data, text data, speech data, time-series, graph data for graph neural networks, agent action, recommendation systems, functional data): we base this review on counterfactual reasoning on that suggestion list and focus on the imaging modality.

Hendricks et al. [552] explored the interface of image and natural language domains by introducing *textual counterfactual explanations* for image classification. In their work, the authors proposed a method to generate counterfactual explanations by inspecting which features are missing in a given image that may contribute to a different classification output if they were present and generating text explanations that aim to explain why the image does not belong to a counterfactual class. Goyal et al. [553] showcased a technique that produces visual counterfactual explanations by comparing a query image against an image from a different label and identifying which regions of the last could be changed so that the model classifies it as the label of the query image. The authors also performed human studies on the task of bird classification, concluding that these examples of counterfactual explanations seem to guide humans and contribute to their performance improvements on that classification task. Liu et al. [554] proposed an introspection technique for deep neural networks that uses generative models with an actionable latent feature editing mechanism [555] to generate counterfactual explanations, taking into account decision boundaries between classes and their relationship to the input data by providing actionable feedback and while producing counterfactual prototypes and explanations based on criticisms.

Balasubramanian et al. [556] proposed a methodology to generate counterfactual explanations from a latent representation (Latent-CF) that consists of training an auto-encoder to achieve meaningful latent representations of the data and perturbing these representations to generate the

desired counterfactual. Akula et al. [557] explored the basis of counterfactual explanations and concept activation vectors [25] to propose the *conceptual and counterfactual explanations* (CoCoX) framework that relies on *fault-lines* (i.e., semantic features that humans use when they visualise an alternative to a model prediction) to explain model decisions. Wang and Vasconcelos [558] proposed to generate *self-aware discriminant counterfactual explanations* (SCOUT), resulting from a combination of self-awareness mechanism (i.e., an ability to measure its limitations) and attributive explanations (i.e., explanations based on saliency maps), which aim to quantify the confidence of the predictions of a model, and connecting the attributive to counterfactual explanations to highlight the image regions that are informative of the query image but not of the comparing one. Feghahati et al. [559] assume the existence of a generative model on the input space (but not necessarily on the input dataset used to train the model) capable of creating synthetic samples similar to the inputs and then use the back-propagation algorithm to understand how to modify the inputs to change the classification result. Smith and Ramamoorthy [560] studied the generation of counterfactual explanations in robot control. They implemented a generative model (cGen) trained to produce minimal and realistic counterfactual perturbations to a known input. The authors stated that such perturbations can be used in the design process of image-based robust control as a means to determine the ability of the controller to return to a working regime by modifications in the input space rather than by adaptation. Zhao [561] also benefited from the current knowledge on generative models and proposed the *fast real-time counterfactual explanation* (FRACE) framework, which uses a generator to generate a perturbation that causes the query image to be classified as counterfactual. According to the author, this method is faster than other methods in generating counterfactual explanations because it is optimisation-free during inference time.

Ali et al. [562] compared the softmax-based counterfactual explanations with their Bayesian model uncertainty counterfactual explanations and suggested that the latter produces more concise and human-understandable counterfactuals. Alipour et al. [563] tackled the generation of counterfactual explanations for visual question answering (VQA) models and compared retrieval-based against generative adversarial networks (GANs) based methods and concluded that the latter could also be effective at improving the mental model of humans. Elliot et al. [564] proposed a regularisation method for the perceptual loss [565] that tackles adversarial perturbations and aims to produce perturbations that remain hidden to the human eye while being semantically meaningful (i.e., they try to highlight objects of interest in the images and leave the background unchanged), and making this connection to the domain of counterfactual explanations in image space. Gulshad and Smeulders [566] also aimed to produce counterfactual visual explanations using factual and counterfactual attributes and examples. The authors also studied how the predicted attributes change after a given perturbation (intended to generate counterfactual explanations). Hvilshøj et al. [567] focused on generating *efficient counterfactuals from invertible neural networks* (ECINN) that explore the generative abilities of invertible neural networks [568, 569, 570, 571] for image classification to generate counterfactual examples efficiently. Hölting et al. [572] proposed the *deep deterministic uncertainty-based counterfactual explanations* (DeDUCE) framework, taking

advantage of the concept of *deep deterministic uncertainty* [573] measurement of epistemic uncertainty, as a valid target to generate counterfactuals that are similar to the training data [574, 575]. Kenny and Keane [576] addressed counterfactual and semi-counterfactual (i.e., a perturbation is applied to the input such that the prediction of the model does not change) examples with their method *plausible exceptionality-based contrastive explanations* (PIECE), which automatically models the distributions of latent features to detect “exceptional features” in a query instance and modifies this feature set to be “normal” in the process of explanation generation. Labaien et al. [577] reflected on the impracticability of the real-time feedback of counterfactual explanations since this process generally consists of solving an optimisation problem for each input requiring an explanation. As such, the authors presented two methods: one called *deep guided counterfactual explanations* (DGCEx), which consists of training an auto-encoder plugged into a classifier, to generate straightforward counterfactual explanations; another called *distribution aware deep guided counterfactual explanations* (DA-DGCEx), built on top of DGCEx and aiming to improve the counterfactual generation process by adding a term in the loss function that penalises counterfactual instances falling outside the data distribution. Li et al. [578] relaxed the problem of generating counterfactuals (through input perturbation) by assuming no explicit causal relations. Their method, *directional counterfactual generation* (DCFG), introduced the concept of *directional counterfactual* that combines *latent subspace disentanglement* and counterfactual explanation to provide progressive explanations of the classifiers taken by unintelligible, black-box models in the low-rank sub-spaces, and uses noise injection and principal component analysis to identify the relevant low-rank sub-spaces. Popescu et al. [579] tackled the generation of counterfactual explanations from the perspective of input perturbation. They consider the case where the perturbation is done using generative in-filling methods [580] with the Knockoffs framework [581, 582] to achieve a causal sense while still maintaining the compactness of the explanations. Rodriguez et al. [583] explored how to interpret machine learning models by identifying sets of attributes with high impact on the output and proposed the *diverse valuable explanations* (DiVE) method that leverages the Fisher information matrix of its latent space to focus its search on the less influential factors of variation of the model to discover spurious correlations and produce more informative multiple counterfactual explanations. Thiagarajan et al. [584] tackled a different use case. In their paper, they assumed they have access only to the trained deep classifier and not the training data and proposed a method, *deep inversion for synthesising counterfactuals* (DISC), inspired by deep inversion [585], that uses strong image priors and incorporates a manifold consistency objective to achieve visually meaningful explanations. Van Looveren et al. [586] studied the generation of counterfactual explanations with a conditional generative model that allows the generation of counterfactual instances with a single forward pass. The authors also state that their method is flexible to the type of generative model and the task so that it is compatible with multiple data modalities, architectures, or learning tasks. White et al. [587] extended the *counterfactual local explanations via regression* (CLEAR) [588] into *CLEAR Image*, that uses generative models to produce counterfactual and contrastive explanations while identifying what the authors call as *causal over-determination* (i.e., the multiple sufficient

causes for image classification). Yang et al. [589] also apply generative models to generate counterfactuals specifically for raw data instances. They proposed the *attribute-informed perturbation* (AIP) framework, which uses gradient-based optimisation and iteratively updates the parameters of the generative models in the attribute-informed latent space, according to the counterfactual loss concerning the desired prediction label. Mothilal et al. [590] addressed two properties that proper counterfactual explanations should address: *feasibility* (of the counterfactual actions given user context and constraints) and *diversity* (among the counterfactuals presented). Considering the latter, they proposed a framework (DiCE) that generates and evaluates a diverse set of counterfactual explanations based on determinantal point processes [591]. Following and extending the latter, Hsieh et al. [592] proposed a methodology designed for process prediction (in their case, event logs) that supports deriving milestone-aware counterfactual explanations at key intermediate steps along process execution to promote interpretability.

Akula et al. [593] developed the counterfactual explanations with the theory-of-mind [594] (CX-ToM) framework that models explanations as a dialogue between the machine and the human (i.e., modelling the human's intention, the machine's mind as inferred by the human, and the human's mind as inferred by the machine). Laird et al. [595] designed a framework that aims to explain model mistakes in terms of high-level concepts that may be easy for humans to understand. Using the hypothesis that generating image counterfactuals is non-trivial due to the underlying causal structure on the features of the image, Dash et al. [596] developed a method (ImageCFGen) that incorporates a structural causal model in a GAN-based framework to generate counterfactuals that should respect causal relationships between the attributes of an image. Filandrianos et al. [597] approached counterfactual explanations from the perspective of the concepts that characterise the data while aiming to find the minimum conceptual edits we need to apply to make the model alter its predicted class. Moreover, the authors also propose a methodology to produce *global* counterfactual explanations for a specific label by accumulating multiple counterfactual explanations from different regions of the dataset. Ghandeharioun et al. [598] also studied the relation between concepts and counterfactual explanations and designed a methodology (DISSECT) that uses generative models and concept disentanglement to generate what the authors call *concept traversals* (i.e., sequences of generated examples with increasing levels of the presence of concepts that may influence the output of a classifier), which may help users analysing some internal mechanisms of the model related to the learning of concepts. Guyomard et al. [599] investigated the application of an auto-encoder architecture to generate class prototypes, used to obtain counterfactual examples close to the distribution of data and the distribution of the target class. Jacob et al. [600] developed the *steering counterfactual explanations with semantics* (STEEX) model that focuses on region-based counterfactual explanations, where the semantic changes are applied to specific regions of the image, but the general scene layout is preserved. Jeanneret et al. [601] leveraged the capabilities of the recently proposed diffusion models [602, 603] to generate counterfactual examples, *diffusion models for counterfactual explanations* (DiME), and developed a different evaluation metric, *correlation difference*, to assess misleading correlations on counterfactual explanations. Khorram et al. [604] worked

on the latent space of generative models and developed the *cycle-consistent counterfactuals by latent transformations* (C3LT) methodology that aims to generate visual counterfactuals by manipulating that latent space, and uses *cycle consistency* [283] between the query and the counterfactual latent representation to find appropriate solutions. Mertes et al. [605] also studied the applicability of generative adversarial models to generate visual counterfactual explanations and introduced a framework based on adversarial image-to-image translation techniques (GANterfactual). The authors also reported user studies using a medical use case and compared their counterfactual explanations against post-hoc xAI techniques based on saliency maps. Oh et al. [606] also worked on the line of post-hoc techniques. They proposed the *born identity network* (BIN), consisting of a *counterfactual map generator*, which is a generative model that can synthesise a counterfactual map conditioned on an arbitrary target label, and the *target attribution network*, that supports the counterfactual map generator by conditioning a target label that the last will use to generate synthetic counterfactual representation maps. Oh et al. [607] later explored counterfactual reasoning to the task of Alzheimer’s disease diagnosis. They proposed the *learn-explain-reinforce* (LEAR) framework, which produces counterfactual maps through explanation-guided model refinement. Their methodology uses a classification model for the disease diagnosis, a generative model to produce the post-hoc counterfactuals, and a combination of both to improve the generalisation capacity of the classification model (i.e., based on the explanations). Sanchez et al. [608] tackled counterfactual estimation from the perspective of causality [609, 610]. Assuming that quantifying the causal effect of interventions for high-dimensional data with neural networks is challenging, the authors developed a deep structural causal (Diff-SCM) model inspired by generative energy-based models [611, 612]: inference is performed by iteratively sampling gradients of the marginal and conditional distributions entailed by the causal model; and counterfactual estimation is achieved by firstly inferring latent variables with deterministic forward diffusion, then intervening on a reverse diffusion process using the gradients of an anti-causal predictor concerning the input. Vandenhende et al. [613] developed a method to produce counterfactuals consisting of a framework to generate both discriminative and semantically consistent counterfactuals by constraining the identified class-specific image regions that alter a model’s decision to allude to the same semantic parts and expanding the search space by using multiple distractor images from a given class leading to more discriminative explanations with fewer regions to replace, but constraining the processing to only the most similar ones. Vermeire et al. [614] also studied a model-agnostic method that only requires access to the model (and not the training data) and can generate instance-level explanations for image classification. Their algorithm, an adapted version of the *search for evidence counterfactual* (SDEC) [615], relies on *evidence counterfactuals* (i.e., an irreducible set of features that, in case they were not present, would alter the classification of a given data point) to generate the counterfactual explanations. Yang et al. [616] worked on a gradient-based counterfactual explanation method, *discriminative gradients* (DiscGrad), that tries to derive the explainable discriminative attributes by considering not only the predicted class but also the counterfactual classes.

Recently, the field of counterfactual reasoning has been delving into the study of causal-inspired

mechanisms. Wang et al. [617] developed a framework, MedSeqCF, that represents electronic health records as medical events and uses a text style-transfer technique to generate counterfactuals. Singla et al. [618] proposed a GAN-inspired algorithm that iteratively perturbs a query image to produce a counterfactual (e.g., transform the positive class into the negative class), preserving several anatomical and spatial details of these images, by adding a specialised reconstruction loss during training. Pombo et al. [619] designed a 3D generative model for volumetric, unpaired, image-to-image translation, applied to brain imaging for tailored data augmentation; this model learns to generate counterfactual volumetric brain imaging. From a fundamental perspective, Nasr-Esfahany et al. [620] studied counterfactual identifiability in causal models with bijective generation mechanisms and concluded that these models enable efficient counterfactuals with the aid of deep-based conditional generative models. Hundogan et al. [621] focused on predictive process analytics (i.e., predicting future states of a system) and used evolutionary methods to generate viable counterfactual sequences. Vlontzos et al. [622] studied the application of deep twin networks to learn counterfactual ordering for causal models with categorical variables, which should show desirable properties that causal mechanisms should have. Dantas et al. [623] developed a generative adversarial counterfactual model for satellite image time series in a multi-class classification task that does not require a prior assumption on the targeted class for a given counterfactual explanation. Ying et al. [624] studied counterfactual data augmentation in the training of fair recommendation systems, where the goal of their system is to ensure that minority users are not jeopardised by mainstream users.

For better comprehension, Table 6.1 presents an overview of the counterfactual generation approaches through time.

6.3.2 Prototype-based Deep Learning

The ideas behind *prototypical* learning are not new. However, the concept of *prototypes* changes across domains and applications. For instance, Takigawa et al. [625] proposed a framework that builds a set cover with *convex sets*, each containing training points of a specific class (only). The author considers these sets as the *prototypes* of those classes. Bien and Tibshirani [626] refer that prototype methods aim to create a minimal subset of samples that may serve as *representative* samples chosen from the dataset while increasing its interpretative value. They also proposed three properties a good prototype set should satisfy: covering as many training points of a given class as possible (and as few training points of the others) and being sparse (i.e., using as few data points as possible). Snell et al. [627] discussed the application of *prototypical networks* for the problem of few-shot classification, where the goal was the learning of a metric space that allowed classification using the distances of prototype representations of each class. A few years later, Pan et al. [628] revisited the previous and applied prototypical learning to domain adaptation, Yang et al. [629] took the first steps towards the application of prototypical learning for multi-label classification, and Kim et al. [630] explored federated semi-supervised learning with prototypes.

Table 6.1: Approaches to counterfactual generation through time.

Author(s)	Methodology	Claims / Contributions
Hendricks et al. [552]	Generative	Inspect the missing features in the image
Goyal et al. [553]	Generative	Inspect the features that need to change
Liu et al. [554]	Generative	Inspect the decision boundaries of classes with an attention mechanism
Balasubramanian et al. [556]	Generative	Intervene on latent representations
Akula et al. [557]	Generative	Focus on the semantic features
Wang and Vasconcelos [558]	Generative	Saliency maps and counterfactuals to highlight informative image regions
Feghahati et al. [559]	Generative	Create synthetic samples that illustrate the counterfactual class
Smith and Ramamoorthy [560]	Generative	Produce minimal and realistic counterfactual perturbations
Zhao [561]	Generative	Generator that produces a perturbation
Oh et al. [606]	Generative	Counterfactual generation driven by a conditional GAN
Mothilal et al. [590]	Generative	Counterfactual generation driven by determinantal point processes
Hsieh et al. [592]	Generative	Counterfactual generation for process prediction with intermediate key steps
Ali et al. [562]	Generative	Counterfactual generation driven by Bayesian model uncertainty
Alipour et al. [563]	Generative	GAN based generation of counterfactual explanations for VQA models
Elliot et al. [564]	Generative	Regularisation method that promotes adversarial perturbations
Gulshad and Smeulders [566]	Generative	Factual and counterfactual attributes to produce counterfactual visual explanations
Hvilshøj et al. [567]	Generative	Counterfactual generation driven by invertible neural networks
Höltgen et al. [572]	Generative	Counterfactual generation driven by deep deterministic uncertainty
Kenny and Keane [576]	Generative	Counterfactual generation driven by finding the right set of features to modify
Labaien et al. [577]	Generative	Counterfactual generation driven by an auto-encoder and a classifier
Li et al. [578]	Generative	Latent subspace disentanglement and counterfactual explanations in low-rank sub-spaces
Popescu et al. [579]	Generative	Counterfactual generation driven by perturbation
Rodriguez et al. [583]	Generative	Counterfactual generation driven by the Fisher information matrix
Thiagarajan et al. [584]	Generative	Counterfactual generation driven by deep inversion
Van Looveren et al. [586]	Generative	Counterfactual generation driven by a GAN with a single forward pass
White et al. [587]	Generative	Contrastive and counterfactual explanations generation driven by over-determination
Yang et al. [589]	Generative	Counterfactual generation with an attribute-informed latent space
Akula et al. [593]	Generative	Model explanations as a dialogue between the machine and the human
Laird et al. [595]	Retrieval	High-level concepts to explain models mistakes in an human-understandable manner
Dash et al. [596]	Generative	Counterfactual generation with a structural causal model incorporated in a GAN
Filandrianos et al. [597]	Retrieval	Global counterfactual retrieval by finding the minimum conceptual edits to the inputs
Ghandeharioun et al. [598]	Generative	Counterfactual generation driven by GANs and concept disentanglement
Guyomard et al. [599]	Generative	Auto-encoder based generation of class prototypes to obtain counterfactual examples
Jacob et al. [600]	Generative	Counterfactual generation driven by changes in specific image regions
Jeanneret et al. [601]	Generative	Counterfactual generation driven by diffusion models
Khorram et al. [604]	Generative	Counterfactual generation driven by latent space manipulation and cycle consistency
Mertes et al. [605]	Generative	Counterfactual generation driven by GANs
Oh et al. [607]	Generative	Counterfactual generation driven by a combination of a GAN and a classifier
Sanchez et al. [608]	Generative	Counterfactual generation driven by a deep structural causal model based on diffusion
Vandenhende et al. [613]	Generative	Counterfactual generation by only replacing semantically matching image regions
Vermeire et al. [614]	Generative	Counterfactual generation driven by an irreducible feature set, essential for the classification
Yang et al. [616]	Generative	Counterfactual generation driven by discriminative gradients
Wang et al. [617]	Generative	Counterfactual generation driven by a text-style transfer approach
Singla et al. [618]	Generative	Counterfactual generation driven by a GAN that applies perturbations to the query image
Pombo et al. [619]	Generative	Counterfactual generation driven by a GAN based model for image-to-image translation
Nasr-Esfahany et al. [620]	Generative	Counterfactual identifiability in causal models with bijective generation mechanisms
Hundogan et al. [621]	Generative	Counterfactual generation driven by evolutionary methods
Vlontzos et al. [622]	Generative	Deep twin networks for the learning of counterfactual ordering for causal model
Dantas et al. [623]	Generative	Counterfactual generation driven by a GAN based model
Ying et al. [624]	Generative	Counterfactual data augmentation for fair recommendation systems

Li et al. [631] belong to the pioneers that proposed the creation of a *prototype layer* in deep learning networks. The main idea behind it is that each unit of this layer should store a weight vector that resembles an encoded training input (i.e., a *prototype*); the learning of these prototypes happens during the training of the network. This work paved the way for the ideas described in [5], followed by this thesis. Proposed by Chen et al. [5], the *prototypical part network* (ProtoPNet) aimed to bring interpretability into deep learning architectures through the addition of a *prototype layer* that learns *prototypical parts* of the images and integrates this information in the classification process. In a nutshell, ProtoPNet receives an image and tries to identify the parts of that image that are more similar to some prototypical part of some class, generating a vector of similarity scores. Then, ProtoPNet uses a weighted combination of these scores in the classification layer to output the final prediction (see Figure 6.1). The authors also suggest that this model is inherently interpretable since they consider this process of making predictions a transparent approach. On the other hand, their results suggest that even with the integration of this intermediate prototype layer, the models can still keep their predictive performance.

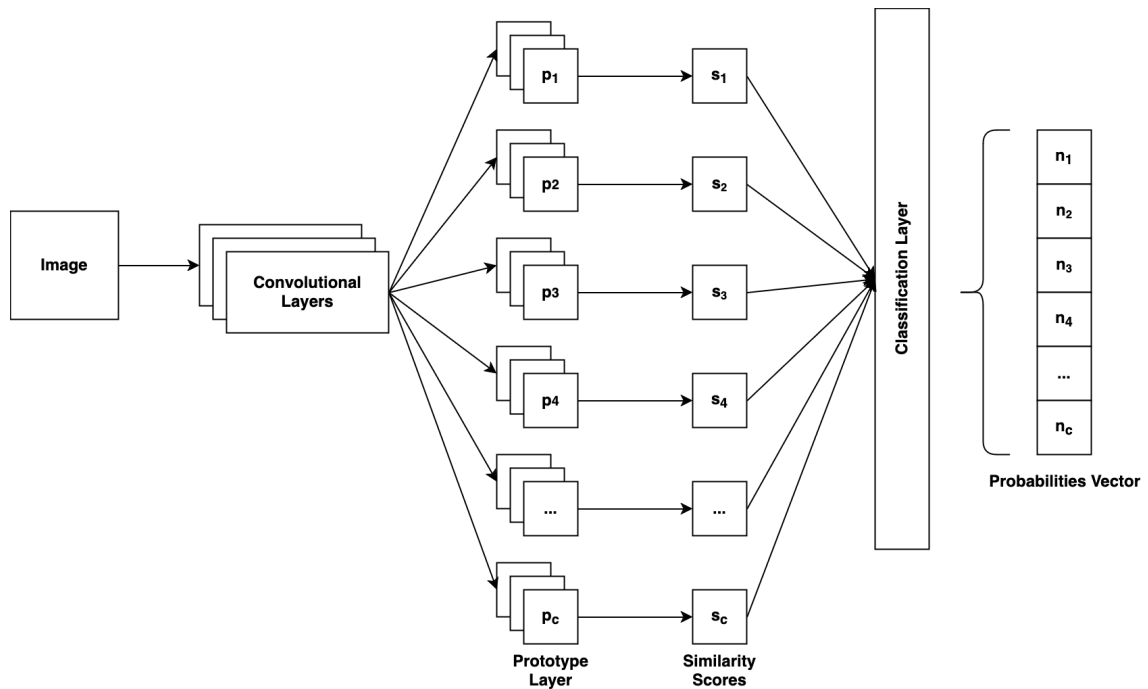


Figure 6.1: Overview of ProtoPNet architecture, according to [5]. The model receives an image and tries to identify the parts of that image that are more similar to some prototypical part of some class, generating a vector of similarity scores. Then, it uses a weighted combination of these scores in the classification layer to output the final prediction. All the descendant models derive from this approach.

After the release of the original ProtoPNet model, the works that followed merely presented small increments, integrations, or different applications. Hence, to avoid redundancy of ideas and to facilitate the contextualisation of the reader, we agglomerate these proposals in Table 6.2, identifying their authors (and references), models, and contributions (or claims).

Table 6.2: ProtoPNet-based models that resulted from small increments, integrations, or different applications.

Author(s)	Model(s)	Claims / Contributions
Xu et al. [632]	APN	Regression and decorrelation of attributes for zero-shot learning
Xu et al. [633]	APN	Regression and decorrelation of attributes for any-shot learning
Arik et al. [634]	ProtoAttend	Attention mechanism to facilitate prototype learning
Kim et al. [635]	XProtoNet	Clinical use case (chest radiography)
Wang et al. [636]	TesNet	Prototypical learning treated as concept learning
Rymarczyk et al. [637]	ProtoPShare	Prototype pruning strategy to merge similar semantic prototypes
Singh and Yow [638]	NP-ProtoPNet	Negative reasoning process
Singh and Yow [639]	Gen-ProtoPNet	Prototypical parts of any spacial dimensions
Singh and Yow [640]	Ps-ProtoPNet	Preservation of the impact of the negative reasoning
Singh [641]	Quasi-ProtoPNet	Only positive reasoning
Stefenon et al. [642]	Semi-ProtoPNet	Different prototype encoding, positive and negative reasoning
Barnett et al. [643]	IAIA-BL	Clinical use case (mammograms)
Carlioni et al. [644]	-	Clinical use case (mammograms)

Contemporaneously, Hase et al. [645] proposed the *hierarchical prototype network* (HPnet), that aims for prototype-based interpretable classification of images at multiple levels of taxonomical specificity even when data points (i.e., images) come from classes that may fall under a broader category in the known taxonomy.

Interestingly, in parallel, several authors began exploring these concepts in applications of few-shot learning [646, 647, 648]. Oreshkin et al. [649] explored metric scaling and task conditioning to improve few-shot performance. Sun et al. [650] proposed a hierarchical attention prototypical network for few-shot text classification. Allen et al. [651] studied infinite mixture modelling [652] and proposed infinite mixture prototypes to represent simple and complex data distributions for few-shot learning adaptively. Yoon et al. [653] devoted efforts to neural networks with task-adaptive projection that employ a linear transformer to project the feature space into another space where classification happens. Ye et al. [654] designed a transformer-based model to adapt instance embeddings to the target classification with a set-to-set function for few-shot learning; Zhen et al. [655] introduced a long-term memory module (i.e., variational semantic memory) into meta-learning for few-shot learning, applying it as an external memory for the probabilistic modelling of prototypes in a hierarchical Bayesian framework.

Recently, Donnelly et al. [656] introduced the *deformable prototypical part network* (Deformable ProtoPNet), in which each prototype is composed of several prototypical parts that adaptively change their relative spatial positions depending on the input image. Naturally, there are already efforts to combine prototype-based architectures and vision transformers. Xue et al. [657] proposed the *prototypical part transformer* (ProtoPFormer), that combines a prototype-based architecture with vision transformers for image classification, with global and local prototypes. Later, Vidal et al. [658] presented a transformer-based multi-prototype network with local and global branches. The prototypes learned by the local branch focus on specific image tokens. The global branch guides the local branch, masking the background to focus the attention on foreground

features. We can observe some influence of these concepts in geometric deep learning. Ragno et al. [659] investigated the use of ProtoPNet and TesNet in the graph domain, developed a method to extract ProtoPNet and TesNet explanations on graph neural networks, and studied how to obtain global and local explanations for the trained models. Wang et al. [660] used prototypical learning strategies to optimise the domain generalisation properties of convolutional neural networks with a siamese architecture that learns a domain invariant subspace, which reduces the inter-domain discrepancy of distributions, and a module of condition recognition that applies a prototypical layer, optimised to learn the prototypes of each class. Fu et al. [661] explored few-shot image classification with their *bidirectional matching prototypical network*, which uses support and query images to generate more accurate prototypes and force the model to learn better generalisation features. Besides, they added a *lightweight coordinate attention feature extractor* to help the model locate the region of interest more accurately. Gogoi et al. [662] proposed the *adaptive prototypical networks*, which uses an encoder for model classification using a support set of images and then fine-tune its weights in the remaining data samples. The authors argue that this strategy is a mechanism to push the embeddings and, thus, the prototypes away from each other, even in the presence of similar-looking classes.

Lately, Wang et al. [663] revisited the fundamentals of support vector machines (SVMs), making an analogy between the prototypes of ProtoPNet and the support vectors in an SVM. On the other hand, they stated that trivial prototypes are far from the classification boundary and support vectors are close to this boundary, meaning that ProtoPNet models may suffer from sub-optimal classification accuracy. To solve this issue, they developed the ST-ProtoPNet model to exploit both support (i.e., hard-to-learn) and trivial (i.e., easy-to-learn) prototypes. The model combines these two sets of prototypes to promote more interpretability and better classification accuracy. In the medical domain, Gallée et al. [664] combined prototypical learning, capsule networks [665] and privileged information [666, 667] for lung nodule classification with improved performance. Ma et al. [668] approached the difficulty of identifying the meaning behind single-image prototypes and modified the prototype geometry to enable the visualisations of prototypical concepts from multiple training images. Lyu and Wang [669] revisited the task of few-shot learning. They proposed *compositional prototypical networks* to learn transferable prototypes (for human-annotated attributes). These models are reusable (e.g., to construct other prototypes for novel classes). An et al. [670] addressed the task of generalised category discovery (i.e., recognise both known and novel categories) with an architecture composed of two components: the decoupled prototypical network, which aims to decouple known and novel prototype categories to better capture high-level semantics; and the semantic-aware prototypical learning to capture better semantic relationships between instances and prototypes.

