

# Visually Explaining Decisions of Deep Neural Network Classifiers in Ophthalmology

## **Dissertation**

der Mathematisch-Naturwissenschaftlichen Fakultät  
der Eberhard Karls Universität Tübingen  
zur Erlangung des Grades eines  
Doktors der Naturwissenschaften  
(Dr. rer. nat.)

vorgelegt von  
Indu Ilanchezian  
aus Madurai, Indien

Tübingen  
2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 16.09.2024

Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Philipp Berens
2. Berichterstatter:	Prof. Dr. Sven Nahnsen

# Abstract

Convolutional Neural Networks (CNNs) have emerged as powerful tools in ophthalmology by exhibiting human-level performance in the classification of various ocular conditions and diseases. However, the decisions of these models are opaque and hard to interpret, which limits their trustworthiness and applicability in clinical settings. In this thesis, we address the challenges through: an inherently interpretable model architecture called BagNet and counterfactual explanations. BagNets use small receptive fields to identify local features in an image that contribute significantly to the model’s decision. On the other hand, counterfactual explanations show the changes required on a given input image to alter the decision of the classifier to a particular class. Importantly, both of these explanation methods share similarities with processes employed by humans to offer explanations for their decisions.

Intriguingly, CNNs demonstrate remarkable accuracy in predicting gender from retinal fundus images even though it was previously unknown to ophthalmologists that retinal fundus images encoded gender information. Here, it would be beneficial to explain the CNN model’s decisions in order to identify features that the model uses for distinguishing between male and female fundus images. To this end, we utilized the local feature importance estimates from BagNets to produce saliency maps that highlight informative patches in fundus images. Our analysis revealed that patches from the optic disc and macula contribute significantly, with the former favoring detection of male fundus images and the latter, female. We conclude that BagNets are feasible alternatives to standard CNN architectures which have the potential to serve as an effective approach to provide explanations in medical image analysis tasks.

Following our study on explanations from BagNets, we investigated the generation of counterfactual images from CNN classifiers to provide explanations. Specifically, we assessed various counterfactual generation techniques across a range of retinal disease classifiers in ophthalmology. The first technique relied on the generation of counterfactual images solely using the gradients of a classifier with respect to the input. Here, adversarially robust models offered more interpretable gradients than a standard classifier although at the expense of reduced accuracy. We combined the strengths of both approaches by ensembling a standard CNN with an adversarially robust one. Our ensemble method achieved high accuracies comparable to the standard CNN while also generating meaningful visual counterfactuals. However, a notable limitation of this classifier-only approach is a lack of realism of the generated counterfactuals.

To achieve realism, the second technique employed a diffusion model alongside adversarially robust and plain classifiers trained on retinal disease classification tasks from color fundus photographs and optical coherence tomography (OCT) B-scans. The gradients of the classifiers guide the diffusion model effectively, enabling it to add or eliminate disease-related lesions in a realistic manner. In a user evaluation, domain experts rated the counterfactuals generated using this approach as significantly more realistic than those produced by the classifier-only method and found them indistinguishable from real images. We conjecture that such realistic counterfactual explanations hold significant promise for assisting clinicians in decision-making processes.

To summarize, BagNets provide saliency map based explanations by highlighting image regions that have a substantial impact on the model’s final decision. In contrast, counterfactuals illustrate the actual visual features that are relevant to the classifier’s decision making process by generating varied versions of the input image corresponding to each class in the task. Overall, both of these methods offer visual explanations pertaining to the model’s decisions albeit through different mechanisms.

# Zusammenfassung

Faltendes Neuronales Netzwerk (Convolutional Neural Networks, CNN) haben sich in der Augenheilkunde als leistungsstarke Werkzeuge erwiesen, da sie bei der Klassifizierung von Augenkrankheiten eine zu klinischen Experten identische Genauigkeit aufzeigen. Die Entscheidungen dieser Modelle sind jedoch undurchsichtig und schwer zu interpretieren, was ihre Vertrauenswürdigkeit und Anwendbarkeit im klinischen Umfeld einschränkt. In dieser Arbeit adressieren wir diese Problematik: durch eine inhärent interpretierbare Modellarchitektur namens BagNet und kontrafaktische Erklärungen. BagNets verwenden kleine rezeptive Felder, um lokale Merkmale in einem Bild zu identifizieren, die wesentlich zur Entscheidung des Modells beitragen. Andererseits zeigen kontrafaktische Erklärungen, welche Änderungen an einem gegebenen Bild erforderlich sind, um die Entscheidung des Klassifizierers zu ändern. Wichtig ist, dass beide Erklärungsmethoden Ähnlichkeiten mit Prozessen aufweisen, die Menschen zur Erklärung ihrer Entscheidungen nutzen..

Erstaunlicherweise zeigen CNNs eine bemerkenswerte Genauigkeit bei der Vorhersage des Geschlechts aus Netzhautfundusbildern, obwohl Augenärzten bisher nicht bekannt war, dass Netzhautfundusbilder Geschlechtinformationen enthalten. In diesem Fall ist es vorteilhaft eine Erklärung zur Entscheidung des CNN-Modells zu liefern. Eine mögliche Erklärung bieten Merkmale, die das Modell zur Unterscheidung zwischen männlichen und weiblichen Fundusbildern verwendet. Zur Identifikation dieser Merkmale nutzen wir BagNets, welche Auffälligkeiten im Fundusbild mittels Salienzkarten hervorheben können. Unsere Analyse zeigte, dass der Sehnervenkopf und die Makula wichtige Merkmalsbereiche waren, wobei der Sehnervenkopf bei Fundusbildern von Männern und die Makula bei Fundusbildern von Frauen einen größeren Einfluss auf das Klassifikationsergebnis hatte. Die BagNet Architekturen bieten somit zusammenfassend eine brauchbare Alternative zu Standard-CNNs, da sie eine fundierte Erklärung für die automatische medizinische Bildanalyse liefern.

Als nächsten Schritt in Richtung erklärbares maschinelles Lernen, untersuchten wir nach der Merkmalsextraktion mit BagNets die Erzeugung kontrafaktischer Bilder mit CNN-Klassifikatoren. Konkret bewerten wir verschiedene Techniken zur Erzeugung kontrafaktischer Bilder für eine Reihe von Klassifikatoren für Netzhauterkrankungen in der Augenheilkunde. Die erste Technik beruht auf der Generierung kontrafaktischer Bilder allein anhand der Gradienten eines Klassifikators in bezüglich der Eingangsbilder. Hier bieten adversarial robuste Modelle besser interpretierbare Gradienten als ein Standardklassifikator, allerdings auf Kosten einer geringeren Genauigkeit. In dieser Arbeit stellen wir eine Lösung vor, die die Stärken beider Ansätze kombiniert, indem wir ein Standard-CNN mit einem adversarial robusten CNN kombinieren. Unsere Ensemble-Methode erreicht hohe Genauigkeiten, die mit denen des Standard-CNN vergleichbar sind, und erzeugt gleichzeitig aussagekräftige visuelle Kontrafakturen. Eine auffallende Einschränkung dieses reinen Klassifikator-Ansatzes ist jedoch die mangelnde Realitätsnähe der erzeugten Kontrafakturen.

In einem nächsten Schritt adressierten wir daher die realistische Erzeugung von Kontrafakturen mittels Diffusionsmodellen. Dafür kombinierten wir Diffusionsmodelle mit Klassifikatoren, die zur Klassifikation von Netzhauterkrankungen auf Farbfundusfotografien und B-Scans der optischen Kohärenz fotografie (OCT) trainiert wurden. Die Gradienten der Klassifikatoren ermöglichtem dem Diffusionsmodell krankheitsbedingte Läsionen auf realistische Weise hinzuzufügen oder zu entfernen. Eine Nutzerbewertung mit klinischen Experten ergab, dass kontrafaktische Bilder die mit diesem Ansatz erzeugt wurden deutlich realistischer sind als Bilder die mit der reinen Klassifikatormethode erzeugt wurden. Dies zeigte sich auch dadurch, dass Experten die Bilder dieser Methode nicht von echten Bildern unterscheiden konnten. Abschließend vermuten wir, dass solche realistischen kontrafaktischen Erklärungen eine vielversprechende Unterstützung für klinische Entschei-

dungsprozesse bieten.

Zusammenfassend lässt sich sagen, dass BagNets auf Salienzkarten basierende Erklärungen liefern, indem sie Bildregionen hervorheben, die einen wesentlichen Einfluss auf die endgültige Entscheidung des Klassifikators haben. Im Gegensatz dazu veranschaulichen Kontrafakturen die tatsächlichen visuellen Merkmale, welche für den Entscheidungsprozess des Klassifikators relevant sind. Auf diese Weise können sie verschiedene Versionen des Eingabebildes erzeugen, die je nach der betrachteten Klasse unterschiedliche charakteristische Merkmale abbilden. Obwohl sich die vorgestellten Methoden in ihren zugrundeliegenden Mechanismen unterscheiden, liefern sie beide wertvolle visuelle Erklärungen für die Modellentscheidungen.

# Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Philipp Berens, for his invaluable guidance, support, and encouragement throughout the journey of my doctoral research. His expertise, insightful feedback, and unwavering commitment have been instrumental in shaping this thesis. Furthermore, I wish to extend my appreciation to him for his kindness and for going the extra mile to establish a welcoming and comfortable work environment for an international research team.

I am deeply grateful to my co-supervisor and TAC advisor, Prof. Dr. Matthias Hein for his guidance, mentorship, and valuable inputs throughout this research endeavor. I am also extremely grateful to Prof. Dr. Georg Martius for being a part of my TAC committee and offering constructive feedback.