Recently, several approaches have tried to combine different neural network architectures with ProtoPNet to understand the impact of the predictive power of models and their explanations. Lim et al. [671] studied the impact of including self-supervised training in the training of a ProtoPNet and reported that, if properly optimised, this strategy benefits the predictive performance of these

models. Fernandes et al. [672] combined the Faster-RCNN [673] and ProtoPNet and developed the interpretable-Faster-RCNN (iFaster-RCNN), an end-to-end object detection capable of explaining its predictions via prototypical learning. Yu et al. [674] tackled the task of prototype-based meta-learning, using a diffusion model (ProtoDiff) during the meta-training phase to generate efficient class representations.

As expected, while there is work to improve the quality of the prediction of the models, there is also work devoted to assessing the limitations and evaluating the quality of their explanations. Hoffmann et al. [675] analysed the latent space of a ProtoPNet and performed experiments to highlight sources of interpretability corruption. Huang et al. [676] went further and proposed two evaluation metrics (i.e., consistency and stability) to assess the explanation consistency across images and the explanation robustness against perturbations. In their work, they also proposed a prototype network with shallow-deep feature alignment (SDFA) and score aggregation (SA) modules to improve the interpretability of prototypes. Bontempelli et al. [677] proposed a human-in-the-loop framework where a human supervisor can give feedback to a ProtoPNet regarding what prototypes should be kept or forgotten.

From this background knowledge of counterfactual examples, we can deduce that there are two approaches to counterfactuals. Firstly, we can retrieve the counterfactual examples using generative methods. In this approach, the paradigm is to intervene in the inputs (directly or via a latent representation) and obtain an adversarial sample (with a different label), which will be counterfactual. On the other hand, we can use retrieval methods, where, given an input (and its corresponding label), we search for the nearest sample in the database with a different label and use it as the counterfactual sample. Nevertheless, retrieval-based methods seem to be less explored. In this sense, our work provides a timely opportunity to investigate a retrieval-based approach (supported by prototypical learning) and understand whether it is a viable alternative to the current paradigm based on generative methods.

6.4 B-cos Networks for Medical Image Classification

The results of this work have been disseminated in the form of a paper at an international conference: I. Rio-Torto, T. Gonçalves, J. S. Cardoso, and L. F. Teixeira, “On the Suitability of B-cos Networks for the Medical Domain,” in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2024 [47]. The code related to the implementation of this work is publicly available in a GitHub repository¹.

6.4.1 Description and Contributions

Although there are several approaches to explaining deep learning models, searching for a gold standard is still one of the open challenges in explainable artificial intelligence (xAI). There

¹<https://github.com/icrto/medical-bcos>

are essentially two lines of thought: create models inherently interpretable by design (i.e., *in-model*) [11], or use explanatory frameworks from the algorithms (i.e., *post-model*) [678]. In medical AI, the community tends to prefer in-model approaches [643], as the consequences of wrong clinical decisions might significantly impact patients' lives. In addition, the medical field is challenging and often highly subjective [679], so clinical decision support tools should be as transparent and auditable as possible. On the other hand, *post-model* methods [7] still are the *de facto* standard for interpretability in the general domain and, consequently, in the medical domain [680], mainly due to their model-agnostic nature and not requiring retraining of models.

Although it is still unclear which methods are better for specific domains (e.g., healthcare or medical), the research community is doing their best to tackle these questions. A recent and popular approach is the Quantus evaluation toolkit, developed by Hedström et al. [681], which contains a series of tests to assess the quality of saliency maps. Aiming to contribute to the fields of intrinsic interpretability and assessment of the quality of explanations, Böhle et al. [682] proposed the B-cos transform as a replacement of all linear transforms of a deep neural network, to induce an alignment between the network's weights and task-relevant input patterns. The intuition is that a sequence (network) of such transformations can be reduced to a single linear transformation, faithfully summarising all the model computations. That said, we can consider that a B-cos network constitutes an inherently interpretable model. In this work, we extend their experiments and investigate the application of B-cos networks to the medical domain. We consider that our contributions are as follows:

1. Application of the B-cos network to several medical use cases (i.e., diabetic retinopathy, cervical cytology, chest X-rays, and skin lesions) and multiple tasks (i.e., multi-class and multi-label classification), thus expanding the proposal of the original paper (i.e., binary classification on natural images);
2. Evaluation of an extended set of properties of the generated visual explanations using an independent framework [681];
3. Integration of the B-cos layer into a state-of-the-art custom model for the medical domain [683];
4. Investigation of the effect of different pre-training strategies on the performance of B-cos networks.

6.4.2 Methodology

This Section presents the details of the implementation of this work.

6.4.2.1 Data

Given the diversity of use cases and tasks in the medical domain, we tested B-cos networks on different tasks (i.e., multi-class and multi-label classification) and applications using different

datasets (see Chapter 3): DTU/Herlev Pap Smear Database (cervical cytology), VinDr-CXR (chest X-rays), APTOS2019 (diabetic retinopathy), and ISIC2018 (skin lesions).

6.4.2.2 B-cos Networks

In a B-cos network [682], we replace all linear transforms of a deep neural network with the B-cos transform, hoping that this new module will optimise the weights to align with task-relevant input patterns and inherently produce explanatory saliency maps given the completely linear nature of the model.

6.4.2.3 Explanation Quality Assessment Metrics

Recognising the importance of a proper methodology to assess the quality of xAI methods, we decided to evaluate the approach proposed by Böhle et al. [682] with the framework developed by Hedström et al. [681]. It consists of a battery of quantitative measurements of different properties that a high-quality explanation should have. We consider the following properties and metrics [681]:

- Faithfulness (do features deemed relevant affect model decisions more strongly?): Selectivity (drop in performance when features deemed relevant by an xAI method are iteratively removed) and Faithfulness Correlation (correlation between the sum of the subset of randomly replaced attributions and the difference in function output);
- Robustness (are explanations stable when subject to slight perturbations in the input?): Maximum Sensitivity (maximum sensitivity of an explanation using Monte Carlo sampling) and Relative Representation Stability (distance between the internal model representation of an explanation and its perturbed version);
- Complexity (are explanations concise?): Sparsity (Gini Index to measure if only highly attributed features are predictive of the model output) and Complexity (entropy of the contribution of all features to the total magnitude of the attribution);
- Localisation (is the explanatory evidence centred around a given region of interest?): Focus (quantifies the precision of the explanation by creating mosaics of images from different classes) and Attribution Localisation (ratio of positive attributions within a targeted object and the total positive attributions).

6.4.2.4 Pre-training Strategies

In their work, Böhle et al. [682] compared B-cos architectures with their original counterparts, but only on ImageNet [362]. Thus, their experiments involved training both types of networks from scratch. In the medical imaging domain, it is common practice to employ transfer learning [684]. Therefore, we test different pre-training strategies for B-cos and original networks: no pre-training, pre-training on ImageNet, and initialising B-cos networks from differently trained baselines.

6.4.2.5 B-cos Version of a Custom Medical Imaging Model

To understand the feasibility of integrating the B-cos transform in medical deep learning models, we adopted the work of Liu et al. [683] as a use case. Their Clinical-Inspired Network (CI-Net) is an attention-based model that aims to mimic the clinical diagnostic process. This model achieved state-of-the-art performance on six benchmark databases for skin lesion analysis. Contrary to commonly used networks like the DenseNet [359], the CI-Net contains two branches and a distinguish module, which learns to discriminate images from different classes, making it more challenging to integrate the B-cos transform into this architecture.

6.4.2.6 Implementation Details

We followed the training pipelines described by Böhle et al. [682] (all models except CI-Net-based models) and Liu et al. [683] (CI-Net-based models). Regarding data processing and augmentation, we resized all the images to 448×448 , applied random affine transforms (i.e., rotation, translation, scale, and shear) according to specific literature on each use case (more details in the code), and normalised the inputs accordingly. We divided all datasets into stratified train-validation-test splits, with 20% of the data for testing on datasets without a default test set and 10% of the remaining data for validation. For ISIC2018, we used the official split. We saved the best models according to the best F1 score on the validation set. For the explanation quality metrics, we evaluated all models under the same conditions (i.e., only on the set of correctly predicted images by all models under test). For the multi-label classification scenario (i.e., VinDr-CXR), an image is considered correctly classified only when the model predicts all labels correctly, and we explain only the label with the lowest index. For datasets where all models correctly predict more than 100 images, we evaluated only 100 (for the DTU/Herlev Pap Smear Database, we evaluated 40 images).

6.4.3 Results and Discussion

Table 6.3 presents the results on model performance and explanation quality assessment obtained for each model across the datasets used in this work.

6.4.3.1 Training

Except for the CI-Net, we focus on DenseNet-121 [359], since it is a popular architecture among the medical imaging community and Böhle et al. [682] already showed that their conclusions hold for different architectures. Another important aspect is that B-cos networks take twice as long to train and require about twice as much memory.

6.4.3.2 Pre-training Strategies

For ISIC2018 and APTOS2019, we study the effect of different pre-training strategies. In both cases, the performance improves on baseline and B-cos models when ImageNet pre-training is used (see

model 3 vs model 6 or model 4 vs model 7). Interestingly, on APTOS2019, transferring knowledge from the baseline model already adapted to the final domain worsens the results compared to the B-cos network trained from scratch (see models 13 and 16 vs model 12), and, on both datasets, it falls short of the performance obtained when transferring knowledge from a B-cos network trained on another domain (see models 5 and 8 vs models 7, 13 and 16 vs model 15). These results show that it is preferable to leverage knowledge from another B-cos network that already learned the weight-input alignment pressure (even if on another domain) than to learn from a non-B-cos network (i.e., non-aligned) on the same domain.

6.4.3.3 Adaptation of a Custom Model

Transforming a custom model into a B-cos network is a straightforward process given the code made available by Böhle et al. [682]. As expected, this custom architecture outperforms DenseNet121 on the baseline and B-cos versions regarding the *Localisation* metric. Interestingly, the baseline CI-Net is not better regarding classification than the DenseNet121, but the B-cos version achieves the best results (see models 9 and 10).

6.4.3.4 Explanation Quality Assessment

We started by applying our evaluation protocol to ImageNet-trained networks made available by the original authors to verify if the results they reported held for different properties of explanations (Böhle et al. et al. only evaluate the *Localisation* metric in their work [682], with a metric similar to *Focus*). In four out of seven metrics, we can see that their conclusions hold (i.e., post-hoc methods applied on B-cos networks yield better explanations; in turn, the inherent B-cos explanations surpass other methods on B-cos and non-B-cos networks).

In the medical use cases of this work, we obtain results with some variations. For ISIC2018 and APTOS2019, we observe that the trend on *Selectivity* is reversed compared to the trend on ImageNet. For the DTU/Herlev Pap Smear Database and VinDr-CXR, explanations obtained from B-cos networks are better than explanations from the original networks. Nevertheless, the inherent B-cos explanations do not surpass other xAI methods (e.g., 4.56 for Grad-CAM vs 5.04 on VinDr-CXR). Concerning *Faithfulness Correlation*, B-cos explanations surpass other methods except on VinDr-CXR. Focusing on *Robustness*, B-cos networks and explanations are generally better for *RRS* but worse for *Maximum Sensitivity*. The same applies to *Sparseness* and *Complexity*. Finally, on *Localisation* metrics, B-cos explanations surpass the best explanation methods on B-cos and non-B-cos models for three out of four datasets, but only for ImageNet pre-trained models (check models 6/7, 9/10, 14/15, 19/20). Further, on VinDr-CXR, where local bounding box annotations are available, the *Attribution Localisation* metric shows the trend the original authors found. These results indicate that B-cos explanations are generally more focused than other explanations. Figure 6.2 presents the qualitative results that support these findings. Visually, Input×Gradient produces sparser maps, while Grad-CAM and B-cos explanations are more concentrated around a single area. Furthermore, B-cos maps tend to have less highly activated pixels than Grad-CAM's.

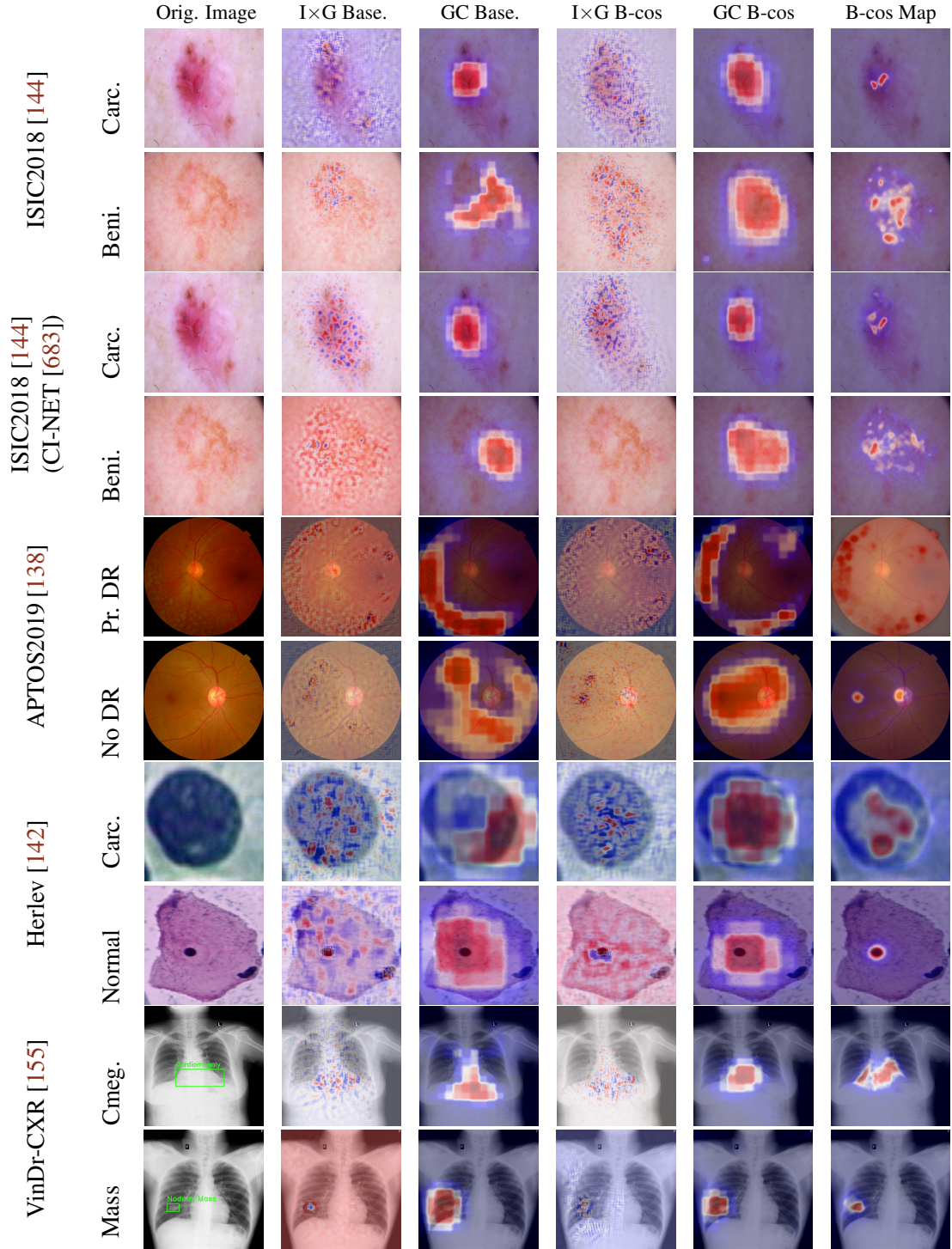


Figure 6.2: Comparing explanations from Input×Gradient [6] (I×G), Grad-CAM [7] (GC) and B-cos on original (Base.) and B-cos networks. The colour map ranges from blue to red, blue(red) meaning less(more) relevant pixels. Herlev stands for DTU/Herlev Pap Smear Database.

Table 6.3: Results on model performance and xAI quality assessment obtained for each model across several datasets. For all models we show the results for the best xAI method. For B-cos networks we also show results for the intrinsic B-cos explanations. We compare against the following xAI methods: G - Grad, GC - Grad-CAM [7], I×G - Input×Gradient [6], IG - Integrated Gradients [8]. For each metric $\uparrow$ ($\downarrow$) means higher (lower) is better. Underlined numbers represent the best result for non-B-cos models on a given dataset, while bold numbers highlight the best B-cos results. Herlev stands for DTU/Herlev Pap Smear Database, DN121 stands for DenseNet121 and RN18 IN corresponds to an ImageNet trained ResNet18.

Dataset	ID	Model	Pre-training	Task Performance			Faithfulness		Robustness		Complexity		Localisation	
				(B)ACC $\uparrow$	F1 $\uparrow$	AUC $\uparrow$	Select. $\downarrow$	Correl. $\uparrow$	Sens. $\downarrow$	RRS $\downarrow$	Sparse. $\uparrow$	Complex. $\uparrow$	Focus $\uparrow$	Att. Loc. $\uparrow$
ImageNet [362]	1	DN121	-	72.6	72.3	<u>99.8</u>	7.63 (IG)	0.073 (I×G)	<u>0.545</u> (GC)	1.62 (G)	0.610 (I×G)	<u>10.5</u> (GC)	0.722 (GC)	-
	2	B-cos DN121	-	73.6	73.2	99.2	5.92 (G) 4.85	0.138 (IG) 0.043	0.994 (GC) 1.54	1.22e-4 (G) 6.92e-5	0.604 (IG) 0.624	10.3 (GC) 10.1	0.748 (GC) 0.851	-
	3	DN121	-	76.7	62.5	92.4	<u>25.9</u> (IG)	0.063 (GC)	1.80 (IG)	<u>0.012</u> (IG)	0.583 (G)	11.6 (IG)	0.707 (GC)	-
	4	B-cos DN121	-	74.9	57.8	89.1	48.1 (GC) 91.4	0.487 (I×G) -0.020	0.964 (GC) 1.13	4.33e-6 (I×G) 8.87e-8	0.612 (I×G) 0.443	11.9 (GC) 11.8	0.434 (GC) 0.585	-
ISIC2018 [144]	5	B-cos DN121	3	75.5	59.6	90.4	35.7 (IG) 70.7	0.297 (I×G) 0.057	2.19 (GC) 1.94	2.84e-6 (IG) 4.81e-8	0.657 (IG) 0.550	11.4 (G) 11.6	0.539 (GC) 0.614	-
	6	DN121	1	<u>85.6</u>	<u>78.5</u>	<u>97.6</u>	100.3 (IG)	<u>0.288</u> (GC)	<u>1.32</u> (G)	0.507 (IG)	0.606 (GC)	<u>11.7</u> (IG)	0.772 (GC)	-
	7	B-cos DN121	2	82.3	71.2	93.7	65.6 (IG) 46.6	0.114 (I×G) 0.024	2.44 (G) 6.62	9.31e-7 1.70e-5	0.707 (GC) 0.767	11.5 (I×G) 10.9	0.619 (I×G) 0.796	-
	8	B-cos DN121	6	77.1	62.0	90.9	26.8 (GC) 48.4	0.170 (I×G) 0.027	2.68 (GC) 1.62	2.58e-7 (G) 1.33e-7	0.631 (I×G) 0.616	11.5 (G) 11.5	0.668 (GC) 0.656	-
	9	Baseline CI-Net [683]	RN18 IN	84.2	74.2	96.3	30.3 (I×G)	0.173 (GC)	2.00 (IG)	0.250 (G)	<u>0.716</u> (GC)	11.4 (IG)	<u>0.887</u> (GC)	-
	10	B-cos CI-Net [683]	RN18 IN	82.9	71.4	94.4	52.8 (I×G) 77.3	0.144 (IG) 0.061	1.40 (G) 3.51	1.55e-5 (G) 1.10e-5	0.744 (GC) 0.709	11.5 (I×G) 11.2	0.591 (IG) 0.889	-
APTOS2019 [138]	11	DN121	-	79.3	57.6	93.0	<u>32.3</u> (I×G)	0.049 (GC)	1.49 (G)	<u>0.038</u> (IG)	0.791 (GC)	<u>11.4</u> (IG)	0.695 (GC)	-
	12	B-cos DN121	-	79.3	58.6	90.9	88.1 (GC) 82.3	0.365 (IG) 0.019	29.1 (G) 18.4	1.02e-5 (I×G) 1.91e-5	0.723 (IG) 0.735	11.4 (GC) 11.0	0.413 (GC) 0.585	-
	13	B-cos DN121	11	75.7	46.2	88.1	41.5 (I×G) 51.7	0.104 (I×G) 0.027	1.66 (GC) 1.46	1.08e-6 (IG) 7.79e-7	0.721 (GC) 0.723	11.1 (G) 11.0	0.549 (GC) 0.642	-
	14	DN121	1	<u>82.9</u>	<u>66.2</u>	<u>94.7</u>	37.1 (IG)	<u>0.180</u> (GC)	<u>1.41</u> (G)	0.156 (G)	<u>0.802</u> (GC)	11.2 (IG)	<u>0.845</u> (GC)	-
	15	B-cos DN121	2	81.3	61.9	92.6	56.1 (I×G) 69.3	0.061 (GC) 0.034	2.61 (IG) 5.10	9.64e-7 (G) 1.35e-6	0.784 (GC) 0.826	11.2 (I×G) 10.5	0.729 (GC) 0.870	-
	16	B-cos DN121	14	78.9	54.8	91.4	51.0 (IG) 53.6	0.0578 (GC) 0.046	1.61 (I×G) 4.99	1.05e-6 (G) 1.67e-5	0.792 (GC) 0.808	10.8 (I×G) 10.6	0.791 (GC) 0.818	-
Herlev [142]	17	DN121	1	<u>66.2</u>	<u>71.0</u>	<u>94.3</u>	10.1 (I×G)	0.143 (GC)	<u>0.912</u> (GC)	0.0530 (Grad)	0.553 (I×G)	<u>10.6</u> (GC)	<u>0.732</u> (GC)	-
	18	B-cos DN121	2	63.5	67.1	91.4	8.67 (I×G) 9.65	0.124 (GC) 0.144	2.03 (IG) 2.68	1.29e-6 2.24e-6	0.557 (IG) 0.572	10.3 (GC) 10.2	0.606 (GC) 0.666	-
VinDr-CXR [155]	19	DN121	1	<u>96.2</u>	<u>28.4</u>	74.8	9.13 (I×G)	<u>0.119</u> (GC)	<u>1.00</u> (I×G)	2.97e-4 (I×G)	0.740 (GC)	<u>11.6</u> (IG)	0.277 (I×G)	0.137 (GC)
	20	B-cos DN121	2	95.6	22.7	77.8	4.56 (GC) 5.04	0.093 (GC) 0.066	2.58 (I×G) 13.2	1.18e-6 (IG) 1.60e-5	0.763 (GC) 0.726	11.5 (G) 11.0	0.297 (I×G) 0.301	0.237 (GC) 0.287

6.4.4 Conclusions and Future Work

This work presents the first study on the applicability and suitability of B-cos networks, a weight-input aligned architecture, for the medical domain. Although these models achieve better explanation quality than other xAI methods on ImageNet-trained models, it is not clear that the same happens for the medical use cases considered, especially when considering all explanation quality assessment metrics. However, B-cos explanations yield more localised saliency maps, a finding that corroborates the original authors' conclusions.

Given these results and the slight drop in classification performance, it is difficult to conclude that B-cos networks produce better explanations in the medical domain. Further work should be devoted to integrating B-cos networks in different tasks (e.g., image segmentation), studying different xAI quality assessment metrics with more post-hoc methods, and a deep focus on model optimisation or regularisation to promote the integration of clinical-domain knowledge into the models (e.g., integrating clinical metadata through data fusion techniques, using class weights in the loss function).

6.5 B-cos Networks for Medical Image Retrieval

The results of this work have not been disseminated in any other venue besides this thesis. The code related to the implementation of this work is publicly available in a GitHub repository².

6.5.1 Description and Contributions

In Section 5.4, we presented a pioneering approach to multi-modal retrieval for the aesthetic assessment of breast cancer surgery outcomes, and in Section 5.5, we did an extended version of that study using only attention-based deep learning models (in our case, transformers). From these works, we observed that using triplet loss optimisation with multi-modal transformer-based architectures leads to good results in retrieving similar cases regarding breast aesthetic assessment outcomes. In this work, we try to go one step further concerning interpretable retrieval by using B-cos networks, which are inherently interpretable models by design. We build upon the framework proposed in [46] and extend the experiments on the B-cos-based architectures (and their original counterparts), comparing seventeen different models and two different optimisation settings (i.e., fine-tuned, and multi-modal). We validate the results published in the original paper and report new ones with the introduced architectures.

6.5.1.1 Brief Note About the Models

In this Section, to contextualise the reader, we present a brief note about the models used in this study. Following the work presented in Section 6.4, we decided to keep DenseNet [359] as the baseline architecture given its wide use in algorithms for healthcare applications, and its integration

²<https://github.com/TiagoFilipeSousaGoncalves/bcs-aesth-mm-bcos-mir>

in the battery of experiments reported by Böhle et al. [682]. As such, we used several variations of the DenseNet architecture (i.e., DenseNet121, DenseNet161, DenseNet169, DenseNet201) along with their B-cos counterparts (i.e., B-cos DenseNet121, B-cos DenseNet161, B-cos DenseNet169, B-cos DenseNet201). Additionally, we experimented with initialisations from scratch, with pre-trained weights on ImageNet (P) and with pre-trained weights on ImageNet and fine-tuned for a long number of epochs (PL).

6.5.2 Methodology

We follow the methodology described in Section 5.4.2.

6.5.3 Results and Discussion

Table 6.4 presents the retrieval performance metrics, Contrastive Accuracy (CAcc) and Adjusted Discounted Cumulative Gain (ADCG) for the models tested in this work, categorised by aesthetic quality (i.e., related to catalogues with Excellent and Good or Fair and Poor aesthetic outcomes). We adopt the same structure presented in Section 5.5.

In agreement with the results presented in Sections 5.4–5.5, we observe that the multi-modal approaches often achieve the best results on the training set. Once again, contrastively, the results on the test set are less consistent with the image (fine-tuned) approaches being equally competitive with multi-modal strategies, suggesting that the latter might be more prone to over-fitting. However, in this study, we notice that the multi-modal approach seems to benefit the B-cos models, preventing the decrease in performance that could arise from the additional degree of complexity of the B-cos operator added to the baseline. In any case, this framework tends to overfit since the performance for both metrics decreases systematically from the training to the test set.

Although we consider that the size of the current dataset (see Chapter 3) might not be enough to draw robust conclusions regarding the path that leads to the best results, we question whether we should use image (fine-tuned) models over multi-modal approaches. While the performance results have differences (sometimes, marginal differences), one could prefer the simplicity and easiness of training of the baseline models over the complexity and training times of their B-cos counterparts. Moreover, similar to previous conclusions, it is not clear whether it is preferable to use the image (fine-tuned) or the multi-modal. If we look at the performance metrics, results suggest that using only the image information is sufficient for the model to perform a good retrieval. Naturally, one can argue that the multi-modal model has more information about the patients, and even if it suffers from a marginal performance decrease, it is still more interpretable.

6.5.4 Conclusions and Future Work

In this study, we obtain similar performance results to the work presented in Section 5.5 with DenseNet architectures and their B-cos counterparts. These results are interesting because they

Table 6.4: Retrieval performance metrics for B-cos networks (and their original counterparts), categorised by aesthetic quality. Each entry is of the form (Excellent/Good, Fair/Poor). Best results, per model, highlighted in bold. Models with a (P) refer to initialisation with pre-trained weights on ImageNet, and with a (PL) refer to initialisation with pre-trained weights on ImageNet and fine-tuned for a long number of epochs.

Model	Type	Train		Test	
		CAcc (%) ↑	ADCG (%) ↑	CAcc (%) ↑	ADCG (%) ↑
DenseNet121	Image (Fine-tuned)	(93.87, 93.93)	(97.30, 97.41)	(69.62, 73.28)	(73.87, 76.63)
	Multi-modal	(98.53, 98.97)	(99.37, 99.46)	(68.58, 71.62)	(72.13, 76.36)
DenseNet121 (P)	Image (Fine-tuned)	(96.70, 94.80)	(98.54, 97.76)	(70.66, 72.68)	(73.14, 78.36)
	Multi-modal	(98.71, 98.32)	(99.37, 99.34)	(68.41, 72.30)	(70.45, 76.05)
DenseNet161	Image (Fine-tuned)	(82.11, 83.04)	(89.00, 90.14)	(68.71, 70.86)	(70.87, 76.73)
	Multi-modal	(98.67, 97.28)	(99.41, 98.75)	(67.67, 71.92)	(70.83, 75.30)
DenseNet161 (P)	Image (Fine-tuned)	(91.39, 97.96)	(95.48, 99.21)	(69.91, 75.19)	(72.26, 80.05)
	Multi-modal	(99.02, 97.97)	(99.46, 99.29)	(68.25, 74.85)	(69.51, 79.79)
DenseNet169	Image (Fine-tuned)	(95.95, 96.10)	(98.38, 98.56)	(69.08, 73.11)	(71.34, 73.90)
	Multimodal	(99.04, 98.34)	(99.45, 99.24)	(68.29, 71.33)	(68.86, 73.83)
DenseNet169 (P)	Image (Fine-tuned)	(97.03, 97.30)	(99.02, 98.92)	(69.83, 74.55)	(71.30, 79.04)
	Multi-modal	(98.63, 98.12)	(99.44, 99.15)	(67.00, 73.79)	(69.07, 78.05)
DenseNet201	Image (Fine-tuned)	(89.96, 92.04)	(95.00, 96.49)	(68.91, 71.79)	(70.64, 75.92)
	Multi-modal	(98.48, 97.04)	(99.38, 98.74)	(67.54, 72.30)	(69.69, 76.20)
DenseNet201 (P)	Image (Fine-tuned)	(94.25, 90.41)	(97.66, 95.86)	(69.21, 73.70)	(70.44, 77.19)
	Multi-modal	(98.77, 98.74)	(99.37, 99.43)	(68.75, 73.24)	(68.30, 77.55)
B-cos DenseNet121	Image (Fine-tuned)	(96.92, 96.38)	(98.83, 98.79)	(65.17, 68.27)	(68.58, 69.78)
	Multi-modal	(98.00, 97.97)	(99.13, 99.19)	(66.33, 68.73)	(70.45, 72.59)
B-cos DenseNet121 (P)	Image (Fine-tuned)	(93.72, 95.56)	(97.14, 98.34)	(59.05, 66.53)	(64.28, 70.19)
	Multi-modal	(96.51, 96.10)	(98.65, 98.45)	(61.17, 68.73)	(63.57, 72.49)
B-cos DenseNet121 (PL)	Image (Fine-tuned)	(94.90, 91.04)	(97.88, 95.92)	(63.50, 65.72)	(66.00, 66.96)
	Multi-modal	(97.84, 97.61)	(99.19, 99.12)	(66.13, 67.97)	(67.92, 68.80)
B-cos DenseNet161	Image (Fine-tuned)	(97.70, 97.76)	(99.22, 99.10)	(66.46, 72.10)	(69.29, 77.38)
	Multi-modal	(97.93, 98.29)	(99.25, 99.32)	(65.38, 71.67)	(67.35, 77.42)
B-cos DenseNet161 (P)	Image (Fine-tuned)	(92.89, 93.97)	(96.98, 97.54)	(62.96, 66.69)	(65.09, 70.97)
	Multi-modal	(97.94, 95.30)	(99.30, 98.36)	(65.71, 71.16)	(66.14, 76.32)
B-cos DenseNet169	Image (Fine-tuned)	(97.17, 97.14)	(98.81, 99.09)	(68.04, 69.07)	(68.99, 72.67)
	Multi-modal	(97.94, 98.14)	(99.12, 99.26)	(69.00, 68.78)	(73.93, 72.43)
B-cos DenseNet169 (P)	Image (Fine-tuned)	(96.47, 95.68)	(98.70, 98.56)	(63.30, 65.29)	(67.51, 71.61)
	Multi-modal	(95.33, 95.85)	(98.10, 98.45)	(63.46, 70.60)	(63.53, 74.71)
B-cos DenseNet201	Image (Fine-tuned)	(97.50, 97.27)	(99.15, 99.12)	(65.71, 71.50)	(67.30, 74.61)
	Multi-modal	(98.51, 96.92)	(99.37, 98.82)	(69.00, 72.01)	(71.87, 73.91)
B-cos DenseNet201 (P)	Image (Fine-tuned)	(96.38, 92.74)	(98.65, 96.83)	(63.30, 66.06)	(64.42, 69.34)
	Multi-modal	(97.88, 96.51)	(99.27, 98.87)	(65.33, 70.09)	(67.08, 74.36)

challenge the conclusions presented in Sections 5.4–5.5, where we concluded that transformer-based models lead to better results on retrieval. Additionally, these results also highlight at least one limitation of the comparative study showed in Section 5.4, suggesting that ResNet or VGG-based models have architectural limitations that led us to conclude (wrongly) that CNNs can not compete with transformers. In contrast, we again observe that only using image features might be equivalent to using multi-modal data, undermining the intuition of using more information about the patient to achieve refined results.

Focusing now on the purpose of this study (i.e., comparing baseline CNNs with their B-cos counterparts), we notice that multi-modal approaches seem to improve the predictive performance of B-cos models. Nevertheless, like in previous studies, the differences in results are marginal in most cases. Apart from gathering more data and building new catalogues or training different models (i.e., segmentation-based and end-to-end multi-modal approaches), further work should focus on generating the visual explanations of these models (i.e., saliency maps) and understanding if the conclusions of the work in Section 6.4 are still valid for B-cos networks (and their baseline counterparts).

6.6 Explaining Counterfactuals with Prototypes

The results of this work have not been disseminated in any other venue besides this thesis. The code related to the implementation of this work is publicly available in a GitHub repository³.

6.6.1 Description and Contributions

Aiming to bring interpretability into deep learning architectures, Chen et al. [5] proposed the *prototypical part network* (ProtoPNet) that uses a *prototype layer* to learn *prototypical parts* of the images and integrates this information into the classification process. Their work resulted in a new paradigm for intrinsically interpretable deep neural networks where the main focus is to provide the user with a case-based reasoning framework. On the other hand, the scientific community generically agrees that a model needs to provide human causally understandable explanations (also known as *causability*) to show a certain degree of human-level explainability [685]. The class of algorithms that may present an answer to this challenge are the counterfactual reasoning models (also known as *counterfactuals*).

In this work, we establish an interface between counterfactual reasoning and intrinsically interpretable deep learning models (based on prototype learning) and propose future research directions in these two distinct (but complementary) areas. We present a series of experiments aimed at answering the following research questions:

1. Can we ensure that deep learning models keep their predictive performance after integrating a prototype learning layer [5, 656] into their architectures?

³<https://github.com/TiagoFilipeSousaGoncalves/proto-counterfactuals>

2. Even though these models promote learning meaningful prototypes during training, how much can we rely on these prototypes, and how should we evaluate the quality of the prototypes these models learned?
3. Can we use prototype-based deep learning models to generate counterfactual explanations that rely on the prototypes learned by such models? If so, can we devise a straightforward manner to get the most suitable counterfactual, supported in prototypes, given a query image?

This work tackles counterfactual reasoning from a different perspective than the one usually followed by literature by bringing prototype-based deep learning models into the discussion. As such, the main contributions of this work are:

1. A comparative study with three different use cases (i.e., bird species classification, glaucoma diagnosis, and skin lesion analysis) using three different architectures (i.e., the combination of three different backbones and two prototype-based models) that focus on their predictive power;
2. An objective analysis of the prototypes learned by the prototype-based models in terms of their class identity (i.e., evaluating whether the most activated prototypes of a given image come from examples of the same class) and in terms of their level of coherence (i.e., evaluation of the class identities of prototypes of the same image across backbones and prototype layers);
3. A study of two counterfactual retrieval methods based on the vector of probabilities of the classification layer of a deep learning model: we consider the index of the second-highest value as the counterfactual label; then, we experiment with two different feature spaces to perform the retrieval of the image of the counterfactual label more similar to the query image; and, finally, we use the most activated prototypes as explanations of the predictions of the model;
4. An objective analysis of the counterfactual labels across model architectures that aims to understand their coherence.

The results obtained in this work allowed us to devise future research lines on generating counterfactual explanations using prototype-based deep learning algorithms while reinforcing the need for objective evaluation metrics to evaluate such explanations and standardise concepts related to prototype-based learning. We provide an overview of our proposed framework in Figure 6.3.

6.6.2 Methodology

The literature review in Section 6.3 has revealed a gap between prototypical deep learning algorithms and their application in generating counterfactual explanations. Hence, we propose a new framework to obtain counterfactuals and their respective explanations. We start by training

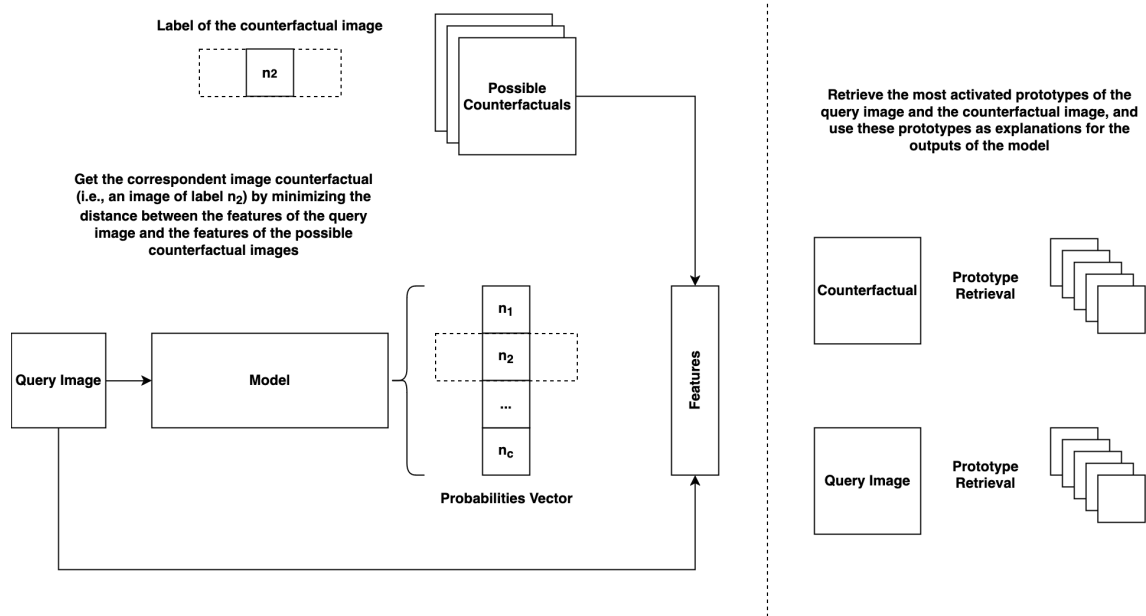


Figure 6.3: Overview of our proposed framework to obtain image counterfactuals and the open problem of using prototypes to explain that counterfactual and justify the model's prediction. Given a query image, we perform model inference to get the vector of probabilities and then select the index of the second-highest value as the counterfactual label. Afterwards, we use the model to extract features of the query image and all the possible counterfactual images in the database and minimise the Euclidean distance between these possible pairs (i.e., the most similar counterfactual corresponds to the minimum distance). To explain the model's predictions, we retrieve the prototypes of the query and counterfactual images.

a prototypical learning model on a given task. Then, given a query image, we use that model to perform inference and obtain the vector of probabilities. We use the index of the second-highest value as the counterfactual label. Afterwards, we use the model to extract features of the query image and all the possible counterfactuals in the database and minimise a distance metric (in our case, Euclidean distance) between these possible pairs: the most similar counterfactual will correspond to the minimum distance. The remaining problem is the strategy we should use to explain these counterfactuals. In our approach, we retrieve the learned prototype of the query and its correspondent counterfactual image and use that information to explain the outputs of this framework (see Figure 6.3).

6.6.2.1 Data

In this work, we used CUB-200-2011, PAPILA and PH2 datasets (see Chapter 3). Regarding data manipulation strategies, we refer the reader to the guidelines described by Chen et al. [5]. We cropped images according to their regions of interest (ROIs) using the bounding-box annotations (CUB-200-2011) or their segmentation masks (PAPILA and PH2). Then, we split the data into train and test sets using the official splits (CUB-200-2011) or with a proportion of 70%-train/30%-test (PAPILA and PH2). Data augmentation strategies included random rotations, skew and shear distortions, and left-right flips.

6.6.2.2 Models

We used DenseNet121 [359], ResNet34 [263], and VGG16 [513] as feature extraction layers in three different types of architectures: **Baseline**, consisting of the feature extraction layers and a fully connected layer for classification; **ProtoPNet**, consisting of the feature extraction layers and the prototypical architecture [5]; and **Deformable ProtoPNet**, consisting of the feature extraction layers and the deformable prototypical architecture [656]. We chose these two prototypical architectures for our experiments because of their author affiliations (i.e., they belong to the same research group). In this work, we consider Deformable ProtoPNet a directly improved version of ProtoPNet.

6.6.2.3 Training and Testing Settings

Baseline We initialised all the models using their pre-trained weights on ImageNet [101]. In the first five epochs, we froze the feature extraction layers and trained the classification layer, using the adaptive moment estimation (Adam) [400] optimiser with a learning rate of 1×10^{-4} . During the remaining epochs, we trained all the layers of the model, using the Adam optimiser with a learning rate of 1×10^{-4} and weight decay of 1×10^{-3} for the feature extraction layers and a learning rate of 1×10^{-4} for the classification layer. We updated the classification layer for 20 iterations, using the Adam optimiser with a learning rate of 1×10^{-4} , every ten epochs. We used the cross-entropy loss function and saved the best model weights with the best accuracy in the test set.

ProtoPNet All the strategies follow the guidelines described by Chen et al. [5]. In the first five epochs, we trained the additional convolutional and prototype layers, using the Adam optimiser, with a learning rate of 3×10^{-3} and a weight decay of 1×10^{-3} for the additional convolution layers, and a learning rate of 3×10^{-3} for the prototype layer. During the remaining epochs, we trained all the layers of the model, using the Adam optimiser with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-3} for the feature extraction layers, a learning rate of 3×10^{-3} and a weight decay of 1×10^{-3} for the additional convolutional layers, and a learning rate of 3×10^{-3} for the prototype layer. We updated the classification layer for 20 iterations using the Adam optimiser, with a learning rate of 1×10^{-4} , every ten epochs (along with the *pushing* operation). We used the loss function described in [5] and saved the best model weights with the best accuracy in the test set.

Deformable ProtoPNet All the strategies follow the guidelines described by Donnelly et al. [656]. In the first five epochs, we trained the additional convolutional and deformable prototype layers, using the Adam optimiser with a learning rate of 3×10^{-3} and a weight decay of 1×10^{-3} for the additional convolutional layers and a learning rate of 3×10^{-3} for the deformable prototype layer. In the following five epochs, we trained additional convolutional, feature extraction and the deformable prototype layers, using the Adam optimiser with a learning rate of 3×10^{-3} and a weight decay of 1×10^{-3} for the additional convolutional layers, a learning rate of 1×10^{-4} and a weight decay of 1×10^{-3} for the feature extraction layers, and a learning rate of 3×10^{-3} for the deformable prototype layer. During the remaining epochs, we trained all the layers of the model, using the Adam optimiser with a learning rate of 3×10^{-3} and a weight decay of 1×10^{-3} for the additional convolutional layers, a learning rate of 1×10^{-4} and a weight decay of 1×10^{-3} for the feature extraction layers, a learning rate of 3×10^{-3} for the deformable prototype layer, a learning rate of 1×10^{-4} for the convolutional offset layer (i.e., the layer which pre-processes the representation maps that will be the input of the deformable convolution), and a learning rate of 1×10^{-4} for the classification layer. We performed the *pushing* operation every ten epochs. We used the loss function described in [656] and saved the best model weights with the best accuracy in the test set.

6.6.2.4 Prototype Retrieval and Analysis

We retrieved the *top-k* (with $k = 10$) most activated prototypes per image (of the train and test sets). We also saved the class identities of each prototype (i.e., the ground-truth class of the image from where we retrieved the prototype). From this data, we extracted the following statistics: the **average number of class-identity prototypes** per correctly classified image, and the **level of label coherence of the prototypes** among backbone models across architectures.

Average number of class-identity prototypes We limit the analysis to the set of correctly classified images. Then, from the *top-k* (with $k = 10$) most activated prototypes, we count the

number of prototypes that originated from the class identity of that image. With this experiment, we try to understand if most prototypes activated by that image are also from that same class.

Level of label coherence of the prototypes For this experiment, we restricted our assessment to the images correctly classified by the three backbones in ProtoPNet and Deformable ProtoPNet (we can only retrieve prototypes from these two architectures). From this subset, we get the 3 vectors (one per backbone) of *top-k* (with $k = 10$) most activated prototypes per image and compute the average Wasserstein distance value per image (resulting from the average of Wasserstein distances of all the pairs of most activated prototypes vectors). The Wasserstein distance (also known as earth mover’s distance) is a distance metric between two distributions, interpreted as the minimum cost of transforming one distribution into the other [686]. Hence, the lower the distance, the higher the similarity between such distributions. In this experiment, we aim to evaluate the stability of the model (i.e., ProtoPNet or Deformable ProtoPNet) about the backbone in terms of the learned prototypes. Therefore, we consider each backbone as a prototype rater and use the average Wasserstein distance between these 3 vectors as *proxy* to the coherence of the learned prototypes (i.e., the **lower** the distance values, the **higher** the consistency).

6.6.2.5 Counterfactual Retrieval and Analysis

We started by determining the label of the counterfactual of a given query image by observing the vector of probabilities of the classification layer: the index of the highest value is the predicted label, followed by the index of the counterfactual. Then, we search the database for the most similar image (to the query image) of the counterfactual label. To evaluate image similarity, we compute the l_2 -norm between the latent space of the query image and the candidate counterfactuals. The minimum distance corresponds to the counterfactual image. A secondary outcome of this work is related to the latent space we may use to retrieve counterfactual(s): the **backbone features** space and the **prototypical features** space. The backbone feature space is the embedding that results from the last convolutional layer of the backbone of a given architecture; this feature space often appears in retrieval tasks in computer vision [49]. The prototypical feature space is the embedding that results from the prototype layer of a given architecture; this embedding encodes the *distances* of the prototypes to their identity classes. Lastly, we use the *top-k* (with $k = 10$) most activated prototypes to explain the differences between the query image and its counterfactual to the user.

Level of label coherence of the counterfactuals In this experiment, we try to assess if a given architecture (i.e., Baseline, ProtoPNet, or Deformable ProtoPNet) contributes to the label consistency of the retrieved counterfactuals. Hence, we consider each backbone a counterfactual rater and compute the Fleiss’ Kappa (a method to measure agreement between n raters, where $n > 2$). The value of the Fleiss’ Kappa ranges from 0 (i.e., no agreement or random agreement) to 1 (i.e., perfect agreement) [687, 688, 689]. This test is slightly different from the level of label coherence of prototypes: in this case, we aim to evaluate the consistency between the counterfactual

labels obtained from each backbone (i.e., the *raters*), where, in the first case, we want to see how different the distributions of *top-k* (with $k = 10$) most activated prototypes are.

6.6.3 Results and Discussion

This Section discusses the results obtained in the experimental part of this work. Figures 6.4 – 6.5, Figures 6.6 – 6.7, and Figures 6.8 – 6.9 show the behaviour of our proposal for CUB-200-2011, PAPILA and PH2 databases, respectively, using the ProtoPNet and Deformable ProtoPNet architectures with DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective *top-k* (with $k = 10$) most activated prototypes to explain their differences.

6.6.3.1 Predictive Performance of Algorithms

Table 6.5 presents the accuracy results on the test set of each database. Across databases, results suggest that prototypical learning architectures usually achieve better performances. Interestingly, the ProtoPNet architecture seems to be the architecture that better interacts with the different backbones (i.e., 4 ProtoPNet-based models among the best achieved accuracy results).

Table 6.5: Accuracy results on the test set of each database. Best values per backbone highlighted in bold.

Database	Backbone	Architecture		
		Baseline	ProtoPNet	Deformable ProtoPNet
CUB-200-2011	DenseNet121	74.58%	76.37%	77.46%
	ResNet34	71.85%	77.74%	74.54%
	VGG16	62.34%	73.84%	71.51%
PAPILA	DenseNet121	75.00%	75.68%	80.41%
	ResNet34	76.35%	72.97%	75.68%
	VGG16	77.03%	75.68%	74.32%
PH2	DenseNet121	66.67%	73.33%	66.67%
	ResNet34	68.33%	75.00%	58.33%
	VGG16	68.33%	63.33%	70.00%

6.6.3.2 Quality of the Learned Prototypes

Table 6.6 presents the average number of class-identity prototypes per correctly classified image of each database. Across the three backbones, it is clear that, almost in every setting, the prototypes learned by the Deformable ProtoPNet architecture seem to connect better with their class identities. These differences increase when the number of classes increases (i.e., when we look at the results obtained for the CUB-200-2011 database). Table 6.7 presents the level of label coherence of the

prototypes per architecture on each database. Once again, the Deformable ProtoPNet shows the best results, showing that this architecture promotes more robustness in the learning process of the prototypes.

Table 6.6: Average number of class-identity prototypes per correctly classified image of each database. Best values per backbone highlighted in bold.

Database	Backbone	Architecture	
		ProtoPNet	Deformable ProtoPNet
CUB-200-2011	DenseNet121	1.62	6.89
	ResNet34	1.97	4.20
	VGG16	3.73	9.49
PAPILA	DenseNet121	7.89	9.88
	ResNet34	7.35	9.95
	VGG-16	9.33	8.00
PH2	DenseNet121	8.99	9.57
	ResNet34	9.47	9.60
	VGG16	6.76	9.87

Table 6.7: Label coherence of the prototypes among backbone models across architectures on each database. Best value highlighted in bold.

Database	Architecture	Average Wasserstein Distance
CUB-200-2011	ProtoPNet	29.73
	Deformable ProtoPNet	27.25
PAPILA	ProtoPNet	0.35
	Deformable ProtoPNet	0.20
PH2	ProtoPNet	0.26
	Deformable ProtoPNet	0.05

6.6.3.3 Quality of the Counterfactuals

Table 6.8 presents the level of label coherence of the counterfactuals per architecture. Interestingly, across databases, the Baseline architecture obtains the best level of coherence regarding the label of its generated counterfactuals.

6.6.4 Conclusions and Future Work

6.6.4.1 Conclusions

In this work, we presented a comparative study of two state-of-the-art prototype-based deep learning architectures (i.e., ProtoPNet and Deformable ProtoPNet) using three backbones (i.e., DenseNet121,

Table 6.8: Label coherence of the counterfactuals among backbone models across architectures on each database. Best values highlighted in bold.

Database	Architecture	Average Fleiss Kappa
CUB-200-2011	Baseline	0.259
	ProtoPNet	0.149
	Deformable ProtoPNet	0.004
PAPILA	Baseline	0.495
	ProtoPNet	0.440
	Deformable ProtoPNet	0.263
PH2	Baseline	0.519
	ProtoPNet	0.037
	Deformable ProtoPNet	0.250

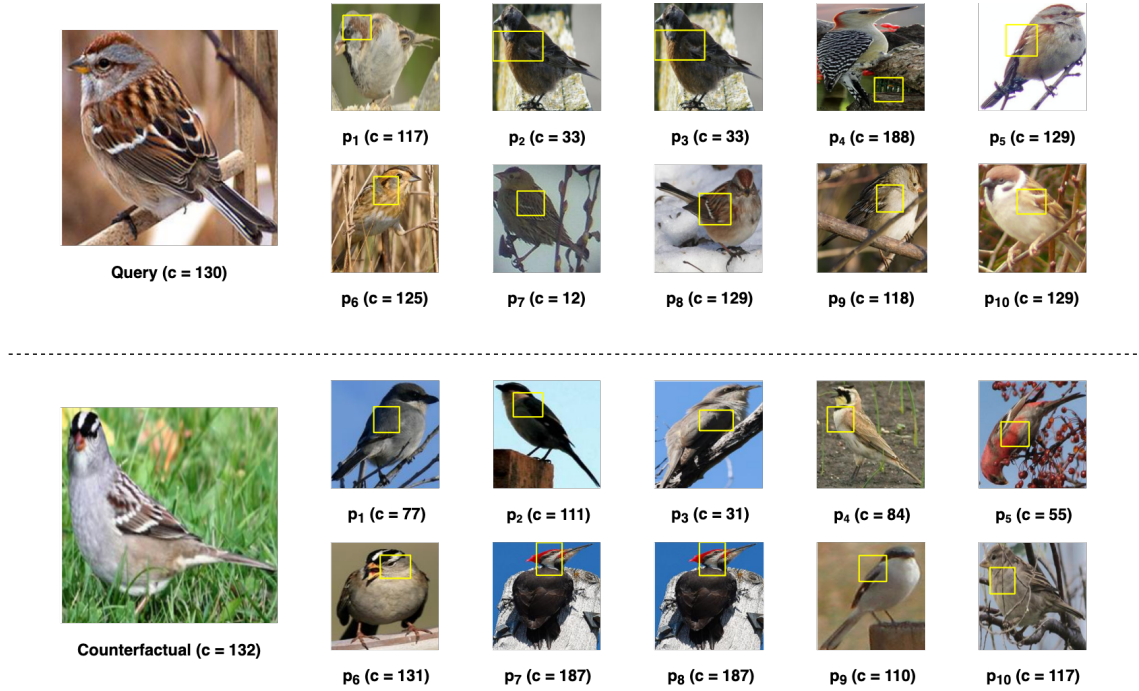


Figure 6.4: Example of our proposal using the CUB-200-2011 database and the ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective *top-k* (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for **prototype** and c stands for **class**.



Figure 6.5: Example of our proposal using the CUB-200-2011 database and the Deformable ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective *top-k* (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for **prototype** and c stands for **class**.

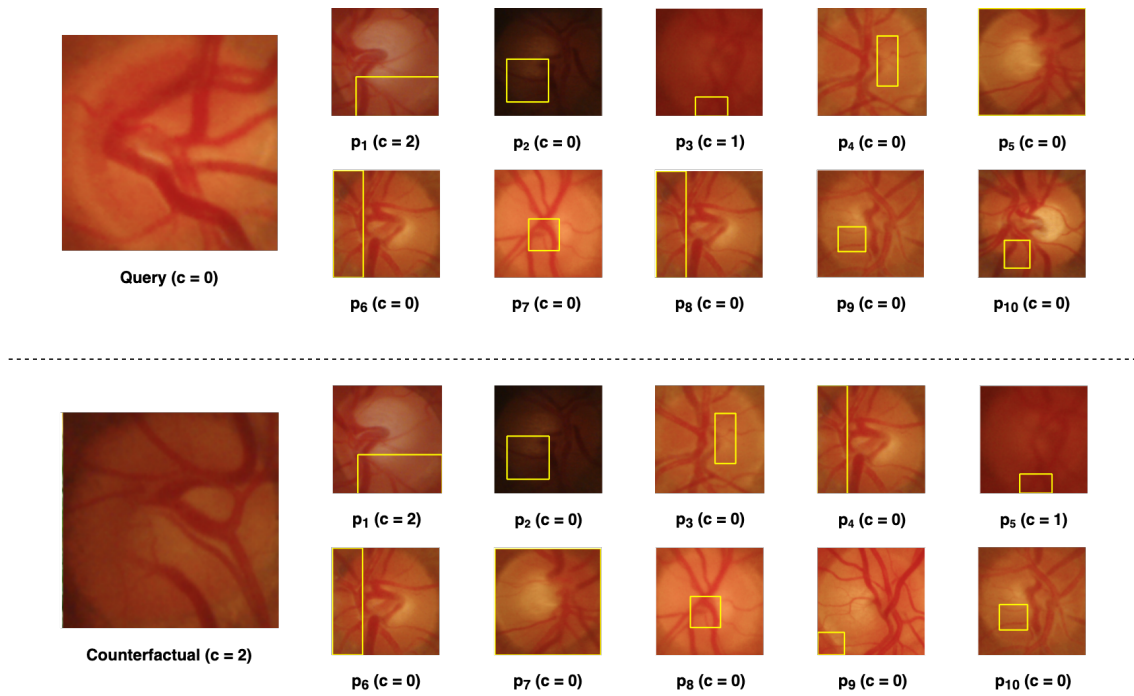


Figure 6.6: Example of our proposal using the PAPILA database and the ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective *top-k* (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for **prototype** and c stands for **class**.

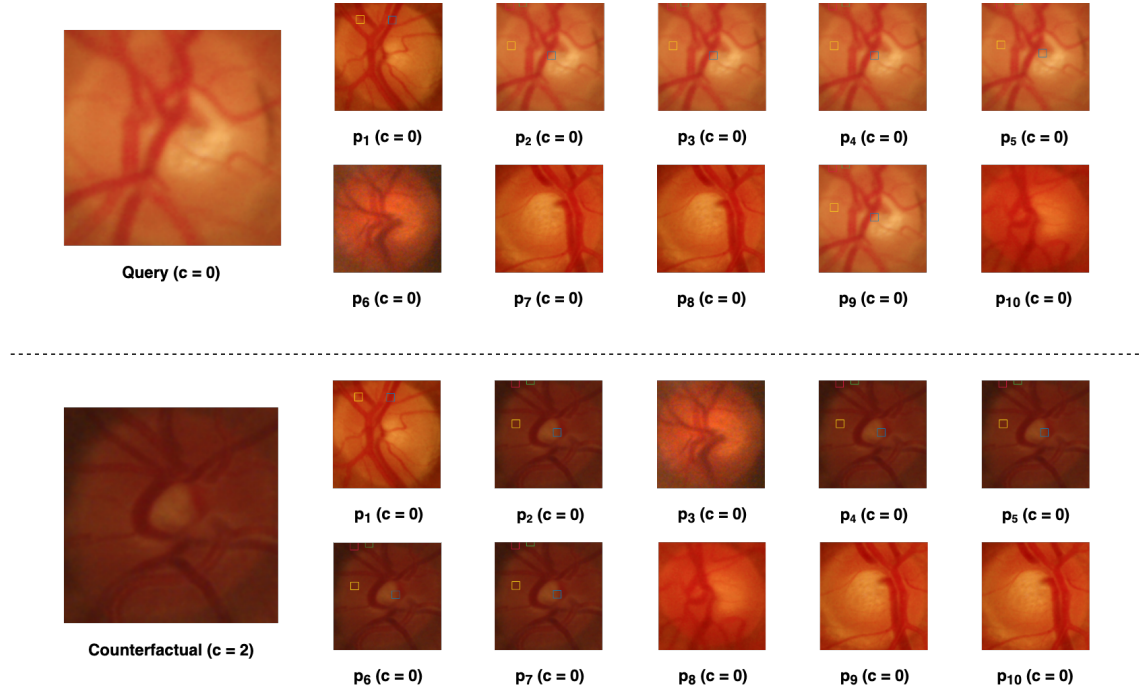


Figure 6.7: Example of our proposal using the PAPILA database and the Deformable ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective *top-k* (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for **prototype** and c stands for **class**.