I had the great pleasure of working with Murat Seçkin Ayhan, Lisa Koch, Dmitry Kobak, Valentyn Boreiko, Kerol Djoumessi, Ziwei Huang, Sarah Müller and all members of the Hertie Institute for AI in Brain Health. I would like to extend my heartfelt thanks for their camaraderie and support. Their active engagement in numerous fruitful discussions and collaborative teamwork with me has enriched my research experience. Special thanks are due to the ophthalmologists Dr. Hanna Faber, Dr. Focke Ziemssen and Dr. Laura Kühlewein for their time, invaluable contributions and feedback to my research analysis. Their participation has undoubtedly enhanced the significance of this research work.

I would like to thank Leila Masri and Sara Sorce for the motivational discussions and creating a lively atmosphere at the IMPRS-IS doctoral program.

Many thanks to Valeska Botzenhardt for facilitating a smooth onboarding process, assisting with settling down in Tübingen, and providing invaluable administrative support. I also wish to express my gratitude to Chiu Yi Lam and Sibylle Kleine for their help and support with administrative tasks.

The completion of this dissertation would not have been possible without my husband Nishanth whose patience, encouragement and belief in me have been a constant source of strength and motivation. He has shown unwavering love, support, and understanding throughout this challenging journey. I am deeply thankful to my parents for their unconditional love, encouragement, and sacrifices throughout the four years of my academic journey. Their constant support and belief in my abilities have been my guiding light. I would also like to express my heartfelt gratitude to my in-laws for their kindness, encouragement, and support.

Finally, I wish to thank and appreciate all the individuals who have contributed in any way to this thesis, whether through discussions, feedback, or moral support. Your contributions have been invaluable and have enriched the quality of this work.

*“Ennenpa ēnai eluttenpa ivvirantum  
kañnenpa vālum uyirkku”*

*Letters and numbers,  
equivalently literature and science,  
are to a human being,  
as are the two eyes to the living.*

Thiruvalluvar, Thirukkural

# Contents

<b>1</b>	<b>Introduction</b>	<b>10</b>
1.1	A brief history of Deep Neural Networks . . . . .	10
1.2	Deep Neural Networks in Ophthalmology . . . . .	11
1.3	Human and DNN explanations . . . . .	13
1.4	Outline . . . . .	17
<b>2</b>	<b>Theoretical and mathematical background</b>	<b>18</b>
2.1	Supervised classification with CNNs . . . . .	18
2.2	Training and loss functions . . . . .	19
2.3	Testing and performance measures . . . . .	20
2.4	CNN architectures . . . . .	21
2.4.1	Inceptionv3 . . . . .	21
2.4.2	ResNet . . . . .	22
<b>3</b>	<b>Explaining CNN decisions</b>	<b>23</b>
3.1	BagNet based explanations . . . . .	23
3.2	Counterfactual Explanations . . . . .	24
3.2.1	Plain and adversarially robust classifiers . . . . .	24
3.2.2	Sparse Visual Counterfactuals . . . . .	25
3.2.3	Diffusion Models . . . . .	25
3.2.4	Diffusion Visual Counterfactuals . . . . .	26
<b>4</b>	<b>Interpretable gender classification from retinal fundus images using BagNets</b>	<b>27</b>
4.1	Abstract . . . . .	27
4.2	Introduction . . . . .	27
4.3	Related Work . . . . .	28
4.4	Methods . . . . .	29
4.4.1	Data and preprocessing . . . . .	29
4.4.2	Network architecture and training . . . . .	29
4.4.3	Generation of saliency maps . . . . .	30
4.4.4	Embedding of image patches . . . . .	30
4.5	Results . . . . .	31
4.6	Discussion . . . . .	32
<b>5</b>	<b>Visual explanations for the detection of diabetic retinopathy from retinal fundus images</b>	<b>34</b>
5.1	Abstract . . . . .	34
5.2	Introduction . . . . .	34
5.3	Methods . . . . .	35
5.3.1	Datasets . . . . .	35
5.3.2	Plain, robust and ensemble models . . . . .	36
5.3.3	Generating visual counterfactual explanations (VCEs) . . . . .	36
5.3.4	Saliency maps . . . . .	37
5.3.5	Model evaluation . . . . .	37
5.4	Results . . . . .	37
5.4.1	Ensembling plain and adversarially trained DNNs . . . . .	37



5.4.2	VCEs as an alternative to saliency maps . . . . .	39
5.4.3	Sparsity versus Realism of VCEs . . . . .	39
5.4.4	VCEs for different budgets . . . . .	39
5.5	Discussion . . . . .	40
<b>6</b>	<b>Generating Realistic Counterfactuals for Retinal Fundus and OCT Images using Diffusion Models</b>	<b>41</b>
6.1	Abstract . . . . .	41
6.2	Introduction . . . . .	41
6.3	Methods . . . . .	42
6.3.1	Datasets . . . . .	42
6.3.2	Generating realistic counterfactual retinal images . . . . .	43
6.3.3	Diffusion Models . . . . .	44
6.3.4	Plain and adversarially robust classifiers . . . . .	44
6.3.5	Diffusion Visual Counterfactuals . . . . .	44
6.3.6	Prior work: Sparse Visual Counterfactuals . . . . .	45
6.3.7	User study . . . . .	46
6.4	Results . . . . .	46
6.4.1	Fundus diffusion counterfactuals are realistic . . . . .	47
6.4.2	Realistic counterfactual examples require robust classifiers . . . . .	49
6.4.3	Influence of regularisation strength on diffusion counterfactuals . . . . .	50
6.4.4	Diffusion counterfactuals for the multiclass DR grading task . . . . .	51
6.4.5	Diffusion counterfactuals of OCT scans are also realistic . . . . .	52
6.5	Discussion . . . . .	54
<b>7</b>	<b>Conclusion</b>	<b>57</b>
7.1	Summary and contributions . . . . .	57
7.2	Future work . . . . .	57
<b>A</b>	<b>Supplementary materials to Chapter 5</b>	<b>59</b>
<b>B</b>	<b>Supplementary materials to Chapter 6</b>	<b>62</b>