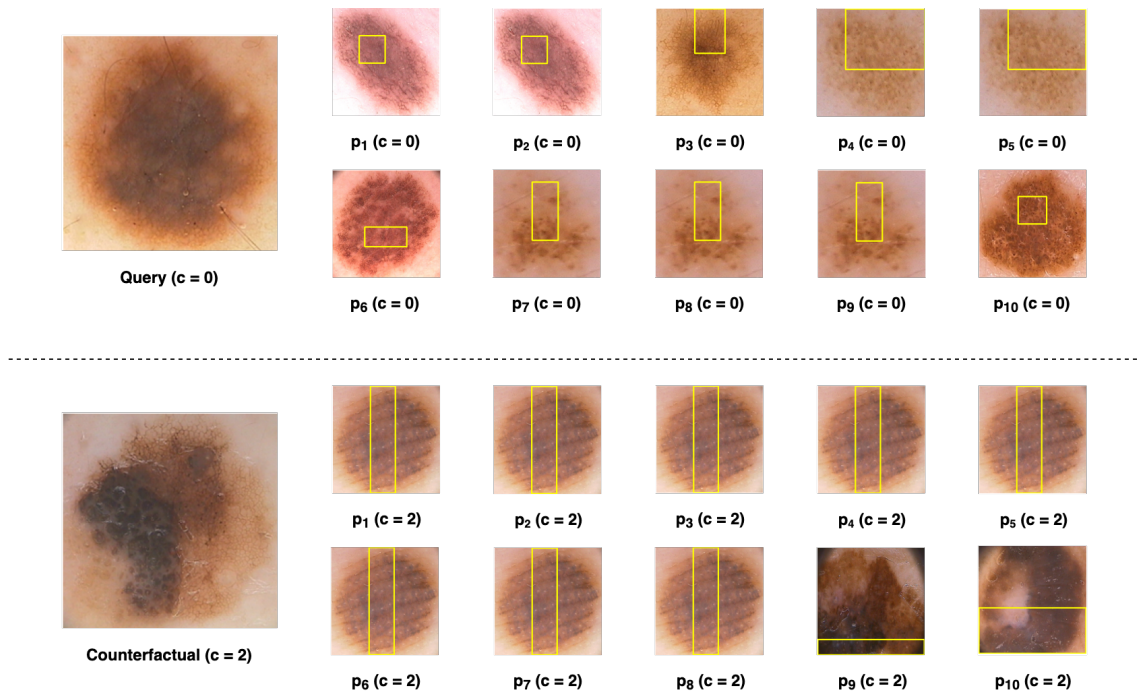


Figure 6.8: Example of our proposal using the PH2 database and the ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective *top-k* (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for **prototype** and c stands for **class**.

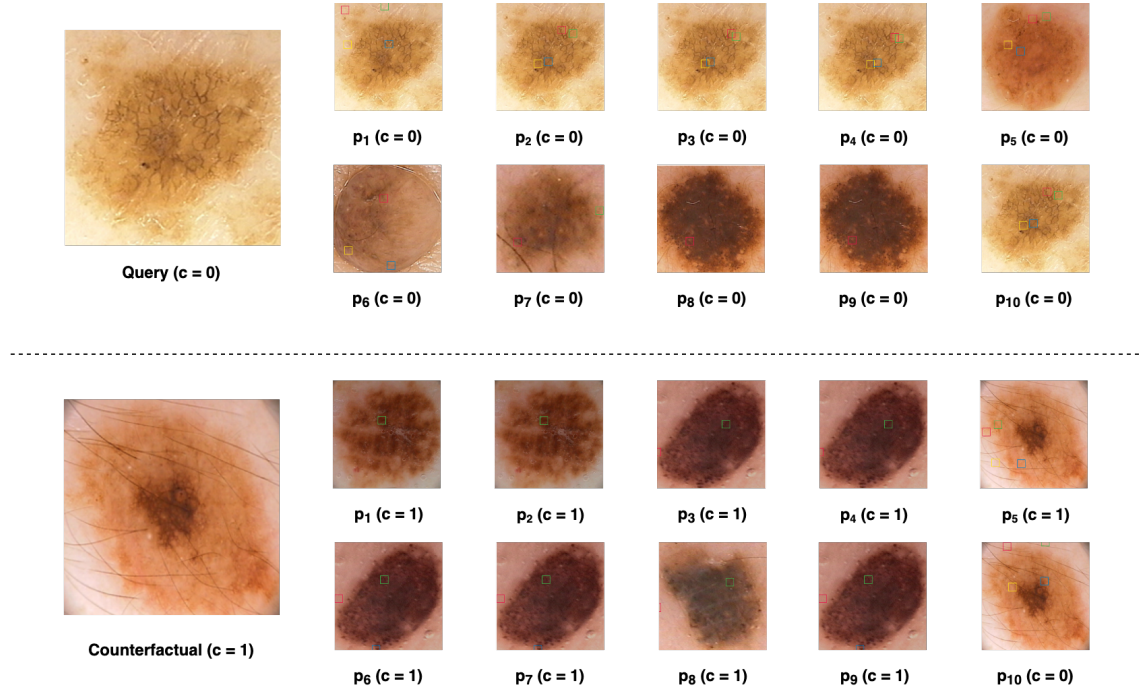


Figure 6.9: Example of our proposal using the PH2 database and the Deformable ProtoPNet architecture, using DenseNet121 as the backbone. Given a query image, we search the database for the most similar counterfactual image (using either the backbone features or the prototypical features) and use their respective *top-k* (with $k = 10$) most activated prototypes to explain their differences. Note: p stands for **prototype** and c stands for **class**.

ResNet34, and VGG16) with three different use cases (i.e., bird species identification, glaucoma detection, and skin lesion analysis). Our goal was to understand the potential and limitations of generating counterfactual explanations using prototype-based deep learning algorithms while highlighting the need for objective evaluation metrics to evaluate such explanations and standardise concepts related to prototype-based learning.

Predictive Performance of Algorithms Regarding the predictive performance of deep learning models, we concluded that the usage of prototype layers, together with the selection of an appropriate backbone, does not harm the accuracy of the models, which means that this could be an interesting baseline for the training of future deep learning models in domains that require special care in terms of transparency and explainability. On the other hand, there is no clear evidence that these frameworks improve predictive performance. We argue that the research community should also work on this issue while focusing on improving transparency and explainability.

Quality of the Learned Prototypes When analysing the prototypes learned by each framework, we found that the number of class identity prototypes per correctly classified image tends to decrease as the number of classes increases. Nevertheless, results suggest that the Deformable ProtoPNet promotes the learning of prototype feature spaces that create models that better connect data points of a class and its identity prototypes. We argue that this is important for training future models. It is not enough to create an interpretable prototype space; we should ensure that the prototypes are meaningful for the generated explanations. We also evaluated the level of coherence of the label of the prototypes when using a particular type of prototype layer. We found that the Deformable ProtoPNet generates prototypes that tend to be more consistent with their source labels. We argue that this is the preferred case because ensuring that prototypes of factual or counterfactual explanations belong to the same label (i.e., promoting sparsity) will make them easier to understand. While these models learn how to weigh the contributions of each prototype to the final prediction, they should pay attention to the sources of these prototypes and ensure that they have a proper meaning to the user.

Quality of the Counterfactuals Lastly, we analysed the level of coherence of the label of the counterfactuals generated by each architecture. In our approach, we solely rely on the classification layer of each architecture, meaning that we straightforwardly obtain counterfactuals. Nevertheless, we observe that the Baseline architectures seem to be the most coherent models concerning the counterfactuals, meaning that they are more predictable when compared to the other architectures (i.e., given a query image of a specific label, the user will know, with higher confidence, the source label of the counterfactual that these models will output). Once again, if we aim to increase the interpretability of our models, this is also a desired property.

6.6.4.2 Future Work

Future work should address the optimisation of prototype-based deep learning models in light of the above conclusion. For example, we must ensure that prototypes per image have a stronger connection to their class identities (although previous works have a term in the loss function that promotes this property, our results suggest that this may not be sufficient). These issues related to the quality of the learned prototypes are in agreement with recent literature. Therefore, we need to develop architectures or training strategies that promote robustness in prototype learning and complement these with an objective metric that allows practitioners to assess their quality.

As for counterfactual learning and retrieval, even though most state-of-the-art approaches focus on generative methods to produce the counterfactuals, we suggest exploring a different framework. While current methods perform local interventions on the inputs (e.g., changing feature values, altering image regions), we focus our efforts on retrieving the counterfactuals at a global level (i.e., getting an example from the database that contrasts better with our query input) and using its learned prototypes to explain this decision. Once again, we still lack objective metrics to assess the quality of these counterfactuals or a mean of comparison between generation and retrieval methods. The same applies to the latent space we use to retrieve our counterfactuals. In this work, we also suggest that a prototypical feature space may be an alternative or complement to the backbone feature space since it may encode local similarity information that may allow us to obtain similar cases that are more informative. On the other hand, we still lack objective metrics to assess these hypotheses.

We foresee a future research direction that focuses on studying adversarial interventions on the prototype layer of these architectures to gain more control over the type of prototypes learned by these frameworks. This aspect also contributes to the strategy of counterfactual retrieval and explanation. Our intuition is that by intervening on the prototype layer, we ensure that the retrieved prototypes may be more meaningful and respect some prior knowledge we may have on the application domain. Related to this topic, we point the reader to [690], whose recent work describes the vulnerabilities of counterfactual explanations and shows how we can manipulate them (e.g., through small perturbations).

Although not connected to the scope of this work in particular, we also suggest that research on prototypical learning for other tasks besides classification should be a priority. For instance, two databases used in this work (i.e., PAPILA and PH2) are also valid for ordinal classification tasks. Therefore, we will also devote efforts to alternative tasks such as ordinal classification, image segmentation, or bounding-box regression in object detection.

6.7 Summary and Conclusions

Developing interpretable algorithms by design is the ultimate goal for those who seek fully interpretable models. Understanding and defining the logical flow of the training and inference procedures of a machine learning algorithm makes it easier for developers to fine-tune or adapt their

models to users or tasks, better define the strengths and limitations of such models, and answer with more confidence to an external audit procedure by external stakeholders. Intuitively, we believe that the more information the user (developer, researcher or end-user) has about the models, the more trust they will have.

In this Chapter, we studied two types of inherently interpretable deep learning models. We dedicated the first two Sections to the B-cos networks [682], which replace all linear transforms of a deep neural network, to induce an alignment between the network's weights and task-relevant input patterns [682]. We evaluated the suitability of B-cos networks for medical applications and concluded that these models achieve more localised visual explanations (i.e., saliency maps) than their CNN counterparts. However, they also suffer from a slight drop in predictive performance, thus making it difficult to conclude that these models produce better overall explanations. While this might be an issue related to hyper-parameter optimisation, the internal mechanism of these models is still interesting for other tasks. With this in mind, we decided to extend the application of these models to the task of medical image retrieval, observing that these models attain similar performances to transformer-based architectures. A future line of research should be devoted to assessing the quality of the explanations of these models to complement and explain the retrieval results. In the third Section, we explored the paradigm of prototypical learning with the ProtoPNet [5] and the Deformable ProtoPNet architectures [656] that use a prototype layer to learn prototypical parts of the images generating a vector of similarity scores that are combined in the classification layer to output the final prediction. We aimed to assess the quality of the prototypes learned by these models. We explored the potential use of the models in counterfactual learning and retrieval, comparing different embedding spaces.

The main limitation of inherently interpretable models is their training process (requiring more time and effort on hyper-parameter tuning) and their predictive power (often drops compared to the baseline counterparts). On the other hand, assuming that we can reach a point where these models are fully optimised and have the same predictive capacity as the others, the paradigm might change into always using these, at least in applications related to high-stakes decisions (e.g., finance, healthcare, justice).

Chapter 7

Conclusion

7.1 Conclusions and Contributions

The spread and permeation of artificial intelligence (AI) into the quotidian happened (is happening) in a smooth (but disruptive) manner. The (apparent) seamless integration of these algorithms into several dimensions of people’s lives brought several advantages (e.g., quick and almost free access to large repositories of data or computational power, simplification of processes, and access to data-driven decision support systems). However, despite the innovative explosion brought by AI, society started to look at the risks and (potential) consequences of not regulating this realm, especially in applications that mostly require high-stakes decisions, where a mistake might change human life forever. The field of explainable artificial intelligence (xAI) appeared as a framework to guide researchers, developers, social scientists and policymakers toward designing, developing, deploying, monitoring and auditing algorithms that should be transparent, reliable and robust. This doctoral thesis sought to contribute to that body of knowledge.

Using the medical domain as the protagonist of this story, we tackled interpretability in AI for healthcare from three different points of view, self-contained but complementary (and, in some cases, sequential):

- Chapter 4: we studied the role of attention mechanisms in deep learning algorithms for healthcare applications. In this Chapter, we made fundamental and applied contributions. We started with an exhaustive review of attention-based mechanisms for healthcare, followed by a comparative study with different models and applications, to assess the potential benefits of using attention mechanisms and their impact on the visual explanations (i.e., saliency maps) generated by these models. With this work, we concluded that using attention mechanisms often increases model complexity, will not always lead to better performances and does not generate less ambiguous visual explanations. On the other hand, we could not base our judgement based only on these findings. As such, we decided to explore applications of these methods where they could have a potential positive impact. In medical image retrieval, we found that attention-driven models can learn embedding spaces that contain

more meaningful information about the pathology, allowing us to retrieve cases more similar to the query. In a computational pathology task, we confirmed the quality of feature extraction methods produced by attention mechanisms (in this case, a transformer) and their power as an aggregation operator in multiple instance learning (MIL).

- Chapter 5: we studied the role of multi-modal approaches for medical image retrieval tasks and computational pathology. In this Chapter, we made applied contributions. We proposed a model-agnostic framework that allows practitioners to optimise their models (and the learned features) using the triplet loss. We analysed the capacity of traditional machine learning algorithms, CNNs and ViTs for medical image retrieval under different settings (i.e., pre-trained, fine-tuned and multi-modal) and concluded that vision transformers (ViTs) with multi-modal information attained the best results. Later, we extended this study into attention-based mechanisms (to establish the bridge with the previous Chapter) and confirmed that ViTs perform well on medical image retrieval tasks, often benefitting from multi-modal information in the pipeline. We also increased the work on computational pathology (of the previous Chapter). We assessed the potential use of clinical data (combined with imaging data) in multi-modal algorithms to predict gene pathway expression. We concluded that a proper combination (and encoding) of a (specific) set of clinical features with WSI (imaging) information might contribute to algorithms with increased predictive power and transparency. Although preliminary, these results are promising and will pave the way for future research in computational pathology.
- Chapter 6: we studied the role of inherently interpretable deep learning models. In this Chapter, we made fundamental and applied contributions. We dedicated two Sections to the study of B-cos networks. We started by assessing the suitability of B-cos networks for medical applications, concluding that visual explanations obtained from these models are more localised than the ones obtained by their convolutional neural network (CNN) counterparts (although not necessarily better). In any case, these promising results pushed us to extend the application of these models to the task of medical image retrieval (and, once again, establishing the connection to the previous Chapters). Interestingly, these models attain competitive results when compared to the transformer-based architectures. In the last Section, we devoted our time to a line of research that is somehow self-contained but related to retrieval (connecting with previous Chapters) and, naturally, interpretability. In this last work, we experimented with the paradigm of prototypical learning with the prototypical part network (ProtoPNet) and Deformable ProtoPNet architectures, focusing on the quality assessment of the prototypes learned by these models and their potential use in counterfactual learning and retrieval.

Hoping to contribute to transparency and reproducibility in scientific research, we added a code repository to all the works included in the main story of this doctoral thesis (published or not).

Hence, we hope the curious reader, researcher or student takes the time to audit and (hopefully) improve all the methodologies described in this work.

7.2 Final Remarks and Future Work

The work described in this doctoral thesis is a journey through different dimensions of the realm of xAI. In this doctoral research, we explored diverse lines of research and tried to contribute to solving the challenges presented in the state-of-the-art. As the field of machine learning evolves at a super-human pace, the attentive reader will notice that, in the same Chapter, several published works (on the same topic) span across different years. Oppositely, others require extra experiments to achieve higher maturity and pass the sieve of demanding reviewers. Although we made incremental contributions from the fundamental perspective or proposed novel application from the applied perspective, this doctoral work devoted a significant amount of time to auditing and reproducing (often with extra difficulties) the work of other researchers in the same field. Although not exempt from personal frustration, this quasi-exhaustive exploration of the state-of-the-art allowed us to have more intuition on the inner workings of these methods and their potential applications. This thesis narrates a tale of achieving transparent and interpretable algorithms for healthcare applications. However, as shown by the collaborations with colleagues from other research fields or by the methodologies proposed, we ensure the reader that there is space to apply these methodologies in (virtually) any other domain.

Although each Chapter ends with a summary and different future work directions, the exercise of writing a doctoral thesis, revisiting previous work or leaving unfinished work behind (at least, for now) gave us space and time to reflect on different research lines or priorities that could derive from this thesis:

1. Multi-modal, multi-task and multi-label xAI: given the increase in data availability from different modalities, this field of machine learning research grows faster. However, we noticed that most works focus on a single task or try to explain a single modality. Hence, we believe that the community should now move into developing methodologies to create explanations that derive from multiple tasks (or labels) regarding different modalities. Understanding the reasoning process at this level might support research on multi-modal and inherently interpretable machine learning.
2. Fairness, equity, privacy or causality: some secondary works developed during this doctoral research briefly touched the topics of fairness, equity or (data) privacy. Causality is a common factor in these topics, but research in causal computer vision is still open. While the technical communities (computer science, engineering) might model causality with a specific type or number of assumptions, when talking to other communities (philosophy, law, clinical), we might notice some divergence in their understanding of these topics.

3. Deployment of algorithms and user-centric studies: the ideal process of development of xAI algorithms should mimic the traditional pipeline of software development where there is an explicit concern with the validation of the functionalities, the understanding and documentation of the limitations or edge cases of the applications. Moreover, this line also relates to evaluating the quality or relevance of model explanations (visual or not). By doing user studies, we could know better what is fundamental for specific domain communities and prepare further research protocols accordingly and in a data-driven manner. These studies take time to prepare, conduct and validate, but their benefits should be enough to overcome these challenges.

7.3 Funding

This work has received funding from the European Union’s Horizon Europe research and innovation programme under the Grant Agreement 101057389-CINDERELLA project and from the Portuguese Foundation for Science and Technology (FCT) through the PhD Grant “2020.06434.BD”.

References

- [1] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, “An Attentive Survey of Attention Models,” *ACM Trans. Intell. Syst. Technol.*, vol. 12, 10 2021.
- [2] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, “Attention U-Net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [3] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.
- [4] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.
- [5] 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*, vol. 32, 2019.
- [6] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning Important Features Through Propagating Activation Differences,” in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153, 2017.
- [7] 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 (ICCV)*, pp. 618–626, 2017.
- [8] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” in *ICML*, vol. 70, p. 3319–3328, 2017.
- [9] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. D. Fauw, and S. Shetty, “International evaluation of an AI system for breast cancer screening,” *Nature*, vol. 577, pp. 89–94, January 2020.
- [10] J. Su, D. V. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.
- [11] C. Rudin, “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

- [12] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine Learning Interpretability: A Survey on Methods and Metrics,” *Electronics*, vol. 8, no. 8, 2019.
- [13] B. Kim and F. Doshi-Velez, “Interpretable machine learning: the fuss, the concrete and the questions,” *ICML Tutorial on interpretable machine learning*, 2017.
- [14] J. W. Tukey, *Exploratory data analysis*, vol. 2. Reading, Mass., 1977.
- [15] K. R. Varshney, J. C. Rasmussen, A. Mojsilovic, M. Singh, and J. M. DiMicco, “Interactive Visual Salesforce Analytics,” in *ICIS*, 2012.
- [16] B. Kim, J. A. Shah, and F. Doshi-Velez, “Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction,” in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.
- [17] B. Kim, R. Khanna, and O. O. Koyejo, “Examples are not enough, learn to criticize! criticism for interpretability,” *Advances in neural information processing systems*, vol. 29, 2016.
- [18] Rivest, Ronald L., “Learning decision lists,” *Machine Learning*, vol. 2, pp. 229–246, November 1987.
- [19] C. Rudin and B. Ustun, “Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice,” *Interfaces*, vol. 48, no. 5, pp. 449–466, 2018.
- [20] 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, (Cambridge, MA, USA), p. 1952–1960, MIT Press, 2014.
- [21] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, “This looks like that: deep learning for interpretable image recognition,” *arXiv preprint arXiv:1806.10574*, 2018.
- [22] A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, and C. Rudin, “IAIA-BL: A Case-based Interpretable Deep Learning Model for Classification of Mass Lesions in Digital Mammography,” *arXiv preprint arXiv:2103.12308*, 2021.
- [23] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, May 2015.
- [24] 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.
- [25] 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 (ICML)*, pp. 2668–2677, PMLR, 2018.
- [26] Z. Chen, Y. Bei, and C. Rudin, “Concept whitening for interpretable image recognition,” *Nature Machine Intelligence*, vol. 2, no. 12, pp. 772–782, 2020.
- [27] T. Wang and C. Rudin, “Causal rule sets for identifying subgroups with enhanced treatment effect,” *arXiv preprint arXiv:1710.05426*, 2017.

- [28] K. Gan, A. Li, Z. Lipton, and S. Tayur, “Causal Inference with Selectively Deconfounded Data,” in *International Conference on Artificial Intelligence and Statistics*, pp. 2791–2799, PMLR, 2021.
- [29] 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.
- [30] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *International Conference on Machine Learning (ICML)*, pp. 3145–3153, PMLR, 2017.
- [31] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, pp. 818–833, Springer, 2014.
- [32] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” *arXiv preprint arXiv:1412.6806*, 2014.
- [33] 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*, pp. 1135–1144, 2016.
- [34] 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*, vol. 10, no. 7, p. e0130140, 2015.
- [35] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, “Explaining nonlinear classification decisions with deep taylor decomposition,” *Pattern Recognition*, vol. 65, pp. 211–222, 2017.
- [36] 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*, pp. 133–140, Springer, 2018.
- [37] 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)*, pp. 1–8, IEEE, 2019.
- [38] 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*, KDD ’16, (New York, NY, USA), p. 1135–1144, Association for Computing Machinery, 2016.
- [39] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, “Learning how to explain neural networks: PatternNet and PatternAttribution,” in *International Conference on Learning Representations*, 2018.
- [40] B. Mittelstadt, C. Russell, and S. Wachter, “Explaining Explanations in AI,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* ’19, (New York, NY, USA), p. 279–288, Association for Computing Machinery, 2019.

- [41] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, “Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’20, (New York, NY, USA), p. 180–186, Association for Computing Machinery, 2020.
- [42] S. Wickramanayake, W. Hsu, and M. L. Lee, “Towards Fully Interpretable Deep Neural Networks: Are We There Yet?,” *arXiv preprint arXiv:2106.13164*, 2021.
- [43] 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*, vol. 10, no. 4, pp. 441–455, 2021.
- [44] United Nations Department of Economic and Social Affairs Sustainable Development, “The 17 Goals Sustainable Development.” <https://sdgs.un.org/goals>, 2025. [Online. Accessed 22-January-2025].
- [45] United Nations Department of Economic and Social Affairs Sustainable Development, “Transforming our world: the 2030 Agenda for Sustainable Development.” <https://sdgs.un.org/2030agenda>, 2025. [Online. Accessed 22-January-2025].
- [46] Mohammad Hossein Zolfagharnasab and Nuno Freitas and Tiago Goncalves and Eduard Bonci and Carlos Mavioso and Maria J. Cardoso and Helder P. Oliveira and Jaime S. Cardoso, “Predicting Aesthetic Outcomes in Breast Cancer Surgery: A Multimodal Retrieval Approach,” in *Proceedings of the Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care (in MICCAI 2024)*, 2024.
- [47] I. Rio-Torto, T. Gonçalves, J. S. Cardoso, and L. F. Teixeira, “On the Suitability of B-cos Networks for the Medical Domain,” in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2024.
- [48] T. Gonçalves, D. Pulido-Arias, J. Willett, K. V. Hoebel, M. Cleveland, S. R. Ahmed, E. Gerstner, J. Kalpathy-Cramer, J. S. Cardoso, C. P. Bridge, and A. E. Kim, “Deep Learning-based Prediction of Breast Cancer Tumor and Immune Phenotypes from Histopathology,” 2024.
- [49] 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*, vol. 12, no. 1, p. 20732, 2022.
- [50] T. Gonçalves, I. Rio-Torto, L. F. Teixeira, and J. S. Cardoso, “A Survey on Attention Mechanisms for Medical Applications: Are We Moving Toward Better Algorithms?,” *IEEE Access*, vol. 10, pp. 98909–98935, 2022.
- [51] T. Gonçalves, L. Capozzi, A. Rebelo, and J. S. Cardoso, “Deep Image Segmentation based on Mutual Information: A Study,” in *30th Portuguese Conference in Pattern Recognition (RECPAD)*, (Covilhã, Portugal), 2024.
- [52] L. Capozzi, T. Gonçalves, J. S. Cardoso, and A. Rebelo, “Towards Fingerprint Presentation Attack Generation using Generative Adversarial Networks,” in *30th Portuguese Conference in Pattern Recognition (RECPAD)*, (Covilhã, Portugal), 2024.

- [53] M. H. Zolfagharnasab, N. Freitas, T. Gonçalves, E. Bonci, C. Mavioso, M. J. Cardoso, H. P. Oliveira, and J. S. Cardoso, “Multimodal Deep Learning for Predicting Aesthetic Outcomes in Breast Cancer Surgery,” in *30th Portuguese Conference in Pattern Recognition (RECPAD)*, (Covilhã, Portugal), 2024.
- [54] T. Gonçalves, L. Capozzi, A. Rebelo, and J. S. Cardoso, “Optimisation of Deep Neural Networks using a Genetic Algorithm: A Comparative Study,” in *29th Portuguese Conference in Pattern Recognition (RECPAD)*, (Coimbra, Portugal), 2023.
- [55] L. Capozzi, T. Gonçalves, J. S. Cardoso, and A. Rebelo, “Analysis of Neurons’ Information in Deep Spiking Neural Networks using Information Theory,” in *29th Portuguese Conference in Pattern Recognition (RECPAD)*, (Coimbra, Portugal), 2023.
- [56] I. Rio-Torto, C. Patrício, H. Montenegro, T. Gonçalves, and J. S. Cardoso, “Detecting Concepts and Generating Captions from Medical Images,” in *29th Portuguese Conference in Pattern Recognition (RECPAD)*, (Coimbra, Portugal), 2023.
- [57] H. Montenegro, P. Neto, C. Patrício, I. Rio-Torto, T. Gonçalves, and L. F. Teixeira, “Evaluating Privacy on Synthetic Images Obtained using Deep Generative Models,” in *29th Portuguese Conference in Pattern Recognition (RECPAD)*, (Coimbra, Portugal), 2023.
- [58] E. Caldeira, E. M. Castro, and T. Gonçalves, “From Easy to Hard: A Curriculum Learning Approach for Breast Lesion Classification,” in *28th Portuguese Conference in Pattern Recognition (RECPAD)*, (Leiria, Portugal), 2022.
- [59] R. C. Diana Teixeira e Silva and T. Gonçalves, “Preliminary Study on Deep Iterative Semantic Segmentation,” in *28th Portuguese Conference in Pattern Recognition (RECPAD)*, (Leiria, Portugal), 2022.
- [60] T. Gonçalves and J. S. Cardoso, “Preliminary Study on the Impact of Attention Mechanisms for Medical Image Classification,” in *27th Portuguese Conference in Pattern Recognition (RECPAD)*, (Évora, Portugal), 2021.
- [61] T. Gonçalves, A. Hedström, A. Pahud de Mortanges, X. Li, H. Müller, J. S. Cardoso, and M. Reyes, “Chapter 15 - Interpretable AI for medical image analysis: methods, evaluation, and clinical considerations,” in *Trustworthy AI in Medical Imaging* (M. Lorenzi and M. A. Zuluaga, eds.), The MICCAI Society book Series, pp. 315–346, Academic Press, 2025.
- [62] M. Vieira, T. Gonçalves, W. Silva, and A. F. Sequeira, “An End-to-End Framework to Classify and Generate Privacy-Preserving Explanations in Pornography Detection,” in *2024 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–6, 2024.
- [63] M. M. Beirão, J. Matos, T. Gonçalves, C. Kase, L. F. Nakayama, D. de Freitas, and J. S. Cardoso, “Classification of Keratitis from Eye Corneal Photographs using Deep Learning,” in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3060–3065, 2024.
- [64] P. C. Neto, R. M. Mamede, C. Albuquerque, T. Gonçalves, and A. F. Sequeira, “Massively Annotated Datasets for Assessment of Synthetic and Real Data in Face Recognition,” 2024.
- [65] Inês Martins and João Matos and Tiago Goncalves and Leo A. Celi and An-Kwok Ian Wong and Jaime S. Cardoso, “Evaluating the Impact of Pulse Oximetry Bias in Machine Learning

- under Counterfactual Thinking,” in *Proceedings of the Third Workshop on Applications of Medical Artificial Intelligence (AMAI) (in MICCAI 2024)*, 2024.
- [66] Eduarda Caldeira and Pedro C. Neto and Tiago Gonçalves and Naser Damer and Ana F. Sequeira and Jaime S. Cardoso, “Disentangling Morphed Identities for Face Morphing Detection,” *Science Talks*, vol. 10, p. 100331, 2024.
- [67] R. Cruz, D. T. e. Silva, T. Gonçalves, D. Carneiro, and J. S. Cardoso, “Two-Stage Framework for Faster Semantic Segmentation,” *Sensors*, vol. 23, no. 6, 2023.
- [68] E. Castro, P. M. Ferreira, A. Rebelo, I. Rio-Torto, L. Capozzi, M. F. Ferreira, T. Gonçalves, T. Albuquerque, W. Silva, C. Afonso, *et al.*, “Fill in the blank for fashion complementary outfit product Retrieval: VISUM summer school competition,” *Machine Vision and Applications*, vol. 34, no. 1, p. 16, 2023.
- [69] O. Kaidar-Person, M. Antunes, J. S. Cardoso, O. Ciani, H. Cruz, R. Di Micco, O. D. Gentilini, T. Gonçalves, P. Gouveia, J. Heil, P. Kabata, D. Lopes, M. Martinho, H. Martins, C. Mavioso, M. Mika, H. Montenegro, H. P. Oliveira, A. Pfob, N. Rotmensz, T. Schinköthe, G. Silva, R. Tarricone, M.-J. Cardoso, and on behalf of the CINDERELLA Consortium, “Evaluating the ability of an artificial-intelligence cloud-based platform designed to provide information prior to locoregional therapy for breast cancer in improving patient’s satisfaction with therapy: The CINDERELLA trial,” *PLOS ONE*, vol. 18, pp. 1–12, 08 2023.
- [70] P. Serrano e Silva, R. Cruz, A. S. M. Shihavuddin, and T. Gonçalves, “Interpretability-Guided Human Feedback During Neural Network Training,” in *Pattern Recognition and Image Analysis* (A. Pertusa, A. J. Gallego, J. A. Sánchez, and I. Domingues, eds.), (Cham), pp. 276–287, Springer Nature Switzerland, 2023.
- [71] E. Caldeira, P. C. Neto, T. Gonçalves, N. Damer, A. F. Sequeira, and J. S. Cardoso, “Unveiling the Two-Faced Truth: Disentangling Morphed Identities for Face Morphing Detection,” in *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 955–959, 2023.
- [72] H. Montenegro, P. Neto, C. Patrício, I. Rio-Torto, T. Gonçalves, and L. F. Teixeira, “Evaluating Privacy on Synthetic Images Generated using GANs: Contributions of the VCMi Team to ImageCLEFmedical GANs 2023,” *Challenge*, p. 8, 2023.
- [73] I. Rio-Torto, C. Patrício, H. Montenegro, T. Gonçalves, and J. S. Cardoso, “Detecting concepts and generating captions from medical images: Contributions of the VCMi team to Image-CLEFmedical caption 2023,” in *CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Thessaloniki, Greece, 2023*.
- [74] D. T. e Silva, R. Cruz, T. Gonçalves, and D. Carneiro, “Two-stage semantic segmentation in neural networks,” in *Fifteenth International Conference on Machine Vision (ICMV 2022)* (W. Osten, D. P. Nikolaev, and J. J. Zhou, eds.), vol. 12701, p. 127010G, International Society for Optics and Photonics, SPIE, 2023.
- [75] M. Huber, F. Boutros, A. T. Luu, K. Raja, R. Ramachandra, N. Damer, P. C. Neto, T. Gonçalves, A. F. Sequeira, J. S. Cardoso, J. Tremoço, M. Lourenço, S. Serra, E. Cermeño, M. Ivanovska, B. Batagelj, A. Kronovšek, P. Peer, and V. Štruc, “SYN-MAD 2022: Competition on Face Morphing Attack Detection Based on Privacy-aware Synthetic Training Data,” in *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10, 2022.

- [76] P. C. Neto, T. Gonçalves, M. Huber, N. Damer, A. F. Sequeira, and J. S. Cardoso, “OrthoMAD: Morphing Attack Detection Through Orthogonal Identity Disentanglement,” in *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–5, 2022.
- [77] I. Rio-Torto, C. Patrício, H. Montenegro, and T. Gonçalves, “Detecting Concepts and Generating Captions from Medical Images: Contributions of the VCMi Team to ImageCLEFmedical 2022 Caption,” 2022.
- [78] P. Maia, J. Morgado, T. Gonçalves, and T. Albuquerque, “Applying Machine Learning for Traffic Forecasting in Porto, Portugal,” in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (M. Kamp, I. Koprinska, A. Bibal, T. Bouadi, B. Frénay, L. Galárraga, J. Oramas, L. Adilova, Y. Krishnamurthy, B. Kang, C. Largeron, J. Lijffijt, T. Viard, P. Welke, M. Ruocco, E. Aune, C. Gallicchio, G. Schiele, F. Pernkopf, M. Blott, H. Fröning, G. Schindler, R. Guidotti, A. Monreale, S. Rinzivillo, P. Biecek, E. Ntoutsis, M. Pechenizkiy, B. Rosenhahn, C. Buckley, D. Cialfi, P. Lanillos, M. Ramstead, T. Verbelen, P. M. Ferreira, G. Andresini, D. Malerba, I. Medeiros, P. Fournier-Viger, M. S. Nawaz, S. Ventura, M. Sun, M. Zhou, V. Bitetta, I. Bordino, A. Ferretti, F. Gullo, G. Ponti, L. Severini, R. Ribeiro, J. Gama, R. Gavaldà, L. Cooper, N. Ghazaleh, J. Richiardi, D. Roqueiro, D. Saldana Miranda, K. Sechidis, and G. Graça, eds.), (Cham), pp. 299–308, Springer International Publishing, 2021.
- [79] T. Schinköthe, E. Bonci, K. Orit, H. Cruz, R. Di Micco, O. Gentilini, J. Heil, P. Kabata, M. Romariz, T. Gonçalves, *et al.*, “The CINDERELLA APProach: Future Concepts for Patient Empowerment in Breast Cancer Treatment with Artificial Intelligence-Driven Healthcare Platform,” *European Journal of Cancer*, vol. 200, 2024.
- [80] E.-A. Bonci, O. Kaidar-Person, M. Antunes, O. Ciani, H. Cruz, R. Di Micco, O. Gentilini, P. Gouveia, J. Heil, P. Kabata, *et al.*, “CINDERELLA Trial: validation of an artificial-intelligence cloud-based platform to improve the shared decision-making process and outcomes in breast cancer patients proposed for locoregional treatment,” *European Journal of Surgical Oncology*, vol. 50, no. 2, 2024.
- [81] A. Pfob, E. Bonci, O. Kaidar-Person, M. Antunes, O. Ciani, H. Cruz, R. Di Micco, O. Gentilini, J. Heil, P. Kabata, *et al.*, “108TiP CINDERELLA clinical trial: Using artificial intelligence-driven healthcare to enhance breast cancer locoregional treatment decisions,” *ESMO Open*, vol. 9, 2024.
- [82] E.-A. Bonci, O. Kaidar-Person, M. Antunes, O. Ciani, H. Cruz, R. D. Micco, O. D. Gentilini, N. Rotmensz, P. Gouveia, J. Heil, P. Kabata, N. Freitas, T. Gonçalves, M. Romariz, H. Montenegro, H. P. Oliveira, J. S. Cardoso, H. Martins, D. Lopes, M. Martinho, L. Borsoi, E. Listorti, C. Mavioso, M. Mika, A. Pfob, T. Schinköthe, G. Silva, and M.-J. Cardoso, “Abstract PO3-19-11: CINDERELLA Clinical Trial (NCT05196269): using artificial intelligence-driven healthcare to enhance breast cancer locoregional treatment decisions,” *Cancer Research*, vol. 84, pp. PO3–19–11–PO3–19–11, 05 2024.
- [83] E.-A. Bonci, O. Kaidar-Person, M. Antunes, O. Ciani, H. Cruz, R. Di Micco, O. D. Gentilini, J. Heil, P. Kabata, M. Romariz, T. Gonçalves, H. Martins, L. Borsoi, M. Mika, A. Pfob, N. Romem, T. Schinkoethe, G. Silva, M. Bobowicz, M.-J. Cardoso, and n. null, “CINDERELLA clinical trial: Using artificial intelligence-driven healthcare to enhance breast

- cancer locoregional treatment decisions,” *Journal of Clinical Oncology*, vol. 42, no. 16_suppl, pp. TPS621–TPS621, 2024. PMID:.
- [84] Shreyas Bhat Brahmavar and Tiago Goncalves and Tobias R. Bodenmann and Syed Rakin Ahmed and Jay B. Patel and Praveer Singh and Katharina V. Hoebel and Mason C. Cleveland and Felix Dorfner and Dagoberto Pulido-Arias and Philipp Hahnel and Jaime S. Cardoso and Jayashree Kalpathy-Cramer and Priscilla Brastianos and Elizabeth Gerstner and Albert E. Kim and Christopher P. Bridge, “Deep Learning-based Non-Invasive Molecular Profiling of Brain Metastases from MR Imaging,” 2024.
 - [85] I. Martins, J. Matos, T. Goncalves, L. A. Celi, A. I. Wong, and J. S. Cardoso, “Thermometry Bias in Critical Care and Impact in Downstream Machine Learning Tasks,” 2024.
 - [86] M. Beirao, J. Matos, T. Goncalves, C. Kase, L. F. Nakayama, and J. S. Cardoso, “A machine learning system for keratitis diagnosis from cornea images,” 2024.
 - [87] D. P. Arias, T. Goncalves, J. Kalpathy-Cramer, E. Gerstner, J. S. Cardoso, A. Kim, and C. P. Bridge, “Dissecting the Impact of Data Augmentation on Whole Slide Image Classification,” 2024.
 - [88] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory*. Psychology Press, 2005.
 - [89] McCulloch, Warren S and Pitts, Walter, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.
 - [90] B. Farley and W. Clark, “Simulation of self-organizing systems by digital computer,” *Transactions of the IRE Professional Group on Information Theory*, vol. 4, no. 4, pp. 76–84, 1954.
 - [91] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain,” *Psychological review*, vol. 65, no. 6, p. 386, 1958.
 - [92] Kurt Hornik and Maxwell Stinchcombe and Halbert White, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
 - [93] Kurt Hornik, “Approximation capabilities of multilayer feedforward networks,” *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.
 - [94] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *nature*, vol. 323, no. 6088, pp. 533–536, 1986.
 - [95] R. Raina, A. Madhavan, and A. Y. Ng, “Large-scale deep unsupervised learning using graphics processors,” in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML ’09*, (New York, NY, USA), p. 873–880, Association for Computing Machinery, 2009.
 - [96] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, “Deep learning for visual understanding: A review,” *Neurocomputing*, vol. 187, pp. 27–48, 2016. Recent Developments on Deep Big Vision.

- [97] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [98] M. K. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey, “Deep learning of the tissue-regulated splicing code,” *Bioinformatics*, vol. 30, pp. i121–i129, 06 2014.
- [99] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, “Deep neural nets as a method for quantitative structure–activity relationships,” *Journal of Chemical Information and Modeling*, vol. 55, no. 2, pp. 263–274, 2015.
- [100] Kunihiro Fukushima, “Neocognitron: A hierarchical neural network capable of visual pattern recognition,” *Neural Networks*, vol. 1, no. 2, pp. 119–130, 1988.
- [101] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [102] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [103] A. Mumuni and F. Mumuni, “Data augmentation: A comprehensive survey of modern approaches,” *Array*, vol. 16, p. 100258, 2022.
- [104] Alexander Kolesnikov and Alexey Dosovitskiy and Dirk Weissenborn and Georg Heigold and Jakob Uszkoreit and Lucas Beyer and Matthias Minderer and Mostafa Dehghani and Neil Houlsby and Sylvain Gelly and Thomas Unterthiner and Xiaohua Zhai, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [105] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.
- [106] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.
- [107] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.
- [108] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Networks,” *Commun. ACM*, vol. 63, p. 139–144, October 2020.
- [109] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” 2022.
- [110] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, *et al.*, “Accurate structure prediction of biomolecular interactions with AlphaFold 3,” *Nature*, pp. 1–3, 2024.

- [111] D. Valevski, Y. Leviathan, M. Arar, and S. Fruchter, “Diffusion Models Are Real-Time Game Engines,” 2024.
- [112] D. J. Mankowitz, A. Michi, A. Zhernov, M. Gelmi, M. Selvi, C. Paduraru, E. Leurent, S. Iqbal, J.-B. Lespiau, A. Ahern, *et al.*, “Faster sorting algorithms discovered using deep reinforcement learning,” *Nature*, vol. 618, no. 7964, pp. 257–263, 2023.
- [113] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, *et al.*, “Mastering atari, go, chess and shogi by planning with a learned model,” *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.
- [114] S. Mohseni, N. Zarei, and E. D. Ragan, “A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems,” *arXiv:1811.11839 [cs]*, December 2019. arXiv: 1811.11839.
- [115] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, “What do we need to build explainable AI systems for the medical domain?,” *arXiv:1712.09923 [cs, stat]*, December 2017. arXiv: 1712.09923.
- [116] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining Explanations: An Overview of Interpretability of Machine Learning,” *arXiv:1806.00069 [cs, stat]*, February 2019. arXiv: 1806.00069.
- [117] C. Molnar, *Interpretable Machine Learning*. Christoph Molnar, 2020.
- [118] Z. C. Lipton, “The Mythos of Model Interpretability,” *arXiv:1606.03490 [cs, stat]*, Mar. 2017. arXiv: 1606.03490.
- [119] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” *arXiv:1702.08608 [cs, stat]*, Mar. 2017. arXiv: 1702.08608.
- [120] Q. Zhang, Y. N. Wu, and S.-C. Zhu, “Interpretable Convolutional Neural Networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8827–8836, 2018.
- [121] H. Jiang, B. Kim, M. Guan, and M. Gupta, “To Trust Or Not To Trust A Classifier,” in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.
- [122] D. Alvarez Melis and T. Jaakkola, “Towards Robust Interpretability with Self-Explaining Neural Networks,” in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.
- [123] N. Papernot and P. McDaniel, “Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning,” *arXiv preprint arXiv:1803.04765*, 2018.
- [124] I. Rio-Torto, K. Fernandes, and L. F. Teixeira, “Towards a Joint Approach to Produce Decisions and Explanations Using CNNs,” in *Pattern Recognition and Image Analysis* (A. Morales, J. Fierrez, J. S. Sánchez, and B. Ribeiro, eds.), (Cham), pp. 3–15, Springer International Publishing, 2019.

- [125] I. Rio-Torto, K. Fernandes, and L. F. Teixeira, “Understanding the decisions of CNNs: An in-model approach,” *Pattern Recognition Letters*, vol. 133, pp. 373–380, 2020.
- [126] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv:1312.6034 [cs]*, April 2014. arXiv: 1312.6034.
- [127] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “SmoothGrad: removing noise by adding noise,” *arXiv:1706.03825 [cs, stat]*, June 2017. arXiv: 1706.03825.
- [128] J. Zhou, F. Chen, and A. Holzinger, “Towards Explainability for AI Fairness,” in *XxAI - Beyond Explainable AI*, p. 375–386, 2020.
- [129] K. Fernandes, J. S. Cardoso, and B. S. Astrup, “A deep learning approach for the forensic evaluation of sexual assault,” *Pattern Analysis and Applications*, vol. 21, pp. 629–640, August 2018.
- [130] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” *arXiv:1311.2901 [cs]*, November 2013. arXiv: 1311.2901.
- [131] A. Dosovitskiy and T. Brox, “Inverting Visual Representations with Convolutional Networks,” *arXiv:1506.02753 [cs]*, Apr. 2016. arXiv: 1506.02753.
- [132] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, “Network Dissection: Quantifying Interpretability of Deep Visual Representations,” *arXiv:1704.05796 [cs]*, April 2017. arXiv: 1704.05796.
- [133] B. Mittelstadt, C. Russell, and S. Wachter, “Explaining explanations in AI,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 279–288, 2019.
- [134] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, J. Adebayo, M. D. Li, and J. Kalpathy-Cramer, “Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging,” *Radiology: Artificial Intelligence*, vol. 3, no. 6, p. e200267, 2021.
- [135] 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* (D. Stoyanov, Z. Taylor, S. M. Kia, I. Oguz, M. Reyes, A. Martel, L. Maier-Hein, A. F. Marquand, E. Duchesnay, T. Löfstedt, B. Landman, M. J. Cardoso, C. A. Silva, S. Pereira, and R. Meier, eds.), vol. 11038, pp. 133–140, Cham: Springer International Publishing, 2018.
- [136] 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)*, pp. 1–8, July 2019. ISSN: 2161-4393.
- [137] NHS England, “Diabetic eye screening: professional guidance.” <https://www.gov.uk/government/collections/diabetic-eye-screening-commission-and-provide>, 2025. [Online. Accessed 22-January-2025].

- [138] Karthik, Maggie, and S. Dane, "APTOS 2019 Blindness Detection." <https://kaggle.com/competitions/aptos2019-blindness-detection>, 2019. Accessed: 2024-09-18.
- [139] Cinderella Project, "Cinderella Project - Artificial Intelligence." <https://cinderellaproject.eu>, 2025. [Online. Accessed 22-January-2025].
- [140] M. A. Rose, I. Olivotto, B. Cady, C. Koufman, R. Osteen, B. Silver, A. Recht, and J. R. Harris, "Conservative Surgery and Radiation Therapy for Early Breast Cancer: Long-term Cosmetic Results," *Archives of Surgery*, vol. 124, pp. 153–157, 02 1989.
- [141] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," *Computation & Neural Systems Technical Report*, 2011.
- [142] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear Benchmark Data For Pattern Classification," *Nature inspired Smart Information Systems (NiSIS)*, pp. 1–9, 2005.
- [143] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv:1710.05006*, 2018.
- [144] P. Tschandl, C. Rosendahl, and H. Kittler, "Data descriptor: the HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, 2018.
- [145] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, *et al.*, "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv:1902.03368*, 2019.
- [146] ISIC Challenge, "ISIC 2018 Challenge." <https://challenge.isic-archive.com/landing/2018/>, 2025. [Online. Accessed 22-January-2025].
- [147] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, *et al.*, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific data*, vol. 8, no. 1, pp. 1–8, 2021.
- [148] A. Zawacki, B. Helba, G. Shih, J. Weber, J. Elliott, M. Combalia, N. Kurtansky, Noel-Codella, P. Culliton, and V. Rotemberg, "SIIM-ISIC Melanoma Classification." <https://kaggle.com/competitions/siim-isic-melanoma-classification>, 2020. [Online. Accessed 22-January-2025].
- [149] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, pp. 1–8, 2019.
- [150] O. Kovalyk, J. Morales-Sánchez, R. Verdú-Monedero, I. Sellés-Navarro, A. Palazón-Cabanes, and J.-L. Sancho-Gómez, "PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment," *Scientific Data*, vol. 9, no. 1, p. 291, 2022.

- [151] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. 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)*, pp. 5437–5440, IEEE, 2013.
- [152] W. Lingle, B. J. Erickson, M. L. Zuley, R. Jarosz, E. Bonaccio, J. Filippini, J. M. Net, L. Levi, E. A. Morris, G. G. Figler, P. Elnajjar, S. Kirk, Y. Lee, M. Giger, and N. Gruszauskas, “The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA) (Version 3) [Data set]. The Cancer Imaging Archive..” <https://doi.org/10.7937/K9/TCIA.2016.AB2NAZRP>, 2016. Accessed: 2023-11-28.
- [153] National Cancer Institute, “Genomic Data Commons.” <https://gdc.cancer.gov>, 2025. [Online. Accessed 22-January-2025].
- [154] National Cancer Institute, “The Cancer Imaging Archive.” <https://www.cancerimagingarchive.net>, 2025. [Online. Accessed 22-January-2025].
- [155] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. T. Tong, D. H. Dinh, C. D. Do, L. T. Doan, C. N. Nguyen, B. T. Nguyen, Q. V. Nguyen, A. D. Hoang, H. N. Phan, A. T. Nguyen, P. H. Ho, D. T. Ngo, N. T. Nguyen, N. T. Nguyen, M. Dao, and V. Vu, “VinDr-CXR: An open dataset of chest X-rays with radiologist’s annotations,” *Scientific Data*, vol. 9, no. 1, p. 429, 2022.
- [156] C. Koch and S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry,” in *Matters of intelligence*, pp. 115–141, Springer, 1987.
- [157] L. Itti and C. Koch, “A saliency-based search mechanism for overt and covert shifts of visual attention,” *Vision research*, vol. 40, no. 10-12, pp. 1489–1506, 2000.
- [158] S. Minut and S. Mahadevan, “A reinforcement learning model of selective visual attention,” in *Proceedings of the fifth international conference on Autonomous agents*, pp. 457–464, 2001.
- [159] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, “Modeling visual attention via selective tuning,” *Artificial intelligence*, vol. 78, no. 1-2, pp. 507–545, 1995.
- [160] D. H. Ballard, “Animate vision,” *Artificial intelligence*, vol. 48, no. 1, pp. 57–86, 1991.
- [161] C. Bandera, F. J. Vico, J. M. Bravo, M. E. Harmon, and L. C. Baird, “Residual Q-learning applied to visual attention,” in *ICML*, pp. 20–27, Citeseer, 1996.
- [162] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Networks*, vol. 61, pp. 85–117, January 2015.
- [163] S. Jitley, N. A. Lord, N. Lee, and P. Torr, “Learn to pay attention,” in *International Conference on Learning Representations*, 2018.
- [164] I. Makarov, M. Bakhanova, S. Nikolenko, and O. Gerasimova, “Self-supervised recurrent depth estimation with attention mechanisms,” *PeerJ Computer Science*, vol. 8, p. e865, 2022.
- [165] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2012.

- [166] K. Cho, A. Courville, and Y. Bengio, “Describing multimedia content using attention-based encoder-decoder networks,” *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.
- [167] 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.
- [168] X. Yang, “An Overview of the Attention Mechanisms in Computer Vision,” in *Journal of Physics: Conference Series*, vol. 1693, p. 012173, IOP Publishing, 2020.
- [169] G. W. Lindsay, “Attention in psychology, neuroscience, and machine learning,” *Frontiers in computational neuroscience*, vol. 14, p. 29, 2020.
- [170] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.
- [171] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, *et al.*, “A survey on visual transformer,” *arXiv preprint arXiv:2012.12556*, 2020.
- [172] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in vision: A survey,” *ACM Comput. Surv.*, 12 2021.
- [173] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, “Attention mechanisms in computer vision: A survey,” *Computational Visual Media*, 3 2022.
- [174] Y. Xu, H. Wei, M. Lin, Y. Deng, K. Sheng, M. Zhang, F. Tang, W. Dong, F. Huang, and C. Xu, “Transformers in computational visual media: A survey,” *Computational Visual Media*, vol. 8, no. 1, pp. 33–62, 2022.
- [175] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, “Transformers in medical imaging: A survey,” *arXiv preprint arXiv:2201.09873*, 2022.
- [176] M. M. Chun, J. D. Golomb, and N. B. Turk-Browne, “A taxonomy of external and internal attention,” *Annual review of psychology*, vol. 62, pp. 73–101, 2011.
- [177] B. S. Oken, M. C. Salinsky, and S. Elsas, “Vigilance, alertness, or sustained attention: physiological basis and measurement,” *Clinical neurophysiology*, vol. 117, no. 9, pp. 1885–1901, 2006.
- [178] N. Kanwisher and E. Wojciulik, “Visual attention: insights from brain imaging,” *Nature reviews neuroscience*, vol. 1, no. 2, pp. 91–100, 2000.
- [179] A. W. Bronkhorst, “The cocktail-party problem revisited: early processing and selection of multi-talker speech,” *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, 2015.
- [180] F. Huttmacher, “Why is there so much more research on vision than on any other sensory modality?,” *Frontiers in psychology*, vol. 10, p. 2246, 2019.
- [181] E. K. Miller and T. J. Buschman, “Neural Mechanisms for the Executive Control of Attention,” in *The Oxford Handbook of Attention*, Oxford University Press, 01 2014.

- [182] M. Aly and N. B. Turk-Browne, “How hippocampal memory shapes, and is shaped by, attention,” in *The hippocampus from cells to systems*, pp. 369–403, Springer, 2017.
- [183] A. Graves, G. Wayne, and I. Danihelka, “Neural turing machines,” *arXiv preprint arXiv:1410.5401*, 2014.
- [184] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, IEEE, 2013.
- [185] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in *International Conference on Machine Learning*, pp. 1764–1772, PMLR, 2014.
- [186] I. Sutskever, J. Martens, and G. Hinton, “Generating Text with Recurrent Neural Networks,” in *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML’11*, (Madison, WI, USA), p. 1017–1024, Omnipress, 2011.
- [187] A. Graves, “Generating sequences with recurrent neural networks,” *arXiv preprint arXiv:1308.0850*, 2013.
- [188] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.
- [189] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” in *3rd International Conference on Learning Representations, (ICLR) 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.
- [190] T. Luong, H. Pham, and C. D. Manning, “Effective Approaches to Attention-based Neural Machine Translation,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (Lisbon, Portugal), pp. 1412–1421, Association for Computational Linguistics, September 2015.
- [191] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-Based Models for Speech Recognition,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS’15*, (Cambridge, MA, USA), p. 577–585, MIT Press, 2015.
- [192] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep Voice 3: 2000-Speaker Neural Text-to-Speech,” in *International Conference on Learning Representations*, 2018.
- [193] S. Sukhbaatar, E. Grave, G. Lample, H. Jegou, and A. Joulin, “Augmenting self-attention with persistent memory,” *arXiv preprint arXiv:1907.01470*, 2019.
- [194] S. Sukhbaatar, E. Grave, P. Bojanowski, and A. Joulin, “Adaptive Attention Span in Transformers,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 331–335, Association for Computational Linguistics, July 2019.
- [195] A. Fan, T. Lavril, E. Grave, A. Joulin, and S. Sukhbaatar, “Addressing some limitations of transformers with feedback memory,” *arXiv preprint arXiv:2002.09402*, 2020.

- [196] S. Mehta, M. Ghazvininejad, S. Iyer, L. Zettlemoyer, and H. Hajishirzi, “DeLighT: Deep and Light-weight Transformer,” in *International Conference on Learning Representations*, 2021.
- [197] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, “Linformer: Self-attention with linear complexity,” *arXiv preprint arXiv:2006.04768*, 2020.
- [198] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” *arXiv preprint arXiv:2004.05150*, 2020.
- [199] N. Kitaev, L. Kaiser, and A. Levskaya, “Reformer: The Efficient Transformer,” in *International Conference on Learning Representations*, 2020.
- [200] Y. Tay, D. Bahri, D. Metzler, D.-C. Juan, Z. Zhao, and C. Zheng, “Synthesizer: Rethinking Self-Attention for Transformer Models,” in *International Conference on Machine Learning*, pp. 10183–10192, PMLR, 2021.
- [201] S. Merity, “Single headed attention rnn: Stop thinking with your head,” *arXiv preprint arXiv:1911.11423*, 2019.
- [202] 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 (ICML)*, pp. 2048–2057, PMLR, 2015.
- [203] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, “Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5659–5667, 2017.
- [204] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “Gcnet: Non-local networks meet squeeze-excitation networks and beyond,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.
- [205] N. Quader, M. M. I. Bhuiyan, J. Lu, P. Dai, and W. Li, “Weight Excitation: Built-in Attention Mechanisms in Convolutional Neural Networks,” in *European Conference on Computer Vision*, pp. 87–103, Springer, 2020.
- [206] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, “Attention augmented convolutional networks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3286–3295, 2019.
- [207] P. Wang, L. Liu, C. Shen, Z. Huang, A. Van Den Hengel, and H. Tao Shen, “Multi-attention network for one shot learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2721–2729, 2017.
- [208] V. Mnih, N. Heess, A. Graves, *et al.*, “Recurrent models of visual attention,” *Advances in neural information processing systems*, vol. 27, 2014.
- [209] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 764–773, 2017.
- [210] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

- [211] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," *Advances in neural information processing systems*, vol. 31, 2018.
- [212] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proceedings of the IEEE international conference on computer vision*, pp. 4733–4742, 2017.
- [213] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1169–1178, 2018.
- [214] R. Zhang, J. Li, H. Sun, Y. Ge, P. Luo, X. Wang, and L. Lin, "Scan: Self-and-collaborative attention network for video person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4870–4882, 2019.
- [215] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," *Advances in neural information processing systems*, vol. 28, 2015.
- [216] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 510–519, 2019.
- [217] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.
- [218] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11030–11039, 2020.
- [219] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 267–283, 2018.
- [220] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, pp. 7354–7363, PMLR, 2019.
- [221] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1347–1360, 2017.
- [222] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.
- [223] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical lstms with adaptive attention for visual captioning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 5, pp. 1112–1131, 2019.
- [224] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: Spatial-temporal attention mechanism for video captioning," *IEEE transactions on multimedia*, vol. 22, no. 1, pp. 229–241, 2019.