# List of Figures

1.1	McCulloch-Pitt’s neuron model . . . . .	11
1.2	Fukushima’s neocognitron model . . . . .	12
1.3	Imaging modalities in ophthalmology . . . . .	13
1.4	Overview of explanations from BagNets . . . . .	14
1.5	Overview of counterfactual explanations . . . . .	15
4.1	A sketch of gender prediction using BagNet-33 . . . . .	28
4.2	Saliency maps from BagNets . . . . .	30
4.3	t-SNE visualization of image patches and associated class evidence . . . . .	31
4.4	Kernel density estimates of factors affecting gender prediction . . . . .	32
5.1	Visual counterfactual and saliency map obtained from plain, robust and ensemble models . . . . .	35
5.2	Counterfactual Explanations with varying degree of sparsity . . . . .	38
5.3	Counterfactual explanations with varying radii . . . . .	39
6.1	Overview of method for generating diffusion visual counterfactuals . . . . .	45
6.2	Diffusion visual counterfactuals and sparse visual counterfactuals for binary DR classification task . . . . .	47
6.3	Results of user study to evaluate realism of fundus diffusion counterfactuals . . . . .	48
6.4	Comparison of diffusion counterfactuals generated with plain model, robust model and cone projection . . . . .	49
6.5	Effect of regularization strength on generated diffusion counterfactuals . . . . .	50
6.6	Diffusion counterfactuals for 5-class DR grading task . . . . .	52
6.7	Diffusion counterfactuals of OCT images . . . . .	53
6.8	Results of user study to evaluate realism of OCT diffusion counterfactuals . . . . .	53
A.1	Sparse visual counterfactual failure case: visible artifacts with $\ell_2$ - and $\ell_{1.5}$ norms . . . . .	59
A.2	Sparse visual counterfactual failure case: visible artifacts with $\ell_{1.5}$ norm . . . . .	60
A.3	Visual counterfactual explanation for correctly classified and misclassified healthy fundus image . . . . .	60
A.4	Visual counterfactual explanations for misclassified DR and healthy fundus images . . . . .	61
B.1	Web interface for user study to evaluate realism of counterfactuals . . . . .	62
B.2	More examples comparing sparse and diffusion counterfactuals . . . . .	63
B.3	As in Fig. 6.4 for a DR fundus image . . . . .	64
B.4	More examples of DR diffusion counterfactuals from healthy fundus images using plain model, robust model and cone projection . . . . .	65
B.5	Effect of regularization strength on diffusion counterfactuals of fundus images with severe or proliferative DR . . . . .	66
B.6	Effect of regularization strength on diffusion counterfactuals from healthy and DR fundus images . . . . .	67

# List of Tables

4.1	Gender prediction performances of DNNs . . . . .	29
5.1	DR classification performances of plain, robust and ensemble models . . . . .	37
5.2	IoU scores of saliency maps and thresholded visual counterfactual saliency maps .	38
6.1	Summary of retinal datasets used for model development and evaluation . . . . .	43
6.2	Classification and grading performances of plain and robust classifiers . . . . .	46
6.3	Statistical assessment of factors for realism of fundus counterfactuals . . . . .	48
6.4	Quantitative assessment for choice of regularization parameter . . . . .	51
6.5	Statistical assessment of factors for realism of OCT counterfactuals . . . . .	54

# Chapter 1

## Introduction

### 1.1 A brief history of Deep Neural Networks

Deep Neural Networks (DNNs) are the game-changers of this century in the field of artificial and machine intelligence. They are capable of performing a wide variety of tasks ranging from playing games [1, 2], detecting and recognizing objects [3, 4, 5], and processing natural language [6, 7]. The foundations of DNNs were laid a century earlier stemming from inspirations of modelling the biological brain [8]. The fundamental unit of a biological brain is a neuron, which is a highly complex cell due its various types, physiological properties and morphologies. The most simplified representation of a neuron [8] consists of a cell body or soma, several branches protruding from the soma called the dendrites and a long fiber called the axon which extends from a region called the axon hillock in the soma. The axon further branches into the axonal arborization and the tips of these branches are the axon terminals which are in contact with other neurons or effector cells. The dendrites and the soma form the input surface of a neuron. Within a neuron, if the various incoming signals from other cells or neurons at the dendrites and the soma generate a potential difference that exceeds a certain threshold then the electrical change yields a spike or action potential which propagates along the axon and is subsequently transferred to the neighboring neurons or cells. This highly simplified behavior of a single neuron is computationally modelled as the McCulloch-Pitts neuron [9] which has multiple binary inputs and a single binary output that is high when the inputs jointly exceed a certain threshold and is low otherwise (Fig. 1.1). This model was capable of simulating logical gates such as AND, OR or NOT. Rosenblatt's perceptron [10] model extended the McCulloch-Pitts single neuron model to multiple neurons which included a self-learning procedure to estimate the parameters of the model based on error-correction. However, this model was shown to fail at learning complex gates such as the XOR gate due to the absence of any hidden units [11]. This drawback was addressed with the introduction of non-linear activation functions in multi-layer perceptron [12] model and the backpropagation algorithm [12] for efficient learning of parameters. Eventually with growing processing power, the number of layers stacked on an MLP increased forming deeper networks with millions of parameters. It is important to note however that these models do not exactly replicate the behaviour of biological neurons. In fact, a single neuron's behaviour can only be encoded with 5 – 8 layered networks of perceptrons [13] which shows that biological neurons are far more computationally complex than perceptrons.

In a similar vein, computer vision models drew inspirations from the visual cortex in biological vision systems [14]. Specifically, the cat's visual system which consisted of two types of cells called the "simple" cells and the "complex" cells [15] formed the basis of computational vision models. The simple cells showed a preference to certain patterns such as vertical lines, edges or corners. On the other hand, the complex cells showed more spatial invariance and appeared to combine responses from different regions. The functionalities of these different cells were modelled as different types of layers in the Neocognitron [16] model which formed the precursor to the modern convolutional neural network (CNN) architectures [14]. The first layer is the input layer which resembles the raw input registered on the retina. Following the input layer are several blocks with each consisting of two different types of layers: the first one which act like simple cells or S-cells are called the "feature detection" layers and the second one called the "pooling" layer simulates the complex

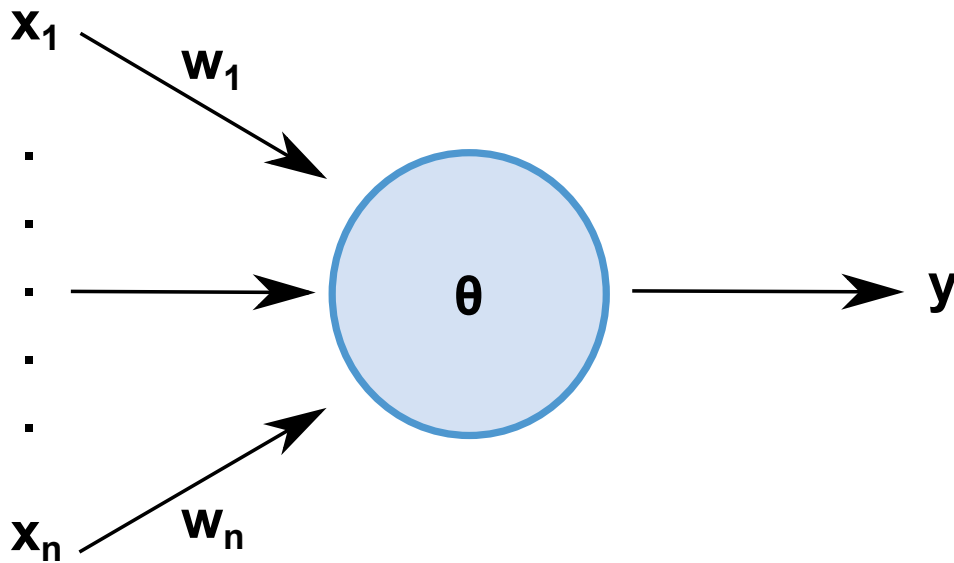


Figure 1.1: Artificial Neuron or the McCulloch-Pitt’s neuron which receives a set of inputs and performs a weighted average of the inputs  $w_1x_1 + w_2x_2 + \dots + w_nx_n$ . When this value is greater than a certain threshold  $\theta$ , the output  $y = 1$ , otherwise  $y = 0$ . This model roughly simulates the behaviour of a biological neuron. Sketch of this artificial neuron is inspired from Figure 3a in The Handbook of Brain Theory and Neural Networks [8]

cells or C-cells which account for spatial invariance (Fig. 1.2). Each of the feature detection layers are modelled mathematically as multiplication with a matrix of weights and the pooling layers correspond to taking an average of the input patches. At the final layer of the model, only certain cells “fire” depending on the input pattern and the average of these output cells determined the final model decision. The Neocognitron model did not have a learning algorithm whereas modern CNNs used the backpropagation algorithm for learning the weights in the different layers [17]. Similar to MLPs, CNNs became deeper with time which drastically improved their performances on detection tasks such as handwritten digit recognition [17, 18] and natural image recognition [4, 19]. In this work, we focus on medical computer vision tasks and all deep networks used are Convolutional Neural Network models. Hence we will use the terms Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) interchangeably.