- [225] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, “Visual transformers: Token-based image representation and processing for computer vision,” *arXiv preprint arXiv:2006.03677*, 2020.
- [226] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*, pp. 213–229, Springer, 2020.
- [227] S. d’Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, “Convit: Improving vision transformers with soft convolutional inductive biases,” *arXiv preprint arXiv:2103.10697*, 2021.
- [228] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, “Deepvit: Towards deeper vision transformer,” *arXiv preprint arXiv:2103.11886*, 2021.
- [229] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, “Training data-efficient image transformers & distillation through attention,” in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357, PMLR, 18–24 Jul 2021.
- [230] B. Gheflati and H. Rivaz, “Vision transformer for classification of breast ultrasound images,” *arXiv preprint arXiv:2110.14731*, 2021.
- [231] M. Ilse, J. Tomczak, and M. Welling, “Attention-based deep multiple instance learning,” in *International conference on machine learning*, pp. 2127–2136, PMLR, 2018.
- [232] Z. Han, B. Wei, Y. Hong, T. Li, J. Cong, X. Zhu, H. Wei, and W. Zhang, “Accurate screening of covid-19 using attention-based deep 3d multiple instance learning,” *IEEE transactions on medical imaging*, vol. 39, no. 8, pp. 2584–2594, 2020.
- [233] S. Perera, S. Adhikari, and A. Yilmaz, “Pocformer: A lightweight transformer architecture for detection of covid-19 using point of care ultrasound,” in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 195–199, IEEE, 2021.
- [234] J. Jiang and S. Lin, “Covid-19 detection in chest x-ray images using swin-transformer and transformer in transformer,” *arXiv preprint arXiv:2110.08427*, 2021.
- [235] C. Liu and Q. Yin, “Automatic Diagnosis of COVID-19 Using a tailored Transformer-Like Network,” in *Journal of Physics: Conference Series*, vol. 2010, p. 012175, IOP Publishing, 2021.
- [236] S. Park, G. Kim, J. Kim, B. Kim, and J. C. Ye, “Federated split task-agnostic vision transformer for COVID-19 CXR diagnosis,” in *Advances in Neural Information Processing Systems* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), 2021.
- [237] C.-C. Hsu, G.-L. Chen, and M.-H. Wu, “Visual transformer with statistical test for covid-19 classification,” *arXiv preprint arXiv:2107.05334*, 2021.
- [238] L. Zhang and Y. Wen, “A transformer-based framework for automatic covid19 diagnosis in chest cts,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 513–518, 2021.

- [239] 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*, pp. 234–241, Springer, 2015.
- [240] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. Prathosh, "xvitcos: Explainable vision transformer based covid-19 screening using radiography," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1–10, 2021.
- [241] X. Gao, Y. Qian, and A. Gao, "Covid-vit: Classification of covid-19 from ct chest images based on vision transformer models," *arXiv preprint arXiv:2107.01682*, 2021.
- [242] D. Shome, T. Kar, S. N. Mohanty, P. Tiwari, K. Muhammad, A. AlTameem, Y. Zhang, and A. K. J. Saudagar, "Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare," *International Journal of Environmental Research and Public Health*, vol. 18, no. 21, p. 11086, 2021.
- [243] A. A. E. Ambita, E. N. V. Boquio, and P. C. Naval, "Covit-gan: Vision transformer for covid-19 detection in ct scan images with self-attention gan for data augmentation," in *International Conference on Artificial Neural Networks*, pp. 587–598, Springer, 2021.
- [244] S. Park, G. Kim, Y. Oh, J. B. Seo, S. M. Lee, J. H. Kim, S. Moon, J.-K. Lim, and J. C. Ye, "Multi-task vision transformer using low-level chest x-ray feature corpus for covid-19 diagnosis and severity quantification," *Medical Image Analysis*, vol. 75, p. 102299, 2022.
- [245] M. Lu, Y. Pan, D. Nie, F. Liu, F. Shi, Y. Xia, and D. Shen, "Smile: Sparse-attention based multiple instance contrastive learning for glioma sub-type classification using pathological images," in *MICCAI Workshop on Computational Pathology*, pp. 159–169, PMLR, 2021.
- [246] H. Chen, C. Li, X. Li, G. Wang, W. Hu, Y. Li, W. Liu, C. Sun, Y. Yao, Y. Teng, *et al.*, "Gashis-transformer: A multi-scale visual transformer approach for gastric histopathology image classification," *arXiv preprint arXiv:2104.14528*, 2021.
- [247] Z. Jiang, Z. Dong, L. Wang, and W. Jiang, "Method for diagnosis of acute lymphoblastic leukemia based on vit-cnn ensemble model," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.
- [248] Y. Zheng, R. Gindra, M. Betke, J. Beane, and V. B. Kolachalama, "A deep learning based graph-transformer for whole slide image classification," *medRxiv*, 2021.
- [249] J. D. Bodapati, N. S. Shaik, and V. Naralasetti, "Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–15, 2021.
- [250] S. Yu, K. Ma, Q. Bi, C. Bian, M. Ning, N. He, Y. Li, H. Liu, and Y. Zheng, "Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 45–54, Springer, 2021.
- [251] R. Sun, Y. Li, T. Zhang, Z. Mao, F. Wu, and Y. Zhang, "Lesion-aware transformers for diabetic retinopathy grading," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10938–10947, 2021.

- [252] J. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, "Vision transformer-based recognition of diabetic retinopathy grade," *Medical Physics*, vol. 48, no. 12, pp. 7850–7863, 2021.
- [253] N. AlDahoul, H. Abdul Karim, M. Joshua Toledo Tan, M. A. Momo, and J. Ledesma Fermin, "Encoding retina image to words using ensemble of vision transformers for diabetic retinopathy grading," *F1000Research*, vol. 10, p. 948, 2021.
- [254] H. Yang, J. Chen, and M. Xu, "Fundus disease image classification based on improved transformer," in *2021 International Conference on Neuromorphic Computing (ICNC)*, pp. 207–214, IEEE, 2021.
- [255] M. Al-Shabi, K. Shak, and M. Tan, "ProCAN: Progressive growing channel attentive non-local network for lung nodule classification," *Pattern Recognition*, vol. 122, p. 108309, 2022.
- [256] J. Moranguinho, T. Pereira, B. Ramos, J. Morgado, J. L. Costa, and H. P. Oliveira, "Attention based deep multiple instance learning approach for lung cancer prediction using histopathological images," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2852–2855, IEEE, 2021.
- [257] L. He, J. C.-W. Chan, and Z. Wang, "Automatic depression recognition using cnn with attention mechanism from videos," *Neurocomputing*, vol. 422, pp. 165–175, 2021.
- [258] Y. Dai, Y. Gao, and F. Liu, "Transmed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, 2021.
- [259] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft Attention Improves Skin Cancer Classification Performance," in *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data* (M. Reyes, P. Henriques Abreu, J. Cardoso, M. Hajij, G. Zamzmi, P. Rahul, and L. Thakur, eds.), (Cham), pp. 13–23, Springer International Publishing, 2021.
- [260] Y. Barhoumi and G. Rasool, "Scopeformer: n-CNN-ViT Hybrid Model for Intracranial Hemorrhage Classification," in *Medical Imaging with Deep Learning*, 2021.
- [261] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.
- [262] S. Liang and Y. Gu, "Computer-aided diagnosis of alzheimer's disease through weak supervision deep learning framework with attention mechanism," *Sensors*, vol. 21, no. 1, p. 220, 2021.
- [263] 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*, pp. 770–778, 2016.
- [264] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 540–549, 2018.
- [265] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "Transbts: Multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 109–119, Springer, 2021.

- [266] Q. Jia and H. Shu, “Bitr-U-Net: a cnn-transformer combined network for mri brain tumor segmentation,” *arXiv preprint arXiv:2109.12271*, 2021.
- [267] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, “A volumetric transformer for accurate 3d tumor segmentation,” *arXiv preprint arXiv:2111.13300*, 2021.
- [268] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, “Swin U-Netr: Swin transformers for semantic segmentation of brain tumors in mri images,” *arXiv preprint arXiv:2201.01266*, 2022.
- [269] Y. Li, W. Cai, Y. Gao, and X. Hu, “More than encoder: Introducing transformer decoder to upsample,” *arXiv preprint arXiv:2106.10637*, 2021.
- [270] B. Chen, Y. Liu, Z. Zhang, G. Lu, and D. Zhang, “TransAttU-Net: Multi-level Attention-guided U-Net with Transformer for Medical Image Segmentation,” *arXiv preprint arXiv:2107.05274*, 2021.
- [271] G. Xu, X. Wu, X. Zhang, and X. He, “LeViT-U-Net: Make Faster Encoders with Transformer for Medical Image Segmentation,” *arXiv preprint arXiv:2107.08623*, 2021.
- [272] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, “Fat-net: Feature adaptive transformers for automated skin lesion segmentation,” *Medical Image Analysis*, vol. 76, p. 102327, 2022.
- [273] A. Vakanski, M. Xian, and P. E. Freer, “Attention-enriched deep learning model for breast tumor segmentation in ultrasound images,” *Ultrasound in Medicine & Biology*, vol. 46, no. 10, pp. 2819–2833, 2020.
- [274] X. Zhu, H. Hu, H. Wang, J. Yao, W. Li, D. Ou, and D. Xu, “Region aware transformer for automatic breast ultrasound tumor segmentation,” in *Medical Imaging with Deep Learning*, 2022.
- [275] Y. Liu, Y. Yang, W. Jiang, T. Wang, and B. Lei, “3d deep attentive U-Net with transformer for breast tumor segmentation from automated breast volume scanner,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 4011–4014, IEEE, 2021.
- [276] T. Prangemeier, C. Reich, and H. Koepl, “Attention-based transformers for instance segmentation of cells in microstructures,” in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 700–707, IEEE, 2020.
- [277] Y. Zhang, R. Higashita, H. Fu, Y. Xu, Y. Zhang, H. Liu, J. Zhang, and J. Liu, “A multi-branch hybrid transformer network for corneal endothelial cell segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 99–108, Springer, 2021.
- [278] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, and yongbing Zhang, “TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification,” in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 2136–2147, Curran Associates, Inc., 2021.

- [279] C. Nguyen, Z. Asad, R. Deng, and Y. Huo, “Evaluating transformer-based semantic segmentation networks for pathological image segmentation,” in *Medical Imaging 2022: Image Processing* (O. Colliot and I. Išgum, eds.), vol. 12032, pp. 942–947, International Society for Optics and Photonics, SPIE, 2022.
- [280] B. Yun, Y. Wang, J. Chen, H. Wang, W. Shen, and Q. Li, “Spectr: Spectral transformer for hyperspectral pathology image segmentation,” *arXiv preprint arXiv:2103.03604*, 2021.
- [281] H. Li, J. Zhang, and B. Menze, “Generalisable cardiac structure segmentation via attentional and stacked image adaptation,” in *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 297–304, Springer, 2020.
- [282] F. Kong and S. C. Shadden, “A generalizable deep-learning approach for cardiac magnetic resonance image segmentation using image augmentation and attention U-Net,” in *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 287–296, Springer, 2020.
- [283] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2223–2232, 2017.
- [284] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, *et al.*, “Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3543–3554, 2021.
- [285] K. Deng, Y. Meng, D. Gao, J. Bridge, Y. Shen, G. Lip, Y. Zhao, and Y. Zheng, “Transbridge: A lightweight transformer for left ventricle segmentation in echocardiography,” in *International Workshop on Advances in Simplifying Medical Ultrasound*, pp. 63–72, Springer, 2021.
- [286] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “TransU-Net: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.
- [287] X. Huang, Z. Deng, D. Li, and X. Yuan, “Missformer: An effective medical image segmentation transformer,” *arXiv preprint arXiv:2109.07162*, 2021.
- [288] C. Yao, J. Tang, M. Hu, Y. Wu, W. Guo, Q. Li, and X.-P. Zhang, “Claw U-Net: A U-Net variant network with deep feature concatenation for scleral blood vessel segmentation,” in *Artificial Intelligence: First CAAI International Conference, CICA 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part II*, (Berlin, Heidelberg), p. 67–78, Springer-Verlag, 2021.
- [289] Y. Chang, H. Menghan, Z. Guangtao, and Z. Xiao-Ping, “Transclaw U-Net: Claw U-Net with transformers for medical image segmentation,” *arXiv preprint arXiv:2107.05188*, 2021.
- [290] Y. Xie, J. Zhang, C. Shen, and Y. Xia, “Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation,” in *International conference on medical image computing and computer-assisted intervention*, pp. 171–180, Springer, 2021.
- [291] Z. Shen, H. Yang, Z. Zhang, and S. Zheng, “Automated Kidney Tumor Segmentation with Convolution and Transformer Network,” in *Kidney and Kidney Tumor Segmentation*

- (N. Heller, F. Isensee, D. Trofimova, R. Tejpaul, N. Papanikolopoulos, and C. Weight, eds.), (Cham), pp. 1–12, Springer International Publishing, 2022.
- [292] Y. Zhang, H. Liu, and Q. Hu, “Transfuse: Fusing transformers and cnns for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 14–24, Springer, 2021.
- [293] Y. Tang, D. Yang, W. Li, H. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, “Self-supervised pre-training of swin transformers for 3d medical image analysis,” *arXiv preprint arXiv:2111.14791*, 2021.
- [294] D. Karimi, S. D. Vasylechko, and A. Gholipour, “Convolution-free medical image segmentation using transformers,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 78–88, Springer, 2021.
- [295] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, “Swin-U-Net: U-Net-like pure transformer for medical image segmentation,” *arXiv preprint arXiv:2105.05537*, 2021.
- [296] D. V. Sang, T. Q. Chung, P. N. Lan, D. V. Hang, D. Van Long, and N. T. Thuy, “AG-CUResNeSt: A Novel Method for Colon Polyp Segmentation,” *arXiv preprint arXiv:2105.00402*, 2021.
- [297] Z. Tang, X. Peng, S. Geng, Y. Zhu, and D. N. Metaxas, “CU-Net: Coupled U-Nets,” in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, p. 305, BMVA Press, 2018.
- [298] Z. Zhang, B. Sun, and W. Zhang, “Pyramid medical transformer for medical image segmentation,” *arXiv preprint arXiv:2104.14702*, 2021.
- [299] A. Lin, B. Chen, J. Xu, Z. Zhang, and G. Lu, “Ds-transU-Net: Dual swin transformer U-Net for medical image segmentation,” *arXiv preprint arXiv:2106.06716*, 2021.
- [300] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, “Medical Image Segmentation using Squeeze-and-Expansion Transformers,” in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (Z.-H. Zhou, ed.), pp. 807–815, International Joint Conferences on Artificial Intelligence Organization, 2021.
- [301] R. Ranjbarzadeh, A. B. Kasgari, S. J. Ghouschi, S. Anari, M. Naseri, and M. Bendeche, “Brain tumor segmentation based on deep learning and an attention mechanism using mri multi-modalities brain images,” *Scientific Reports*, vol. 11, no. 1, pp. 1–17, 2021.
- [302] A. Ribeiro, “Study of attention mechanisms and ensemble methods for medical image semantic segmentation,” Master’s thesis, Universidade do Minho (Portugal), 2019.
- [303] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, “U-Net transformer: Self and cross attention for medical image segmentation,” in *International Workshop on Machine Learning in Medical Imaging*, pp. 267–276, Springer, 2021.
- [304] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, and J. Qin, “Boundary-aware transformers for skin lesion segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 206–216, Springer, 2021.

- [305] Y. Li, S. Wang, J. Wang, G. Zeng, W. Liu, Q. Zhang, Q. Jin, and Y. Wang, "Gt U-Net: A U-Net like group transformer network for tooth root segmentation," in *International Workshop on Machine Learning in Medical Imaging*, pp. 386–395, Springer, 2021.
- [306] Y. Li, G. Zeng, Y. Zhang, J. Wang, Q. Jin, L. Sun, Q. Zhang, Q. Lian, G. Qian, N. Xia, *et al.*, "Agmb-transformer: Anatomy-guided multi-branch transformer network for automated evaluation of root canal therapy," *IEEE Journal of Biomedical and Health Informatics*, 2021.
- [307] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, "Sa-U-Net: Spatial attention U-Net for retinal vessel segmentation," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 1236–1242, IEEE, 2021.
- [308] J. Wang, Y. Zhao, L. Qian, X. Yu, and Y. Gao, "Ear-net: Error attention refining network for retinal vessel segmentation," in *2021 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–7, IEEE, 2021.
- [309] C. Guo, M. Szemenyei, Y. Yi, W. Zhou, and H. Bian, "Residual spatial attention network for retinal vessel segmentation," in *International Conference on Neural Information Processing*, pp. 509–519, Springer, 2020.
- [310] X.-F. Du, J.-S. Wang, and W.-z. Sun, "U-Net retinal blood vessel segmentation algorithm based on improved pyramid pooling method and attention mechanism," *Physics in Medicine & Biology*, vol. 66, no. 17, p. 175013, 2021.
- [311] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.
- [312] J. Gao, Q. Huang, Z. Gao, and S. Chen, "Image segmentation of retinal blood vessels based on dual-attention multiscale feature fusion," *Computational and Mathematical Methods in Medicine*, vol. 2022, 2022.
- [313] A. Amer, T. Lambrou, and X. Ye, "Mda-U-Net: A multi-scale dilated attention U-Net for medical image segmentation," *Applied Sciences*, vol. 12, no. 7, p. 3676, 2022.
- [314] P. Zhao, J. Zhang, W. Fang, and S. Deng, "SCAU-Net: spatial-channel attention U-Net for gland segmentation," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 670, 2020.
- [315] Q. Jin, Z. Meng, C. Sun, H. Cui, and R. Su, "Ra-U-Net: A hybrid deep attention-aware network to extract liver and tumor in ct scans," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 1471, 2020.
- [316] I. Sobirov, O. Nazarov, H. Alasmawi, and M. Yaqub, "Automatic segmentation of head and neck tumor: How powerful transformers are?," in *Medical Imaging with Deep Learning*, 2022.
- [317] X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie, "After-U-Net: Axial fusion transformer U-Net for medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3971–3981, 2022.
- [318] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 2577–2586, Association for Computational Linguistics, July 2018.

- [319] Y. Xiong, B. Du, and P. Yan, “Reinforced transformer for medical image captioning,” in *International Workshop on Machine Learning in Medical Imaging*, pp. 673–680, Springer, 2019.
- [320] J. Lovelace and B. Mortazavi, “Learning to generate clinically coherent chest x-ray reports,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 1235–1243, 2020.
- [321] P. Srinivasan, D. Thapar, A. Bhavsar, and A. Nigam, “Hierarchical x-ray report generation via pathology tags and multi head attention,” in *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [322] Y. Miura, Y. Zhang, E. Tsai, C. Langlotz, and D. Jurafsky, “Improving factual completeness and consistency of image-to-text radiology report generation,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Online), pp. 5288–5304, Association for Computational Linguistics, June 2021.
- [323] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, “Generating radiology reports via memory-driven transformer,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 1439–1449, Association for Computational Linguistics, Nov. 2020.
- [324] Liu, Fenglin and Yin, Changchang and Wu, Xian and Ge, Shen and Zhang, Ping and Sun, Xu, “Contrastive Attention for Automatic Chest X-ray Report Generation,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, (Online), pp. 269–280, Association for Computational Linguistics, August 2021.
- [325] I. Najdenkoska, X. Zhen, M. Worring, and L. Shao, “Variational Topic Inference for Chest X-Ray Report Generation,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, eds.), (Cham), pp. 625–635, Springer International Publishing, 2021.
- [326] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, “Automated radiology report generation using conditioned transformers,” *Informatics in Medicine Unlocked*, vol. 24, p. 100557, 2021.
- [327] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, *et al.*, “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning,” *arXiv preprint arXiv:1711.05225*, 2017.
- [328] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [329] A. B. Amjoud and M. Amrouch, “Automatic generation of chest x-ray reports using a transformer-based deep learning model,” in *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)*, pp. 1–5, IEEE, 2021.
- [330] S. Tipirneni and C. K. Reddy, “Self-supervised Transformer for Multivariate Clinical Time-Series with Missing Values,” *arXiv preprint arXiv:2107.14293*, 2021.
- [331] Y. Wang, Z. Lin, J. Tian, Z. Shi, Y. Zhang, J. Fan, and Z. He, “Confidence-guided radiology report generation,” *arXiv preprint arXiv:2106.10887*, 2021.

- [332] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, “Exploring and distilling posterior and prior knowledge for radiology report generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13753–13762, 2021.
- [333] F. Liu, C. You, X. Wu, S. Ge, X. Sun, *et al.*, “Auto-encoding knowledge graph for unsupervised medical report generation,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [334] H. Nguyen, D. Nie, T. Badamdorj, Y. Liu, Y. Zhu, J. Truong, and L. Cheng, “Automated generation of accurate & fluent medical X-ray reports,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (Online and Punta Cana, Dominican Republic), pp. 3552–3569, Association for Computational Linguistics, Nov. 2021.
- [335] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, “Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 72–82, Springer, 2021.
- [336] B. Hou, G. Kaissis, R. M. Summers, and B. Kainz, “Ratchet: Medical transformer for chest x-ray diagnosis and reporting,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 293–303, Springer, 2021.
- [337] H.-Y. Zhou, X. Chen, Y. Zhang, R. Luo, L. Wang, and Y. Yu, “Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports,” *Nature Machine Intelligence*, pp. 1–9, 2022.
- [338] J. Zhang, Y. Nie, J. Chang, and J. J. Zhang, “Surgical instruction generation with transformers,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 290–299, Springer, 2021.
- [339] X. Wang, Z. Xu, L. Tam, D. Yang, and D. Xu, “Self-supervised image-text pre-training with mixed data in chest x-rays,” *arXiv preprint arXiv:2103.16022*, 2021.
- [340] Z. Shen, R. Fu, C. Lin, and S. Zheng, “Cotr: Convolution in transformer network for end to end polyp detection,” in *2021 7th International Conference on Computer and Communications (ICCC)*, pp. 1757–1761, IEEE, 2021.
- [341] S. Liu, H. Zhou, X. Shi, and J. Pan, “Transformer for polyp detection,” *arXiv preprint arXiv:2111.07918*, 2021.
- [342] T. S. Mathai, S. Lee, D. C. Elton, T. C. Shen, Y. Peng, Z. Lu, and R. M. Summers, “Lymph node detection in T2 MRI with transformers,” in *Medical Imaging 2022: Computer-Aided Diagnosis* (K. Drukker, K. M. Iftekharuddin, H. Lu, M. A. Mazurowski, C. Muramatsu, and R. K. Samala, eds.), vol. 12033, pp. 855–859, International Society for Optics and Photonics, SPIE, 2022.
- [343] K. Gong, C. Catana, J. Qi, and Q. Li, “Direct Reconstruction of Linear Parametric Images From Dynamic PET Using Nonlocal Deep Image Prior,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 3, pp. 680–689, 2022.
- [344] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-Net: learning dense volumetric segmentation from sparse annotation,” in *International Conference on*

- Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, Springer, 2016.
- [345] P. Grundmann, S. Arnold, and A. Löser, “Self-supervised Answer Retrieval on Clinical Notes,” *arXiv preprint arXiv:2108.00775*, 2021.
 - [346] B.-H. Kim and V. Ganapathi, “Read, Attend, and Code: Pushing the Limits of Medical Codes Prediction from Clinical Notes by Machines,” in *Proceedings of the 6th Machine Learning for Healthcare Conference* (K. Jung, S. Yeung, M. Sendak, M. Sjoding, and R. Ranganath, eds.), vol. 149 of *Proceedings of Machine Learning Research*, pp. 196–208, PMLR, 8 2021.
 - [347] Y. Li, G. Zeng, Y. Zhang, J. Wang, Q. Zhang, Q. Jin, L. Sun, Q. Lian, N. Xia, R. Peng, *et al.*, “Anatomy-Guided Parallel Bottleneck Transformer Network for Automated Evaluation of Root Canal Therapy,” *arXiv preprint arXiv:2105.00381*, 2021.
 - [348] X. Ma, G. Luo, W. Wang, and K. Wang, “Transformer Network for Significant Stenosis Detection in CCTA of Coronary Arteries,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, eds.), (Cham), pp. 516–525, Springer International Publishing, 2021.
 - [349] M. Nauta, D. Bucur, and C. Seifert, “Causal discovery with attention-based convolutional neural networks,” *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 312–340, 2019.
 - [350] R. Shrestha, E. Fajardo, N. Gil, K. Fidelis, A. Kryshchuk, B. Monastyrskyy, and A. Fiser, “Assessing the accuracy of contact predictions in casp13,” *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1058–1068, 2019.
 - [351] W. Zheng, Y. Li, C. Zhang, R. Pearce, S. Mortuza, and Y. Zhang, “Deep-learning contact-map guided protein structure prediction in casp13,” *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1149–1164, 2019.
 - [352] J. Hou, T. Wu, R. Cao, and J. Cheng, “Protein tertiary structure modeling driven by deep learning and contact distance prediction in casp13,” *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1165–1178, 2019.
 - [353] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, *et al.*, “Improved protein structure prediction using potentials from deep learning,” *Nature*, vol. 577, no. 7792, pp. 706–710, 2020.
 - [354] M. Filipavicius, M. Manica, J. Cadow, and M. R. Martinez, “Pre-training protein language models with label-agnostic binding pairs enhances performance in downstream tasks,” in *Machine Learning for Structural Biology Workshop, NeurIPS*, 2020.
 - [355] C. Chen, T. Wu, Z. Guo, and J. Cheng, “Combination of deep neural network with attention mechanism enhances the explainability of protein contact prediction,” *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 6, pp. 697–707, 2021.
 - [356] J. Wang, X. Liu, S. Shen, L. Deng, and H. Liu, “DeepDDS: deep graph neural network with attention mechanism to predict synergistic drug combinations,” *Briefings in Bioinformatics*, vol. 23, 1 2022.

- [357] A. Khan and B. Lee, “Gene transformer: Transformers for the gene expression-based classification of lung cancer subtypes,” *arXiv preprint arXiv:2108.11833*, 2021.
- [358] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems* 32 (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.
- [359] 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*, pp. 4700–4708, 2017.
- [360] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, “Captum: A unified and generic model interpretability library for pytorch,” 2020.
- [361] H. Chefer, S. Gur, and L. Wolf, “Transformer interpretability beyond attention visualization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021.
- [362] 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 (CVPR)*, pp. 248–255, IEEE, 2009.
- [363] J. D. Hunter, “Matplotlib: A 2d graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
- [364] J.-B. Cordonnier, A. Loukas, and M. Jaggi, “On the relationship between self-attention and convolutional layers,” in *International Conference on Learning Representations*, 2020.
- [365] Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery,” *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
- [366] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity checks for saliency maps,” *Advances in neural information processing systems*, vol. 31, 2018.
- [367] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, “The (un) reliability of saliency methods,” in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280, Springer, 2019.
- [368] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, “A benchmark for interpretability methods in deep neural networks,” *Advances in neural information processing systems*, vol. 32, 2019.
- [369] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, “Interpreting interpretability: understanding data scientists’ use of interpretability tools for machine learning,” in *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–14, 2020.

- [370] 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*, vol. 22, pp. 1191–1198, Sep 2015.
- [371] 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*, vol. 201, pp. 611–617, Sep 2013.
- [372] J. Vosschenrich, P. Brantner, J. Cyriac, D. T. Boll, E. M. Merkle, and T. Heye, "Quantifying radiology resident fatigue: Analysis of preliminary reports," *Radiology*, vol. 298, no. 3, pp. 632–639, 2021.
- [373] 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*, pp. 305–314, Springer, 2020.
- [374] Radiopaedia, "Radiopaedia.org." <https://radiopaedia.org/>, 2025. [Online. Accessed 22-January-2025].
- [375] Z. Li, X. Zhang, H. Müller, and S. Zhang, "Large-scale retrieval for medical image analytics: A comprehensive review," *Medical Image Analysis*, vol. 43, pp. 66–84, 2018.
- [376] P. Das and A. Neelima, "An overview of approaches for content-based medical image retrieval," *International journal of multimedia information retrieval*, vol. 6, no. 4, pp. 271–280, 2017.
- [377] 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*, vol. 263, pp. 60–86, 2014.
- [378] 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*, vol. 1, pp. 144–148, IEEE, 2014.
- [379] H. R. Tizhoosh, "Barcode annotations for medical image retrieval: A preliminary investigation," in *2015 IEEE international conference on image processing (ICIP)*, pp. 818–822, IEEE, 2015.
- [380] M. Srinivas, R. R. Naidu, C. S. Sastry, and C. K. Mohan, "Content based medical image retrieval using dictionary learning," *Neurocomputing*, vol. 168, pp. 880–895, 2015.
- [381] 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*, pp. 457–465, 2015.
- [382] 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*, vol. 28, no. 1, pp. 110–124, 2016.
- [383] 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*, vol. 66, pp. 148–158, 2017.