Following the initial architectural advancements of these networks, the algorithms to optimize the error or loss functions were further fine-tuned and improved. Several purely technical improvements which had no biological connections such as ReLU non-linearities, dropout, batch normalization, data augmentation were also made to boost the performance of these models [21]. Hence, the artificial neural network models eventually drifted apart from the biological neural networks.

## 1.2 Deep Neural Networks in Ophthalmology

Soon after Deep Neural Networks were shown to be successful in vision tasks, they were also applied to medical images to study their effectiveness in clinical diagnostics [22]. Among the various medical fields, ophthalmology was quick to adopt deep neural networks for image analysis due to two factors. First, the imaging modalities used in ophthalmology were predominantly digitized by the time of introduction of deep learning models and computer aided detection using image processing algorithms was already prevalent in the field [23]. Second, population screening programs for various eye diseases stored retinal images and integrated them to patient electronic health records through picture archiving and communication systems (PACS) [24]. These well-established workflows in ophthalmology provided deep learning researchers with convenient access to extensive medical records and image data and enabled them to train models that demonstrated impressive levels of accuracy.

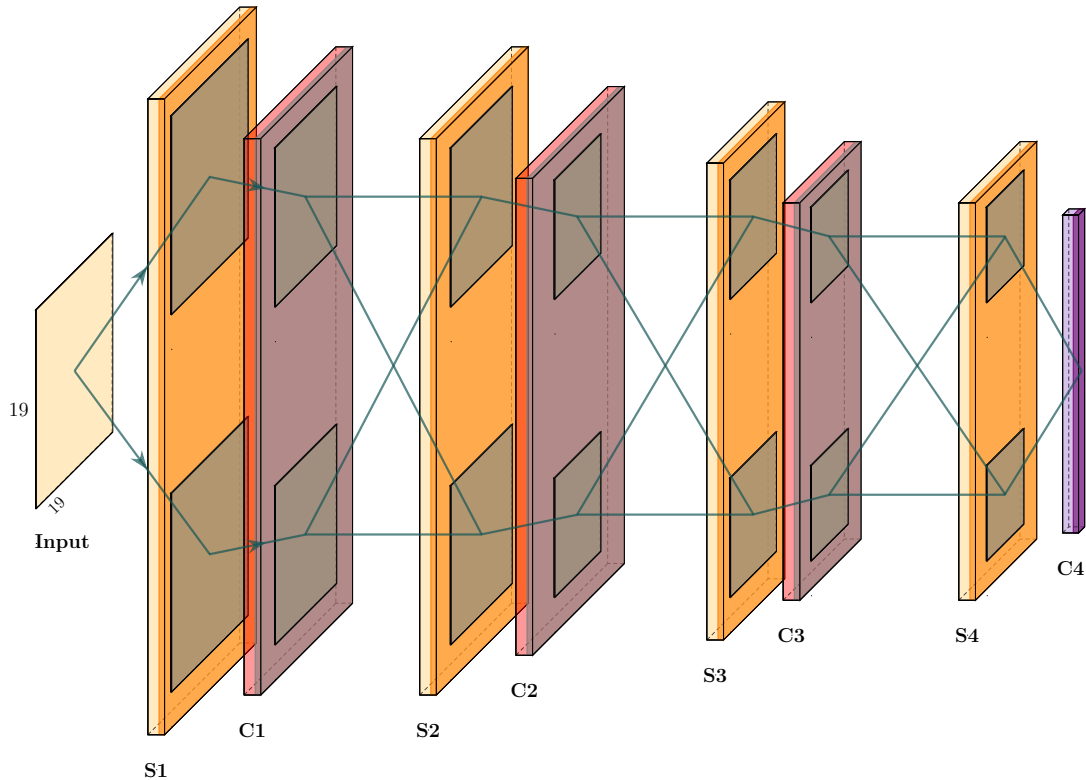


Figure 1.2: Fukushima’s Neocognitron model based on the cat’s visual cortex. The model has a hierarchical structure with alternating S1 cells and C1 cells. While S1 cells are receptive to local features such as edges or corner, C1 cells ensure spatial invariance. The S1 layers are equivalent to convolutional layers and C1 layers to pooling layers in modern CNNs. At the final C layer, each output unit corresponds to one class of the classification task. Architecture figure is reproduced from Fig 2 in [16] and plotted with PlotNeuralNet[20]

Ophthalmologists commonly use two imaging modalities for diagnosis, namely, color fundus photographs and optical coherence tomography (OCT) scans. A color fundus photograph is captured using an ophthalmoscope which records light reflected from the retina on a 2D plane. A typical fundus photograph of a healthy individual shows the optic disc, macula and blood vessels (Fig. 1.3). OCT scans are acquired based on the principle of low-coherence tomography [25]. A high-bandwidth light beam is split into two parts with one part being directed to a mirror and the other to the retina. The light being reflected from the target tissue in the retina is allowed to combine with the beam reflected from the mirror resulting in interference patterns. These patterns are recorded to construct an axial A-scan. Several A-scans are put together to construct a two-dimensional cross-sectional image of the retina called the B-scan. The same is repeated for different depths through various methods such as time-domain OCT or spectral domain OCT which results in a 3D OCT scan [25]. In this work, we use 2D OCT B-scans which show the different layers on a retina such as: inner retina, outer retina, retinal pigment epithelium (RPE)/Bruch’s membrane and the choroid from top to bottom [26] (Fig. 1.3). Both fundus images and OCT scans are most commonly used for diagnosing conditions such as Diabetic Retinopathy (DR), Age Macular Degeneration (AMD), Diabetic Macular Edema (DME) and Epiretinal Membrane (ERM).

Deep learning methods have been used to perform a wide variety of tasks using both of these imaging modalities. With color fundus photographs, deep neural networks have been developed to segment the retinal vessels [27] and optic disc [28] from the fundus image. Among diagnostic tasks, deep learning models are trained to detect diabetic retinopathy [29, 30], Age Macular Degeneration [30, 31] and glaucoma [30] from retinal fundus images. Notably, deep learning models have also shown the capability to perform tasks which are hard for clinicians such as detecting gender, age and other cardiovascular risk factors from retinal fundus images [32]. Miscellaneously, they have also been used to assess the quality of retinal fundus images [33]. With OCT scans, they have been

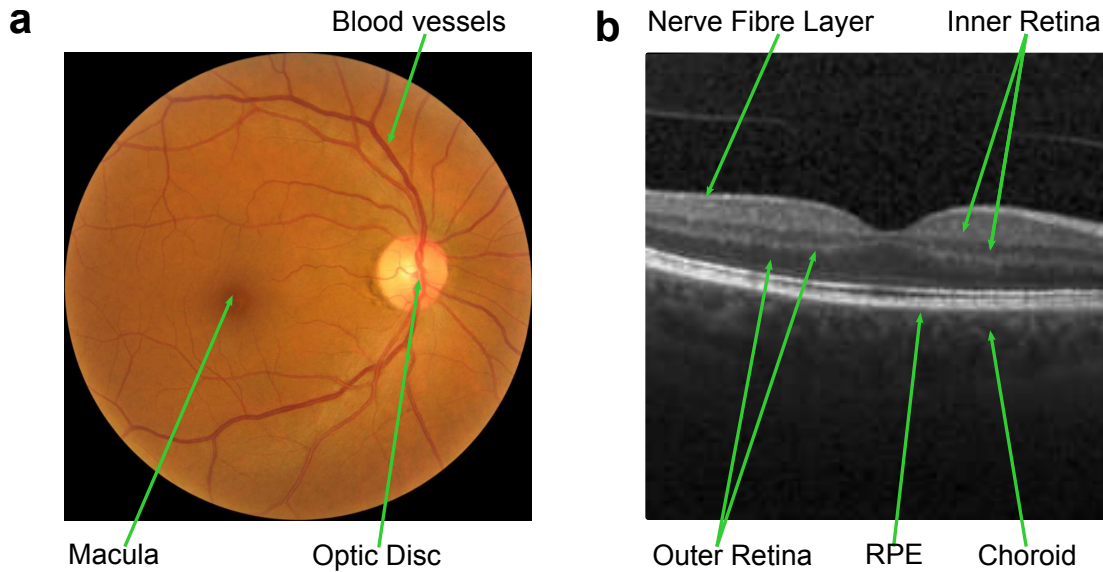


Figure 1.3: Imaging modalities in ophthalmology **a.** An example of a color fundus photograph from a healthy subject showing the macula, optic disc and blood vessels. **b.** An example OCT B-scan of a healthy subject showing the various retinal layers including the nerve fibre layer, outer retina, inner retina, Retinal Pigment Epithelium (RPE) or the Bruch’s membrane and the choroid.

used to segment the various retinal layers and predict the thickness of the retina. Diagnostic tasks performed by DNNs on OCT scans include detection of AMD [34], segmentation of intraretinal fluids [35, 36], drusen, Choroidal Neovascularization (CNV) [37], Diabetic Macular Edema [37] and Epiretinal Membrane [38]. In most of these tasks, DNNs are shown to either perform on par with or better than ophthalmologists and optometrists [29, 34].

One such DNN model which showed outstanding performance at detecting referable Diabetic Retinopathy from retinal fundus images was embedded in a device called the iDX-DR and used to conduct clinical trials [22, 39]. This model passed the performance criteria during clinical trials, leading to its FDA approval for use in the clinics. Recently, it has been shown that devices to detect Diabetic Retinopathy from retinal fundus images using DNNs are among the most widely used AI technologies in a real-world clinical setting [40].