- [384] J. Nowaková, M. Prílepok, and V. Snášel, “Medical image retrieval using vector quantization and fuzzy s-tree,” *Journal of medical systems*, vol. 41, no. 2, pp. 1–16, 2017.
- [385] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid, “Medical image retrieval using deep convolutional neural network,” *Neurocomputing*, vol. 266, pp. 8–20, 2017.
- [386] 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*, vol. 10954, p. 1095410, International Society for Optics and Photonics, 2019.
- [387] 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*, vol. 8, no. 4, p. 462, 2019.
- [388] Y. Cai, Y. Li, C. Qiu, J. Ma, and X. Gao, “Medical image retrieval based on convolutional neural network and supervised hashing,” *IEEE Access*, vol. 7, pp. 51877–51885, 2019.
- [389] A. E. Minarno, K. M. Ghufro, 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)*, pp. 29–34, IEEE, 2021.
- [390] 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)*, pp. 1–4, IEEE, 2021.
- [391] 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*, vol. 13, no. 4, pp. 600–612, 2004.
- [392] 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*, vol. 35, no. 5, pp. 1285–1298, 2016.
- [393] 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*, pp. 589–596, Springer, 2015.
- [394] 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*, pp. 118–130, Springer, 2020.
- [395] 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.
- [396] S. S. Mishra, B. Mandal, and N. B. Puan, “Multi-level dual-attention based CNN for macular optical coherence tomography classification,” *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1793–1797, 2019.
- [397] C. Chen and A. Ross, “An explainable attention-guided iris presentation attack detector,” in *WACV (Workshops)*, pp. 97–106, 2021.

- [398] M. D. Zeiler, “Adadelta: an adaptive learning rate method,” *arXiv preprint arXiv:1212.5701*, 2012.
- [399] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, “Cost-sensitive boosting for classification of imbalanced data,” *Pattern recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.
- [400] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.
- [401] F. Chollet *et al.*, “Keras.” <https://keras.io>, 2015.
- [402] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, “investigate neural networks!,” *J. Mach. Learn. Res.*, vol. 20, no. 93, pp. 1–8, 2019.
- [403] K. Fernandes and J. S. Cardoso, “Hypothesis transfer learning based on structural model similarity,” *Neural Computing and Applications*, vol. 31, no. 8, pp. 3417–3430, 2019.
- [404] D. C. Castro, I. Walker, and B. Glocker, “Causality matters in medical imaging,” *Nature Communications*, vol. 11, no. 1, pp. 1–10, 2020.
- [405] H. Montenegro, W. Silva, and J. S. Cardoso, “Privacy-preserving generative adversarial network for case-based explainability in medical image analysis,” *IEEE Access*, vol. 9, pp. 148037–148047, 2021.
- [406] A. Ziller, D. Usynin, R. Braren, M. Makowski, D. Rueckert, and G. Kaissis, “Medical imaging deep learning with differential privacy,” *Scientific Reports*, vol. 11, no. 1, pp. 1–8, 2021.
- [407] 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*, vol. 2, no. 11, pp. 665–673, 2020.
- [408] D. Mahapatra, A. Poellinger, and M. Reyes, “Interpretability-guided inductive bias for deep learning based medical image,” *Medical image analysis*, vol. 81, p. 102551, 2022.
- [409] D. Mahapatra, A. Poellinger, and M. Reyes, “Graph node based interpretability guided sample selection for active learning,” *IEEE Transactions on Medical Imaging*, 2022.
- [410] H. Montenegro, W. Silva, and J. S. Cardoso, “Towards privacy-preserving explanations in medical image analysis,” *1st Workshop on Interpretable Machine Learning in Healthcare at ICML2021*, 2021.
- [411] V. T. Lin and E. S. Yang, “The Pros and Cons of Incorporating Transcriptomics in the Age of Precision Oncology,” *JNCI: Journal of the National Cancer Institute*, vol. 111, no. 10, pp. 1016–1022, 2019.
- [412] K. A. Khan and R. S. Kerbel, “Improving immunotherapy outcomes with anti-angiogenic treatments and vice versa,” *Nature Reviews Clinical Oncology*, vol. 15, no. 5, pp. 310–324, 2018.

- [413] S. K. Longo, M. G. Guo, A. L. Ji, and P. A. Khavari, “Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics,” *Nature Reviews Genetics*, vol. 22, no. 10, pp. 627–644, 2021.
- [414] D. Montezuma, S. P. Oliveira, P. C. Neto, D. Oliveira, A. Monteiro, J. S. Cardoso, and I. Macedo-Pinto, “Annotating for Artificial Intelligence Applications in Digital Pathology: A Practical Guide for Pathologists and Researchers,” *Modern Pathology*, vol. 36, no. 4, p. 100086, 2023.
- [415] S. P. Oliveira, J. Ribeiro Pinto, T. Gonçalves, R. Canas-Marques, M.-J. Cardoso, H. P. Oliveira, and J. S. Cardoso, “Weakly-Supervised Classification of HER2 Expression in Breast Cancer Haematoxylin and Eosin Stained Slides,” *Applied Sciences*, vol. 10, no. 14, 2020.
- [416] P. C. Neto, S. P. Oliveira, D. Montezuma, J. Fraga, A. Monteiro, L. Ribeiro, S. Gonçalves, I. M. Pinto, and J. S. Cardoso, “iMIL4PATH: A Semi-Supervised Interpretable Approach for Colorectal Whole-Slide Images,” *Cancers*, vol. 14, no. 10, 2022.
- [417] B. Wu and G. Moeckel, “Application of digital pathology and machine learning in the liver, kidney and lung diseases,” *Journal of Pathology Informatics*, vol. 14, p. 100184, 2023.
- [418] C. L. Srinidhi, O. Ciga, and A. L. Martel, “Deep eural network models for computational histopathology: A survey,” *Medical Image Analysis*, vol. 67, p. 101813, 2021.
- [419] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, “Data-efficient and weakly supervised computational pathology on whole-slide images,” *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.
- [420] J. Lipkova, R. J. Chen, B. Chen, M. Y. Lu, M. Barbieri, D. Shao, A. J. Vaidya, C. Chen, L. Zhuang, D. F. Williamson, M. Shaban, T. Y. Chen, and F. Mahmood, “Artificial intelligence for multimodal data integration in oncology,” *Cancer Cell*, vol. 40, no. 10, pp. 1095–1110, 2022.
- [421] R. J. Chen and R. G. Krishnan, “Self-Supervised Vision Transformers Learn Visual Concepts in Histopathology,” *Learning Meaningful Representations of Life, NeurIPS 2021*, 2021.
- [422] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood, “Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16144–16155, June 2022.
- [423] M. Y. Lu, B. Chen, A. Zhang, D. F. K. Williamson, R. J. Chen, T. Ding, L. P. Le, Y.-S. Chuang, and F. Mahmood, “Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19764–19775, June 2023.
- [424] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, “A visual–language foundation model for pathology image analysis using medical Twitter,” *Nature Medicine*, pp. 1–10, 2023.
- [425] A. Janowczyk, R. Zuo, H. Gilmore, M. Feldman, and A. Madabhushi, “HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides,” *JCO Clinical Cancer Informatics*, pp. 1–7, 2019. PMID: 30990737.

- [426] Y. Chen, J. Zee, A. Smith, C. Jayapandian, J. Hodgins, D. Howell, M. Palmer, D. Thomas, C. Cassol, A. B. Farris III, K. Perkinson, A. Madabhushi, L. Barisoni, and A. Janowczyk, "Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies," *The Journal of Pathology*, vol. 253, no. 3, pp. 268–278, 2021.
- [427] cBioPortal for Cancer Genomics, "Normalized expression data for tcga-brca cohort." <http://www.cBioPortal.org>, 2023. Accessed: 2023-03-31.
- [428] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545–15550, 2005.
- [429] F. Coscia, E. Lengyel, J. Duraiswamy, B. Ashcroft, M. Bassani-Sternberg, M. Wierer, A. Johnson, K. Wroblewski, A. Montag, S. D. Yamada, *et al.*, "Multi-level proteomics identifies CT45 as a chemosensitivity mediator and immunotherapy target in ovarian cancer," *Cell*, vol. 175, no. 1, pp. 159–170, 2018.
- [430] M. Yi, D. V. Nissley, F. McCormick, and R. M. Stephens, "ssGSEA score-based Ras dependency indexes derived from gene expression data reveal potential Ras addiction mechanisms with possible clinical implications," *Scientific Reports*, vol. 10, no. 1, p. 10258, 2020.
- [431] S. Kornblith, J. Shlens, and Q. V. Le, "Do Better ImageNet Models Transfer Better?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [432] H.-Y. Zhou, S. Yu, C. Bian, Y. Hu, K. Ma, and Y. Zheng, "Comparing to Learn: Surpassing ImageNet Pretraining on Radiographs by Comparing Image Representations," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racocanu, and L. Joskowicz, eds.), (Cham), pp. 398–407, Springer International Publishing, 2020.
- [433] M. A. Morid, A. Borjali, and G. Del Fiore, "A scoping review of transfer learning research on medical image analysis using ImageNet," *Computers in Biology and Medicine*, vol. 128, p. 104115, 2021.
- [434] X. Li, M. Cen, J. Xu, H. Zhang, and X. S. Xu, "Improving feature extraction from histopathological images through a fine-tuning ImageNet model," *Journal of Pathology Informatics*, vol. 13, p. 100115, 2022.
- [435] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, 2019.
- [436] K. Lawonn, N. N. Smit, K. Bühler, and B. Preim, "A survey on multimodal medical data visualization," in *Computer Graphics Forum*, vol. 37, pp. 413–438, Wiley Online Library, 2018.
- [437] A. P. James and B. V. Dasarthy, "Medical image fusion: A survey of the state of the art," *Information Fusion*, vol. 19, pp. 4–19, 2014.
- [438] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1863–1883, 2018.

- [439] L. Rokach, “Ensemble-based classifiers,” *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010.
- [440] J. Gao, P. Li, Z. Chen, and J. Zhang, “A survey on deep learning for multimodal data fusion,” *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.
- [441] A. E. U. Cerna, M. Pattichis, D. P. VanMaanen, L. Jing, A. A. Patel, J. V. Stough, C. M. Haggerty, and B. K. Fornwalt, “Interpretable neural networks for predicting mortality risk using multi-modal electronic health records,” *arXiv preprint arXiv:1901.08125*, 2019.
- [442] W. Hu, X. Meng, Y. Bai, A. Zhang, G. Qu, B. Cai, G. Zhang, T. W. Wilson, J. M. Stephen, V. D. Calhoun, *et al.*, “Interpretable multimodal fusion networks reveal mechanisms of brain cognition,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1474–1483, 2021.
- [443] M. Blendowski and M. P. Heinrich, “Learning interpretable multi-modal features for alignment with supervised iterative descent,” in *International Conference on Medical Imaging with Deep Learning*, pp. 73–83, PMLR, 2019.
- [444] A. Taleb, C. Lippert, T. Klein, and M. Nabi, “Multimodal self-supervised learning for medical image analysis,” in *International Conference on Information Processing in Medical Imaging*, pp. 661–673, Springer, 2021.
- [445] D. Lahat, T. Adali, and C. Jutten, “Multimodal data fusion: an overview of methods, challenges, and prospects,” *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.
- [446] F. E. White, “Data fusion lexicon,” tech. rep., Joint Directors of Labs Washington DC, 1991.
- [447] D. L. Hall and J. Llinas, “An introduction to multisensor data fusion,” *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.
- [448] R. C. Luo and M. G. Kay, “Multisensor integration and fusion: issues and approaches,” in *Sensor Fusion*, vol. 931, pp. 42–49, International Society for Optics and Photonics, 1988.
- [449] R. C. Luo, C.-C. Yih, and K. L. Su, “Multisensor fusion and integration: approaches, applications, and future research directions,” *IEEE Sensors journal*, vol. 2, no. 2, pp. 107–119, 2002.
- [450] A. N. Steinberg, C. L. Bowman, and F. E. White, “Revisions to the JDL data fusion model,” in *Sensor fusion: Architectures, algorithms, and applications III*, vol. 3719, pp. 430–441, International Society for Optics and Photonics, 1999.
- [451] S. B. Ayed, H. Trichili, and A. M. Alimi, “Data fusion architectures: A survey and comparison,” in *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*, pp. 277–282, IEEE, 2015.
- [452] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.
- [453] H. Noh, P. H. Seo, and B. Han, “Image question answering using convolutional neural network with dynamic parameter prediction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 30–38, 2016.

- [454] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.
- [455] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, “Composing text and image for image retrieval-an empirical odyssey,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6439–6448, 2019.
- [456] M. U. Anwaar, E. Labintcev, and M. Kleinsteuber, “Compositional Learning of Image-Text Query for Image Retrieval,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1140–1149, 2021.
- [457] S. Nawaz, A. Calefati, M. Caraffini, N. Landro, and I. Gallo, “Are these birds similar: Learning branched networks for fine-grained representations,” in *2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pp. 1–5, IEEE, 2019.
- [458] I. Gallo, G. Ria, N. Landro, and R. La Grassa, “Image and Text fusion for UPMC Food-101 using BERT and CNNs,” in *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pp. 1–6, IEEE, 2020.
- [459] J. Ferrando, J. L. Domínguez, J. Torres, R. García, D. García, D. Garrido, J. Cortada, and M. Valero, “Improving accuracy and speeding up document image classification through parallel systems,” in *International Conference on Computational Science*, pp. 387–400, Springer, 2020.
- [460] M. Lorenzi, I. J. Simpson, A. F. Mendelson, S. B. Vos, M. J. Cardoso, M. Modat, J. M. Schott, and S. Ourselin, “Multimodal image analysis in alzheimer’s disease via statistical modelling of non-local intensity correlations,” *Scientific reports*, vol. 6, no. 1, pp. 1–8, 2016.
- [461] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative, *et al.*, “Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis,” *NeuroImage*, vol. 101, pp. 569–582, 2014.
- [462] N. M. Correa, T. Eichele, T. Adalı, Y.-O. Li, and V. D. Calhoun, “Multi-set canonical correlation analysis for the fusion of concurrent single trial erp and functional mri,” *NeuroImage*, vol. 50, no. 4, pp. 1438–1445, 2010.
- [463] J. R. Kettenring, “Canonical analysis of several sets of variables,” *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.
- [464] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, “HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1116–1126, 2018.
- [465] T. Xu, H. Zhang, X. Huang, S. Zhang, and D. N. Metaxas, “Multimodal deep learning for cervical dysplasia diagnosis,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 115–123, Springer, 2016.
- [466] T. Chen, X. Ma, X. Ying, W. Wang, C. Yuan, W. Lu, D. Z. Chen, and J. Wu, “Multi-modal fusion learning for cervical dysplasia diagnosis,” in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1505–1509, IEEE, 2019.

- [467] G. Carneiro, J. Nascimento, and A. P. Bradley, “Unregistered multiview mammogram analysis with pre-trained deep learning models,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 652–660, Springer, 2015.
- [468] M. Liang, Z. Li, T. Chen, and J. Zeng, “Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 4, pp. 928–937, 2014.
- [469] N. Srivastava, R. Salakhutdinov, *et al.*, “Multimodal learning with deep Boltzmann machines,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2949–2980, 2014.
- [470] G. E. Hinton, “A practical guide to training restricted boltzmann machines,” in *Neural Networks: Tricks of the Trade*, pp. 599–619, Springer, 2012.
- [471] M. A. Carreira-Perpinan and G. Hinton, “On contrastive divergence learning,” in *International Workshop on Artificial Intelligence and Statistics*, pp. 33–40, PMLR, 2005.
- [472] X. Cheng, L. Zhang, and Y. Zheng, “Deep similarity learning for multimodal medical images,” *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 248–252, 2018.
- [473] Y. Cao, S. Steffey, J. He, D. Xiao, C. Tao, P. Chen, and H. Müller, “Medical image retrieval: a multimodal approach,” *Cancer informatics*, vol. 13, pp. CIN–S14053, 2014.
- [474] A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai, and I. S. Kohane, “Clinical concept embeddings learned from massive sources of multimodal medical data,” in *Pacific Symposium on Biocomputing 2020*, pp. 295–306, World Scientific, 2019.
- [475] F. Qin, V. Madan, U. Ratan, Z. Karnin, V. Kapoor, P. Bhatia, and T. Kass-Hout, “Improving Early Sepsis Prediction with Multi Modal Learning,” *arXiv preprint arXiv:2107.11094*, 2021.
- [476] M. Jin, M. T. Bahadori, A. Colak, P. Bhatia, B. Celikkaya, R. Bhakta, S. Senthivel, M. Khalilia, D. Navarro, B. Zhang, *et al.*, “Improving hospital mortality prediction with medical named entities and multimodal learning,” *arXiv preprint arXiv:1811.12276*, 2018.
- [477] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling, and I. Soerjomataram, “Current and future burden of breast cancer: Global statistics for 2020 and 2040,” *The Breast*, vol. 66, pp. 15–23, 2022.
- [478] 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*, vol. 49, pp. 123–130, 2020.
- [479] W. Silva, M. Carvalho, C. Mavioso, M. J. Cardoso, and J. S. Cardoso, “Deep Aesthetic Assessment and Retrieval of Breast Cancer Treatment Outcomes,” in *Pattern Recognition and Image Analysis* (A. J. Pinho, P. Georgieva, L. F. Teixeira, and J. A. Sánchez, eds.), (Cham), pp. 108–118, Springer International Publishing, 2022.
- [480] M. Zwakman, A. Tan, C. Boersma, J. Klinkenbijn, E. Noorda, T. De Jong, and A. Francken, “Long-term quality of life and aesthetic outcomes after breast conserving surgery in patients with breast cancer,” *European Journal of Surgical Oncology*, vol. 48, no. 8, pp. 1692–1698, 2022.

- [481] M. K. Kim, T. Kim, H.-G. Moon, U. S. Jin, K. Kim, J. Kim, J. Lee, E. Lee, T.-K. Yoo, D.-Y. Noh, *et al.*, “Effect of cosmetic outcome on quality of life after breast cancer surgery,” *European Journal of Surgical Oncology (EJSO)*, vol. 41, no. 3, pp. 426–432, 2015.
- [482] J. Huan, “Research on the application of artificial intelligence in image and text database retrieval,” *Frontiers in Computing and Intelligent Systems*, vol. 2, 2022.
- [483] W. M. Naqvi, H. Sundus, G. Mishra, R. Muthukrishnan, and P. K. Kandakurti, “Ai in medical education curriculum: The future of healthcare learning,” *European Journal of Therapeutics*, 2024.
- [484] W. Chen, Y. Liu, W. Wang, E. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, “Deep image retrieval: A survey,” *arXiv preprint arXiv:2101.11282*, vol. 1, no. 3, p. 6, 2021.
- [485] X. Li, J. Yang, and J. Ma, “Recent developments of content-based image retrieval (cbir),” *Neurocomputing*, vol. 452, pp. 675–689, 2021.
- [486] S. Agrawal, A. Chowdhary, S. Agarwala, V. Mayya, and S. Kamath S, “Content-based medical image retrieval system for lung diseases using deep cnns,” *International Journal of Information Technology*, vol. 14, no. 7, pp. 3619–3627, 2022.
- [487] F. Z. Boroujeni, D. H. Afshari, and F. Mahmoodi, “Covid-19 image retrieval using siamese deep neural network and hashing technique,” in *2023 14th International Conference on Information and Knowledge Technology (IKT)*, pp. 186–193, IEEE, 2023.
- [488] N. Sampathila, Pavithra, and R. J. Martis, “Computational approach for content-based image retrieval of k-similar images from brain mr image database,” *Expert Systems*, vol. 39, no. 7, p. e12652, 2022.
- [489] L. Si, R. Jin, S. C. Hoi, and M. R. Lyu, “Collaborative image retrieval via regularized metric learning,” *Multimedia Systems*, vol. 12, pp. 34–44, 2006.
- [490] W. F. Lo, N. Mital, H. Wu, and D. Gündüz, “Collaborative semantic communication for edge inference,” *IEEE Wireless Communications Letters*, 2023.
- [491] S. Shirkhani, H. Mokayed, R. Saini, and H. Y. Chai, “Study of ai-driven fashion recommender systems,” *SN Computer Science*, vol. 4, 2023.
- [492] H. A. Al-Jubouri, “Content-based image retrieval: Survey,” 2019.
- [493] S. R. Dubey, “A Decade Survey of Content Based Image Retrieval Using Deep Learning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, 2022.
- [494] K. Yan, Y. Wang, D. Liang, T. Huang, and Y. Tian, “CNN vs. SIFT for Image Retrieval: Alternative or Complementary?,” in *Proceedings of the 24th ACM International Conference on Multimedia*, MM ’16, (New York, NY, USA), p. 407–411, Association for Computing Machinery, 2016.
- [495] H. Wang, Y. Cai, Y. Zhang, H. Pan, W. Lv, and H. Han, “Deep learning for image retrieval: What works and what doesn’t,” in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 1576–1583, 2015.

- [496] V. Bhandi and K. A. Sumithra Devi, "Image Retrieval by Fusion of Features from Pre-trained Deep Convolution Neural Networks," in *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, pp. 35–40, 2019.
- [497] S. Maji and S. Bose, "CBIR using features derived by deep learning," *CoRR*, vol. abs/2002.07877, 2020.
- [498] S. Denner, D. Zimmerer, D. Bounias, M. Bujotzek, S. Xiao, L. Kausch, P. Schader, T. Penzkofer, P. F. Jäger, and K. Maier-Hein, "Leveraging foundation models for content-based medical image retrieval in radiology," 2024.
- [499] C. H. Song, J. Yoon, S. Choi, and Y. Avrithis, "Boosting vision transformers for image retrieval," 2022.
- [500] S. R. Dubey, S. K. Singh, and W.-T. Chu, "Vision Transformer Hashing for Image Retrieval," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2022.
- [501] H.-Y. Zhou, X. Chen, Y. Zhang, R. Luo, L. Wang, and Y. Yu, "Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports," *Nature Machine Intelligence*, vol. 4, no. 1, 2022.
- [502] A. Sain, A. K. Bhunia, P. N. Chowdhury, S. Koley, T. Xiang, and Y.-Z. Song, "Clip for all things zero-shot sketch-based image retrieval, fine-grained or not," 2023.
- [503] Y. Kumar and P. Marttinen, "Improving medical multi-modal contrastive learning with expert annotations," 2024.
- [504] K. You, J. Gu, J. Ham, B. Park, J. Kim, E. K. Hong, W. Baek, and B. Roh, "Cxr-clip: Toward large scale chest x-ray language-image pre-training," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 101–111, Springer, 2023.
- [505] M. Aono, H. Shinoda, T. Asakawa, K. Shimizu, T. Togawa, and T. Komoda, "Multi-stage medical image captioning using classification and clip.," in *CLEF (Working Notes)*, pp. 1387–1395, 2023.
- [506] Y. Gao, H.-Y. Zhou, X. Wang, T. Zhang, R. Tan, L. Han, L. Estacio, A. D'Angelo, J. Teuwen, R. Mann, and T. Tan, "Visualize what you learn: a well-explainable joint-learning framework based on multi-view mammograms and associated reports," 2023.
- [507] G. Wang, Y. Guo, Z. Xu, Y. Wong, and M. S. Kankanhalli, "Semantic-aware triplet loss for image classification," *IEEE Transactions on Multimedia*, vol. 25, 2023.
- [508] Hugging Face, "Hugging Face - The AI Community." <https://huggingface.co/>, 2025. [Online. Accessed 22-January-2025].
- [509] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT Pre-Training of Image Transformers," in *International Conference on Learning Representations*, 2022.
- [510] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Burststein,

- Jill and Doran, Christy and Solorio, Tamar, ed.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.
- [511] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “DINOv2: Learning Robust Visual Features without Supervision,” *Transactions on Machine Learning Research*, 2024.
- [512] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, “Visual Transformers: Token-based Image Representation and Processing for Computer Vision,” 2020.
- [513] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.
- [514] D’Ascoli, Stéphane and Touvron, Hugo and Leavitt, Matthew L and Morcos, Ari S and Biroli, Giulio and Sagun, Levent, “ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases,” in *Proceedings of the 38th International Conference on Machine Learning* (Meila, Marina and Zhang, Tong, ed.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 2286–2296, PMLR, 18–24 Jul 2021.
- [515] Chen, Chun-Fu (Richard) and Fan, Quanfu and Panda, Rameswar, “CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 357–366, October 2021.
- [516] Ding, Mingyu and Xiao, Bin and Codella, Noel and Luo, Ping and Wang, Jingdong and Yuan, Lu, “DaViT: Dual Attention Vision Transformers,” in *Computer Vision – ECCV 2022* (Avidan, Shai and Brostow, Gabriel and Cissé, Moustapha and Farinella, Giovanni Maria and Hassner, Tal, ed.), (Cham), pp. 74–92, Springer Nature Switzerland, 2022.
- [517] Graham, Ben and El-Nouby, Alaaeldin and Touvron, Hugo and Stock, Pierre and Joulin, Armand and Jegou, Herve and Douze, Matthijs, “LeViT: a Vision Transformer in ConvNet’s Clothing for Faster Inference,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (Los Alamitos, CA, USA), pp. 12239–12249, IEEE Computer Society, October 2021.
- [518] Tu, Zhengzhong and Talebi, Hossein and Zhang, Han and Yang, Feng and Milanfar, Peyman and Bovik, Alan and Li, Yinxiao, “MaxViT: Multi-axis Vision Transformer,” in *Computer Vision – ECCV 2022* (Avidan, Shai and Brostow, Gabriel and Cissé, Moustapha and Farinella, Giovanni Maria and Hassner, Tal, ed.), (Cham), pp. 459–479, Springer Nature Switzerland, 2022.
- [519] Li, Yanghao and Wu, Chao-Yuan and Fan, Haoqi and Mangalam, Karttikeya and Xiong, Bo and Malik, Jitendra and Feichtenhofer, Christoph, “MViTv2: Improved Multiscale Vision Transformers for Classification and Detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4804–4814, June 2022.

- [520] Mahmood Lab, “Mahmood Lab - Computational Pathology.” <https://faisal.ai>, 2025. [Online. Accessed 22-January-2025].
- [521] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, *et al.*, “A visual-language foundation model for computational pathology,” *Nature Medicine*, vol. 30, p. 863–874, 2024.
- [522] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, B. Chen, A. Zhang, D. Shao, A. H. Song, M. Shaban, *et al.*, “Towards a General-Purpose Foundation Model for Computational Pathology,” *Nature Medicine*, 2024.
- [523] T. Ding, S. J. Wagner, A. H. Song, R. J. Chen, M. Y. Lu, A. Zhang, A. J. Vaidya, G. Jaume, M. Shaban, A. Kim, D. F. K. Williamson, B. Chen, C. Almagro-Perez, P. Doucet, S. Sahai, C. Chen, D. Komura, A. Kawabe, S. Ishikawa, G. Gerber, T. Peng, L. P. Le, and F. Mahmood, “Multimodal Whole Slide Foundation Model for Pathology,” 2024.
- [524] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, M. Zhao, A. K. Chow, K. Ikemura, A. Kim, D. Pouli, A. Patel, *et al.*, “A multimodal generative AI copilot for human pathology,” *Nature*, vol. 634, no. 8033, pp. 466–473, 2024.
- [525] G. Jaume, A. Vaidya, R. Chen, D. Williamson, P. Liang, and F. Mahmood, “Modeling Dense Multimodal Interactions Between Biological Pathways and Histology for Survival Prediction,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [526] R. J. Chen, M. Y. Lu, D. F. Williamson, T. Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo, and F. Mahmood, “Pan-cancer integrative histology-genomic analysis via multimodal deep learning,” *Cancer Cell*, vol. 40, no. 8, pp. 865–878.e6, 2022.
- [527] R. Li, X. Wu, A. Li, and M. Wang, “HFBSurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction,” *Bioinformatics*, vol. 38, no. 9, pp. 2587–2594, 2022.
- [528] P. Xu, X. Zhu, and D. A. Clifton, “Multimodal Learning With Transformers: A Survey,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, pp. 12113–12132, October 2023.
- [529] R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. Williamson, T. Manz, M. Shady, and F. Mahmood, “Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3995–4005, 2021.
- [530] G. Jaume, A. Vaidya, R. J. Chen, D. F. Williamson, P. P. Liang, and F. Mahmood, “Modeling Dense Multimodal Interactions Between Biological Pathways and Histology for Survival Prediction,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11579–11590, 2024.
- [531] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. van der Laak, “Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology,” *Medical Image Analysis*, vol. 58, p. 101544, 2019.

- [532] K. Faryna, J. van der Laak, and G. Litjens, “Automatic data augmentation to improve generalization of deep learning in H&E stained histopathology,” *Computers in Biology and Medicine*, vol. 170, p. 108018, 2024.
- [533] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, “Fooling lime and shap: Adversarial attacks on post hoc explanation methods,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.
- [534] A. Subramanya, V. Pillai, and H. Pirsiavash, “Fooling network interpretation in image classification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2020–2029, 2019.
- [535] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, and C. Shah, “Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review,” *arXiv preprint arXiv:2010.10596*, 2020.
- [536] I. Y. Chen, S. Joshi, and M. Ghassemi, “Treating health disparities with artificial intelligence,” *Nature Medicine*, vol. 26, pp. 16–17, Jan. 2020.
- [537] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, “Post-comparison mitigation of demographic bias in face recognition using fair score normalization,” *Pattern Recognition Letters*, vol. 140, pp. 332–338, 2020.
- [538] M. E. Kaminski, “The Right to Explanation, Explained,” *Berkeley Tech. LJ*, vol. 34, p. 189, 2019.
- [539] R. Moghaddass, C. Rudin, and D. Madigan, “The Factorized Self-Controlled Case Series Method: An Approach for Estimating the Effects of Many Drugs on Many Outcomes.,” *J. Mach. Learn. Res.*, vol. 17, no. 185, pp. 1–24, 2016.
- [540] C. Farrington, “Relative incidence estimation from case series for vaccine safety evaluation,” *Biometrics*, pp. 228–235, 1995.
- [541] I. Petersen, I. Douglas, and H. Whitaker, “Self controlled case series methods: an alternative to standard epidemiological study designs,” *bmj*, vol. 354, 2016.
- [542] J. Ghosh and D. B. Dunson, “Default prior distributions and efficient posterior computation in Bayesian factor analysis,” *Journal of Computational and Graphical Statistics*, vol. 18, no. 2, pp. 306–320, 2009.
- [543] H. Lakkaraju and C. Rudin, “Learning cost-effective treatment regimes using markov decision processes,” *arXiv preprint arXiv:1610.06972*, 2016.
- [544] B. Ustun and C. Rudin, “Supersparse linear integer models for optimized medical scoring systems,” *Machine Learning*, vol. 102, no. 3, pp. 349–391, 2016.
- [545] B. Ustun, M. B. Westover, C. Rudin, and M. T. Bianchi, “Clinical prediction models for sleep apnea: the importance of medical history over symptoms,” *Journal of Clinical Sleep Medicine*, vol. 12, no. 2, pp. 161–168, 2016.
- [546] H. Zhou, C. Cheng, Z. C. Lipton, G. H. Chen, and J. C. Weiss, “Predicting Mortality Risk in Viral and Unspecified Pneumonia to Assist Clinicians with COVID-19 ECMO Planning,” *arXiv preprint arXiv:2006.01898*, 2020.

- [547] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for Cox’s proportional hazards model via coordinate descent,” *Journal of Statistical Software*, vol. 39, no. 5, p. 1, 2011.
- [548] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, “Concept bottleneck models,” in *International Conference on Machine Learning*, pp. 5338–5348, PMLR, 2020.
- [549] M. A. M. Afnan, C. Rudin, V. Conitzer, J. Savulescu, A. Mishra, Y. Liu, and M. Afnan, “Ethical implementation of artificial intelligence to select embryos in In Vitro Fertilization,” *arXiv preprint arXiv:2105.00060*, 2021.
- [550] Z. C. Lipton, A. Chouldechova, and J. McAuley, “Does mitigating ML’s impact disparity require treatment disparity?,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8136–8146, 2018.
- [551] T. Miller, P. Howe, and L. Sonenberg, “Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences,” *arXiv preprint arXiv:1712.00547*, 2017.
- [552] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, “Generating Counterfactual Explanations with Natural Language,” in *ICML Workshop on Human Interpretability in Machine Learning*, pp. 95–98, 2018.
- [553] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, “Counterfactual Visual Explanations,” in *International Conference on Machine Learning*, pp. 2376–2384, PMLR, 2019.
- [554] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, “Generative Counterfactual Introspection for Explainable Deep Learning,” in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5, IEEE, 2019.
- [555] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, “AttGAN: Facial Attribute Editing by Only Changing What You Want,” *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.
- [556] R. Balasubramanian, S. Sharpe, B. Barr, J. Wittenbach, and C. B. Bruss, “Latent-CF: A Simple Baseline for Reverse Counterfactual Explanations,” *arXiv preprint arXiv:2012.09301*, 2020.
- [557] A. Akula, S. Wang, and S.-C. Zhu, “CoCoX: Generating Conceptual and Counterfactual Explanations via Fault-Lines,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 2594–2601, 2020.
- [558] P. Wang and N. Vasconcelos, “SCOUT: Self-Aware Discriminant Counterfactual Explanations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [559] A. Feghahati, C. R. Shelton, M. J. Pazzani, and K. Tang, “CDeepEx: Contrastive Deep Explanations,” in *ECAI 2020*, pp. 1143–1151, IOS Press, 2020.

- [560] S. C. Smith and S. Ramamoorthy, “Counterfactual Explanation and Causal Inference In Service of Robustness in Robot Control,” in *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 1–8, IEEE, 2020.
- [561] Y. Zhao, “Fast Real-time Counterfactual Explanations,” *arXiv preprint arXiv:2007.05684*, 2020.
- [562] G. Ali, F. Al-Obeidat, A. Tubaishat, T. Zia, M. Ilyas, and A. Rocha, “Counterfactual explanation of Bayesian model uncertainty,” *Neural Computing and Applications*, pp. 1–8, 2021.
- [563] K. Alipour, A. Ray, X. Lin, M. Cogswell, J. P. Schulze, Y. Yao, and G. T. Burachas, “Improving users’ mental model with attention-directed counterfactual edits,” *Applied AI Letters*, vol. 2, no. 4, p. e47, 2021.
- [564] A. Elliott, S. Law, and C. Russell, “Explaining Classifiers using Adversarial Perturbations on the Perceptual Ball,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10693–10702, 2021.
- [565] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution,” in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pp. 694–711, Springer, 2016.
- [566] S. Gulshad and A. Smeulders, “Counterfactual attribute-based visual explanations for classification,” *International Journal of Multimedia Information Retrieval*, vol. 10, pp. 127–140, 2021.
- [567] F. Hvilshøj, A. Iosifidis, and I. Assent, “ECINN: Efficient Counterfactuals from Invertible Neural Networks,” *arXiv preprint arXiv:2103.13701*, 2021.
- [568] L. Dinh, D. Krueger, and Y. Bengio, “NICE: Non-linear Independent Components Estimation,” *arXiv preprint arXiv:1410.8516*, 2014.
- [569] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using Real NVP,” *arXiv preprint arXiv:1605.08803*, 2016.
- [570] D. P. Kingma and P. Dhariwal, “Glow: Generative Flow with Invertible 1×1 Convolutions,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [571] J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J.-H. Jacobsen, “Invertible Residual Networks,” in *International Conference on Machine Learning*, pp. 573–582, PMLR, 2019.
- [572] B. Höltingen, L. Schut, J. M. Brauner, and Y. Gal, “DeDUCE: Generating Counterfactual Explanations Efficiently,” *arXiv preprint arXiv:2111.15639*, 2021.
- [573] J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. Torr, and Y. Gal, “Deterministic Neural Networks with Inductive Biases Capture Epistemic and Aleatoric Uncertainty,” *arXiv preprint arXiv:2102.11582*, 2021.

- [574] L. Smith and Y. Gal, “Understanding Measures of Uncertainty for Adversarial Example Detection,” *arXiv preprint arXiv:1803.08533*, 2018.
- [575] L. Schut, O. Key, R. Mc Grath, L. Costabello, B. Sacaleanu, Y. Gal, *et al.*, “Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties,” in *International Conference on Artificial Intelligence and Statistics*, pp. 1756–1764, PMLR, 2021.
- [576] E. M. Kenny and M. T. Keane, “On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 11575–11585, 2021.
- [577] J. Labaien, E. Zugasti, and X. De Carlos, “DA-DGCEX: Ensuring Validity of Deep Guided Counterfactual Explanations With Distribution-Aware Autoencoder Loss,” *arXiv preprint arXiv:2104.09062*, 2021.
- [578] Y. Li, S. Liu, C. Wu, X. Xi, G. Cao, and W. Cao, “Dcfcg: Discovering directional counterfactual generation for chest x-rays,” in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 972–979, IEEE, 2021.
- [579] O.-I. Popescu, M. Shadaydeh, and J. Denzler, “Counterfactual Generation with Knockoffs,” *arXiv preprint arXiv:2102.00951*, 2021.
- [580] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, “Explaining Image Classifiers by Counterfactual Generation,” *arXiv preprint arXiv:1807.08024*, 2018.
- [581] R. F. Barber and E. J. Candès, “Controlling the false discovery rate via knockoffs,” *The Annals of Statistics*, vol. 43, no. 5, pp. 2055 – 2085, 2015.
- [582] D. S. Watson and M. N. Wright, “Testing conditional independence in supervised learning algorithms,” *Machine Learning*, vol. 110, no. 8, pp. 2107–2129, 2021.
- [583] P. Rodriguez, M. Caccia, A. Lacoste, L. Zamparo, I. Laradji, L. Charlin, and D. Vazquez, “Beyond Trivial Counterfactual Explanations With Diverse Valuable Explanations,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1056–1065, 2021.
- [584] J. Thiagarajan, V. S. Narayanaswamy, D. Rajan, J. Liang, A. Chaudhari, and A. Spanias, “Designing Counterfactual Generators using Deep Model Inversion,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 16873–16884, 2021.
- [585] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, “Dreaming to Distill: Data-Free Knowledge Transfer via DeepInversion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [586] A. Van Looveren, J. Klaise, G. Vacanti, and O. Cobb, “Conditional Generative Models for Counterfactual Explanations,” *arXiv preprint arXiv:2101.10123*, 2021.
- [587] A. White, K. H. Ngan, J. Phelan, S. S. Afgeh, K. Ryan, C. C. Reyes-Aldasoro, and A. d. Garcez, “Contrastive Counterfactual Visual Explanations With Overdetermination,” *arXiv preprint arXiv:2106.14556*, 2021.

- [588] A. White and A. d. Garcez, “Measurable Counterfactual Local Explanations for Any Classifier,” *arXiv preprint arXiv:1908.03020*, 2019.
- [589] F. Yang, N. Liu, M. Du, and X. Hu, “Generative Counterfactuals for Neural Networks via Attribute-Informed Perturbation,” *ACM SIGKDD Explorations Newsletter*, vol. 23, no. 1, pp. 59–68, 2021.
- [590] R. K. Mothilal, A. Sharma, and C. Tan, “Explaining machine learning classifiers through diverse counterfactual explanations,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* ’20*, (New York, NY, USA), p. 607–617, Association for Computing Machinery, 2020.
- [591] A. Kulesza and B. Taskar, “Determinantal Point Processes for Machine Learning,” *Foundations and Trends® in Machine Learning*, vol. 5, no. 2–3, pp. 123–286, 2012.
- [592] C. Hsieh, C. Moreira, and C. Ouyang, “DiCE4EL: Interpreting Process Predictions using a Milestone-Aware Counterfactual Approach,” in *2021 3rd International Conference on Process Mining (ICPM)*, pp. 88–95, 2021.
- [593] A. R. Akula, K. Wang, C. Liu, S. Saba-Sadiya, H. Lu, S. Todorovic, J. Chai, and S.-C. Zhu, “CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models,” *Iscience*, vol. 25, no. 1, p. 103581, 2022.
- [594] J. E. Laird, C. Lebiere, and P. S. Rosenbloom, “A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics,” *Ai Magazine*, vol. 38, no. 4, pp. 13–26, 2017.
- [595] A. Abid, M. Yuksekgonul, and J. Zou, “Meaningfully Debugging Model Mistakes using Conceptual Counterfactual Explanations,” in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 66–88, PMLR, 17–23 Jul 2022.
- [596] S. Dash, V. N. Balasubramanian, and A. Sharma, “Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 915–924, 2022.
- [597] G. Filandrianos, K. Thomas, E. Dervakos, and G. Stamou, “Conceptual Edits as Counterfactual Explanations,” in *AAAI Spring Symposium: MAKE*, 2022.
- [598] A. Ghandeharioun, B. Kim, C.-L. Li, B. Jou, B. Eoff, and R. Picard, “DISSECT: Disentangled Simultaneous Explanations via Concept Traversals,” in *International Conference on Learning Representations*, 2022.
- [599] V. Guyomard, F. Fessant, T. Bouadi, and T. Guyet, “Post-hoc counterfactual generation with supervised autoencoder,” in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I*, pp. 105–114, Springer, 2022.
- [600] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord, “STEEX: Steering Counterfactual Explanations with Semantics,” in *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pp. 387–403, Springer, 2022.

- [601] G. Jeanneret, L. Simon, and F. Jurie, “Diffusion Models for Counterfactual Explanations,” in *Proceedings of the Asian Conference on Computer Vision*, pp. 858–876, 2022.
- [602] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [603] J. Song, C. Meng, and S. Ermon, “Denoising Diffusion Implicit Models,” in *International Conference on Learning Representations*, 2021.
- [604] S. Khorram and L. Fuxin, “Cycle-Consistent Counterfactuals by Latent Transformations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10203–10212, 2022.
- [605] S. Mertes, T. Huber, K. Weitz, A. Heimerl, and E. André, “GANterfactual—Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning,” *Frontiers in Artificial Intelligence*, vol. 5, 2022.
- [606] K. Oh, J. S. Yoon, and H.-I. Suk, “Born Identity Network: Multi-way Counterfactual Map Generation to Explain a Classifier’s Decision,” *arXiv preprint arXiv:2011.10381*, 2020.
- [607] K. Oh, J. S. Yoon, and H.-I. Suk, “Learn-Explain-Reinforce: Counterfactual Reasoning and Its Guidance to Reinforce an Alzheimer’s Disease Diagnosis Model,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [608] P. Sanchez and S. A. Tsafaris, “Diffusion Causal Models for Counterfactual Estimation,” in *Proceedings of the First Conference on Causal Learning and Reasoning* (B. Schölkopf, C. Uhler, and K. Zhang, eds.), vol. 177 of *Proceedings of Machine Learning Research*, pp. 647–668, PMLR, 11–13 Apr 2022.
- [609] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
- [610] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.
- [611] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.
- [612] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-Based Generative Modeling through Stochastic Differential Equations,” in *International Conference on Learning Representations*, 2021.
- [613] S. Vandenhende, D. Mahajan, F. Radenovic, and D. Ghadiyaram, “Making Heads or Tails: Towards Semantically Consistent Visual Counterfactuals,” in *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pp. 261–279, Springer, 2022.
- [614] T. Vermeire, D. Brughmans, S. Goethals, R. M. B. de Oliveira, and D. Martens, “Explainable image classification with evidence counterfactual,” *Pattern Analysis and Applications*, vol. 25, no. 2, pp. 315–335, 2022.

- [615] D. Martens and F. Provost, “Explaining Data-Driven Document Classifications,” *MIS quarterly*, vol. 38, no. 1, pp. 73–100, 2014.
- [616] N. Yang, T. Kang, and K. Jung, “Deriving Explainable Discriminative Attributes Using Confusion About Counterfactual Class,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1730–1734, IEEE, 2022.
- [617] Z. Wang, I. Samsten, V. Kougia, and P. Papapetrou, “Style-transfer counterfactual explanations: An application to mortality prevention of ICU patients,” *Artificial Intelligence in Medicine*, vol. 135, p. 102457, 2023.
- [618] S. Singla, M. Eslami, B. Pollack, S. Wallace, and K. Batmanghelich, “Explaining the black-box smoothly—A counterfactual approach,” *Medical Image Analysis*, vol. 84, p. 102721, 2023.
- [619] G. Pombo, R. Gray, M. J. Cardoso, S. Ourselin, G. Rees, J. Ashburner, and P. Nachev, “Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models,” *Medical Image Analysis*, vol. 84, p. 102723, 2023.
- [620] A. Nasr-Esfahany, M. Alizadeh, and D. Shah, “Counterfactual Identifiability of Bijective Causal Models,” in *Proceedings of the 40th International Conference on Machine Learning* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds.), vol. 202 of *Proceedings of Machine Learning Research*, pp. 25733–25754, PMLR, 23–29 Jul 2023.
- [621] O. Hundogan, X. Lu, Y. Du, and H. A. Reijers, “CREATED: Generating Viable Counterfactual Sequences for Predictive Process Analytics,” in *Advanced Information Systems Engineering* (M. Indulska, I. Reinhartz-Berger, C. Cetina, and O. Pastor, eds.), (Cham), pp. 541–557, Springer Nature Switzerland, 2023.
- [622] A. Vlontzos, B. Kainz, and C. M. Gilligan-Lee, “Estimating categorical counterfactuals via deep twin networks,” *Nature Machine Intelligence*, vol. 5, no. 2, pp. 159–168, 2023.
- [623] C. F. Dantas, D. Marcos, and D. Ienco, “Towards Explainable Land Cover Mapping: a Counterfactual-based Strategy,” *arXiv preprint arXiv:2301.01520*, 2023.
- [624] Y. Ying, F. Zhuang, Y. Zhu, D. Wang, and H. Zheng, “Camus: Attribute-aware counterfactual augmentation for minority users in recommendation,” in *Proceedings of the ACM Web Conference 2023*, pp. 1396–1404, 2023.
- [625] I. Takigawa, M. Kudo, and A. Nakamura, “Convex sets as prototypes for classifying patterns,” *Engineering Applications of Artificial Intelligence*, vol. 22, no. 1, pp. 101–108, 2009.
- [626] J. Bien and R. Tibshirani, “Prototype selection for interpretable classification,” *The Annals of Applied Statistics*, pp. 2403–2424, 2011.
- [627] J. Snell, K. Swersky, and R. Zemel, “Prototypical Networks for Few-shot Learning,” *Advances in neural information processing systems*, vol. 30, 2017.
- [628] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, “Transferrable Prototypical Networks for Unsupervised Domain Adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

- [629] Z. Yang, Y. Han, G. Yu, Q. Yang, and X. Zhang, “Prototypical Networks for Multi-Label Learning,” *arXiv preprint arXiv:1911.07203*, 2019.
- [630] W. Kim, K. Park, K. Sohn, R. Shu, and H.-S. Kim, “Federated Semi-Supervised Learning with Prototypical Networks,” *arXiv preprint arXiv:2205.13921*, 2022.
- [631] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.
- [632] W. Xu, Y. Xian, J. Wang, B. Schiele, and Z. Akata, “Attribute Prototype Network for Zero-Shot Learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 21969–21980, 2020.
- [633] W. Xu, Y. Xian, J. Wang, B. Schiele, and Z. Akata, “Attribute Prototype Network for Any-Shot Learning,” *International Journal of Computer Vision*, vol. 130, no. 7, pp. 1735–1753, 2022.
- [634] S. Ö. Arik and T. Pfister, “ProtoAttend: attention-based prototypical learning,” *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 8691–8725, 2020.
- [635] E. Kim, S. Kim, M. Seo, and S. Yoon, “XProtoNet: Diagnosis in Chest Radiography with Global and Local Explanations,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15719–15728, 2021.
- [636] J. Wang, H. Liu, X. Wang, and L. Jing, “Interpretable Image Recognition by Constructing Transparent Embedding Space,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 895–904, 2021.
- [637] D. Rymarczyk, Ł. Struski, J. Tabor, and B. Zieliński, “ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1420–1430, 2021.
- [638] G. Singh and K.-C. Yow, “These do not Look Like Those: An Interpretable Deep Learning Model for Image Recognition,” *IEEE Access*, vol. 9, pp. 41482–41493, 2021.
- [639] G. Singh and K.-C. Yow, “An Interpretable Deep Learning Model for Covid-19 Detection With Chest X-Ray Images,” *Ieee Access*, vol. 9, pp. 85198–85208, 2021.
- [640] G. Singh and K.-C. Yow, “Object or Background: An Interpretable Deep Learning Model for COVID-19 Detection from CT-Scan Images,” *Diagnostics*, vol. 11, no. 9, p. 1732, 2021.
- [641] G. Singh, “Think positive: An interpretable neural network for image recognition,” *Neural Networks*, vol. 151, pp. 178–189, 2022.
- [642] S. F. Stefenon, G. Singh, K.-C. Yow, and A. Cimatti, “Semi-ProtoPNet Deep Neural Network for the Classification of Defective Power Grid Distribution Structures,” *Sensors*, vol. 22, no. 13, p. 4859, 2022.
- [643] 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*, vol. 3, no. 12, pp. 1061–1070, 2021.

- [644] G. Carloni, A. Berti, C. Iacconi, M. A. Pascali, and S. Colantonio, “On the Applicability of Prototypical Part Learning in Medical Images: Breast Masses Classification Using ProtoPNet,” 2022.
- [645] P. Hase, C. Chen, O. Li, and C. Rudin, “Interpretable Image Recognition with Hierarchical Prototypes,” *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, pp. 32–40, Oct. 2019.
- [646] V. G. Satorras and J. B. Estrach, “Few-Shot Learning with Graph Neural Networks,” in *International Conference on Learning Representations (ICLR)*, 2018.
- [647] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, and H. Larochelle, “Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples,” in *International Conference on Learning Representations (ICLR)*, 2020.
- [648] T. Cao, M. T. Law, and S. Fidler, “A Theoretical Analysis of the Number of Shots in Few-Shot Learning,” in *International Conference on Learning Representations (ICLR)*, 2020.
- [649] B. Oreshkin, P. Rodríguez López, and A. Lacoste, “TADAM: Task dependent adaptive metric for improved few-shot learning,” in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.
- [650] S. Sun, Q. Sun, K. Zhou, and T. Lv, “Hierarchical Attention Prototypical Networks for Few-Shot Text Classification,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 476–485, Association for Computational Linguistics, November 2019.
- [651] K. Allen, E. Shelhamer, H. Shin, and J. Tenenbaum, “Infinite Mixture Prototypes for Few-shot Learning,” in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 232–241, PMLR, 09–15 June 2019.
- [652] N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, *Bayesian Nonparametrics*, vol. 28. Cambridge University Press, 2010.
- [653] S. W. Yoon, J. Seo, and J. Moon, “TapNet: Neural Network Augmented with Task-Adaptive Projection for Few-Shot Learning,” in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 7115–7123, PMLR, 09–15 June 2019.
- [654] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, “Few-Shot Learning via Embedding Adaptation With Set-to-Set Functions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [655] X. Zhen, Y. Du, H. Xiong, Q. Qiu, C. Snoek, and L. Shao, “Learning to Learn Variational Semantic Memory,” in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 9122–9134, Curran Associates, Inc., 2020.

- [656] J. Donnelly, A. J. Barnett, and C. Chen, “Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10265–10275, 2022.
- [657] M. Xue, Q. Huang, H. Zhang, L. Cheng, J. Song, M. Wu, and M. Song, “ProtoPFormer: Concentrating on Prototypical Parts in Vision Transformers for Interpretable Image Recognition,” *arXiv preprint arXiv:2208.10431*, 2022.
- [658] P. L. Vidal, J. de Moura, J. Novo, M. Ortega, and J. S. Cardoso, “Transformer-Based Multi-Prototype Approach for Diabetic Macular Edema Analysis in OCT Images,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.
- [659] A. Ragno, B. La Rosa, and R. Capobianco, “Prototype-based Interpretable Graph Neural Networks,” *IEEE Transactions on Artificial Intelligence*, 2022.
- [660] H. Wang, X. Bai, J. Tan, and J. Yang, “Deep prototypical networks based domain adaptation for fault diagnosis,” *Journal of Intelligent Manufacturing*, vol. 33, no. 4, pp. 973–983, 2022.
- [661] W. Fu, L. Zhou, and J. Chen, “Bidirectional Matching Prototypical Network for Few-Shot Image Classification,” *IEEE Signal Processing Letters*, vol. 29, pp. 982–986, 2022.
- [662] M. Gogoi, S. Tiwari, and S. Verma, “Adaptive Prototypical Networks,” *arXiv preprint arXiv:2211.12479*, 2022.
- [663] C. Wang, Y. Liu, Y. Chen, F. Liu, Y. Tian, D. J. McCarthy, H. Frazer, and G. Carneiro, “Learning Support and Trivial Prototypes for Interpretable Image Classification,” *arXiv preprint arXiv:2301.04011*, 2023.
- [664] L. Gallée, M. Beer, and M. Götz, “Interpretable Medical Image Classification Using Prototype Learning and Privileged Information,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023* (H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, eds.), (Cham), pp. 435–445, Springer Nature Switzerland, 2023.
- [665] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic Routing Between Capsules,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [666] V. Vapnik and A. Vashist, “A new learning paradigm: Learning using privileged information,” *Neural Networks*, vol. 22, no. 5-6, pp. 544–557, 2009.
- [667] V. Vapnik, R. Izmailov, *et al.*, “Learning using privileged information: similarity control and knowledge transfer,” *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 2023–2049, 2015.
- [668] C. Ma, B. Zhao, C. Chen, and C. Rudin, “This Looks Like Those: Illuminating Prototypical Concepts Using Multiple Visualizations,” in *Advances in Neural Information Processing Systems* (A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 39212–39235, Curran Associates, Inc., 2023.
- [669] Q. Lyu and W. Wang, “Compositional Prototypical Networks for Few-Shot Classification,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 9011–9019, June 2023.

- [670] W. An, F. Tian, Q. Zheng, W. Ding, Q. Wang, and P. Chen, “Generalized Category Discovery with Decoupled Prototypical Network,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 12527–12535, June 2023.
- [671] J. Y. Lim, K. M. Lim, C. P. Lee, and Y. X. Tan, “SSL-ProtoNet: Self-supervised Learning Prototypical Networks for few-shot learning,” *Expert Systems with Applications*, vol. 238, p. 122173, 2024.
- [672] L. Fernandes, J. N. D. Fernandes, M. Calado, J. R. Pinto, R. Cerqueira, and J. S. Cardoso, “Intrinsic Explainability for End-to-End Object Detection,” *IEEE Access*, vol. 12, pp. 2623–2634, 2024.
- [673] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [674] Y. Du, Z. Xiao, S. Liao, and C. Snoek, “ProtoDiff: Learning to Learn Prototypical Networks by Task-Guided Diffusion,” in *Advances in Neural Information Processing Systems* (A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 46304–46322, Curran Associates, Inc., 2023.
- [675] A. Hoffmann, C. Fanconi, R. Rade, and J. Kohler, “This Looks Like That... Does it? Shortcomings of Latent Space Prototype Explainability in Deep Networks,” in *ICML 2021 Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*, 05 2021.
- [676] Q. Huang, M. Xue, W. Huang, H. Zhang, J. Song, Y. Jing, and M. Song, “Evaluation and Improvement of Interpretability for Self-Explainable Part-Prototype Networks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2011–2020, October 2023.
- [677] A. Bontempelli, S. Teso, K. Tentori, F. Giunchiglia, and A. Passerini, “Concept-level Debugging of Part-Prototype Networks,” in *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.
- [678] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, “Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications,” *Proceedings of the IEEE*, vol. 109, pp. 247–278, March 2021.
- [679] T. Gonçalves, I. Rio-Torto, L. F. Teixeira, and J. S. Cardoso, “A Survey on Attention Mechanisms for Medical Applications: Are We Moving Toward Better Algorithms?,” *IEEE Access*, vol. 10, pp. 98909–98935, 2022.
- [680] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis,” *Medical Image Analysis*, vol. 79, p. 102470, 2022.
- [681] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin and Marina Marina M.-C. Höhne, “Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond,” *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1–11, 2023.

- [682] M. Böhle, N. Singh, M. Fritz, and B. Schiele, “B-cos Alignment for Inherently Interpretable CNNs and Vision Transformers,” *arXiv:2306.10898*, 2023.
- [683] Z. Liu, R. Xiong, and T. Jiang, “CI-Net: Clinical-Inspired Network for Automated Skin Lesion Recognition,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 619–632, 2023.
- [684] C. Matsoukas, J. F. Haslum, M. Sorkhei, M. Söderberg, and K. Smith, “What Makes Transfer Learning Work for Medical Images: Feature Reuse & Other Factors,” in *CVPR*, pp. 9225–9234, 2022.
- [685] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, “Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications,” *Information Fusion*, vol. 81, pp. 59–83, 2022.
- [686] A. Ramdas, N. G. Trillos, and M. Cuturi, “On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests,” *Entropy*, vol. 19, no. 2, 2017.
- [687] J. L. Fleiss, “Measuring nominal scale agreement among many raters,” *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.
- [688] J. J. Randolph, “Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss’ Fixed-Marginal Multirater Kappa,” *Online submission*, 2005.
- [689] R. Falotico and P. Quatto, “Fleiss’ kappa statistic without paradoxes,” *Quality & Quantity*, vol. 49, pp. 463–470, 2015.
- [690] D. Slack, A. Hilgard, H. Lakkaraju, and S. Singh, “Counterfactual Explanations Can Be Manipulated,” in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 62–75, Curran Associates, Inc., 2021.