### 1.3 Human and DNN explanations

The example of iDX-DR shows that DNNs will become increasingly popular for decision making in a clinical setting and will be allowed to determine treatment paths. Any wrong decisions by the DNN models could result in high costs. Despite their ability to perform comparably with clinicians, DNNs used in practice are black-box in nature owing to their complex architectures. By default, these models cannot provide additional information on the cause of the decision to either a clinician/technician operating the device or a patient who is interested in learning more about the diagnosis. In order for the different stakeholders involved to gain more insights into the model’s decision, it would be beneficial to provide “explanations” along with decisions. Explanations help users to understand model behaviour (Fig. 1.4), diagnose model failures, identify biases that could have possibly been picked up by the model from training data, enhance the trust of users on deep learning models [42] and facilitate novel scientific discoveries [43]. Furthermore, the regulations established by the European Union concerning the ethical use of Artificial Intelligence place a significant emphasis on ensuring transparency of deep learning algorithms in high-risk scenarios [44, 45].

An “explanation” is a common day-to-day phenomenon in human behaviour. According to Halpern and Pearl, “the role of explanation is to provide the information needed to establish causation”

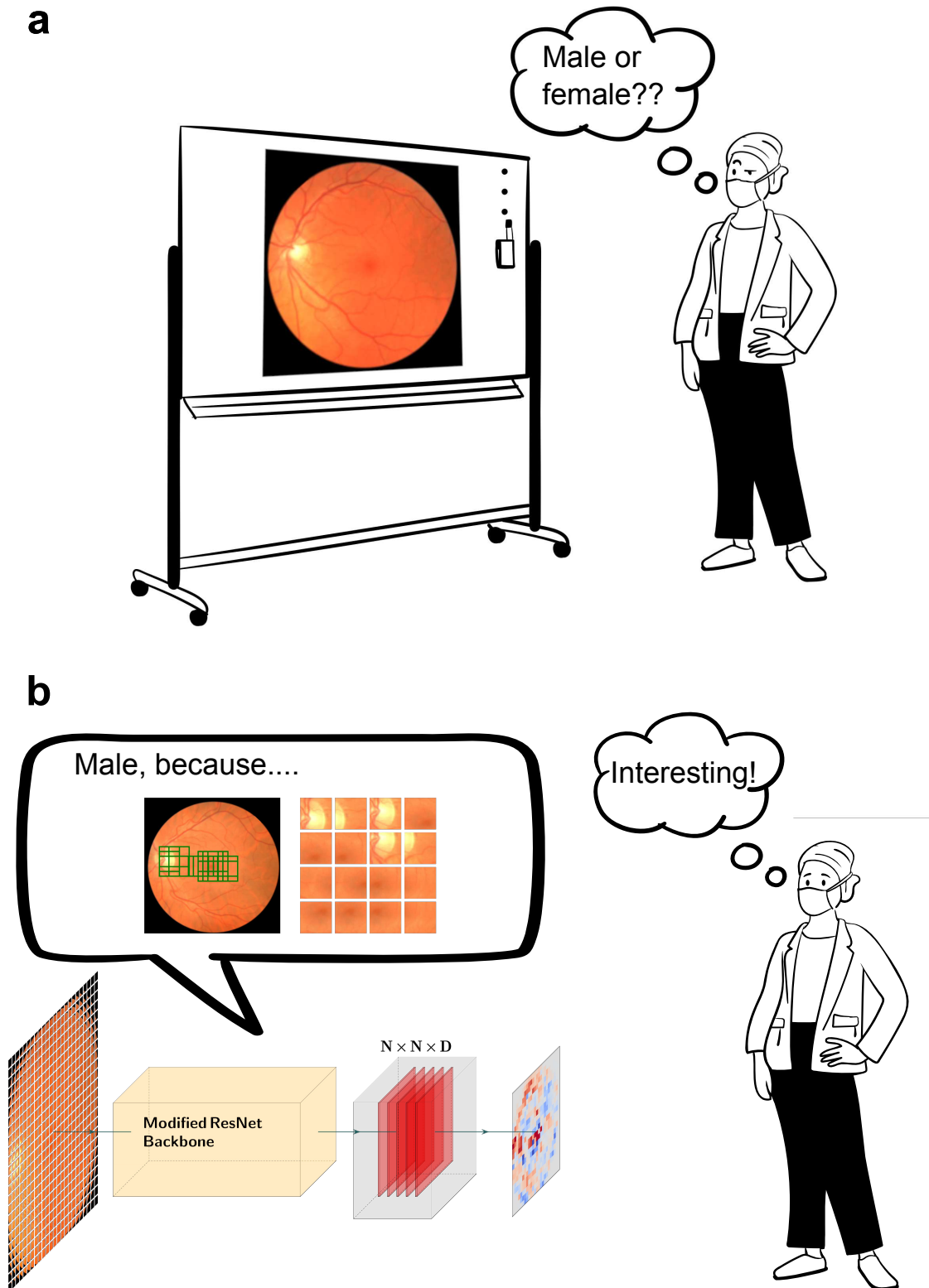


Figure 1.4: Explanations for understanding model behaviour. **a.** For clinicians, it is hardly possible to identify the gender of the subject from their retinal fundus images. **b.** CNN models, on the other hand, can predict gender from fundus images with high accuracy. A BagNet [41] model can provide explanations by highlighting the patches from the image which led to this decision. This is similar to humans providing explanations by specifying the attributes that define a particular class.



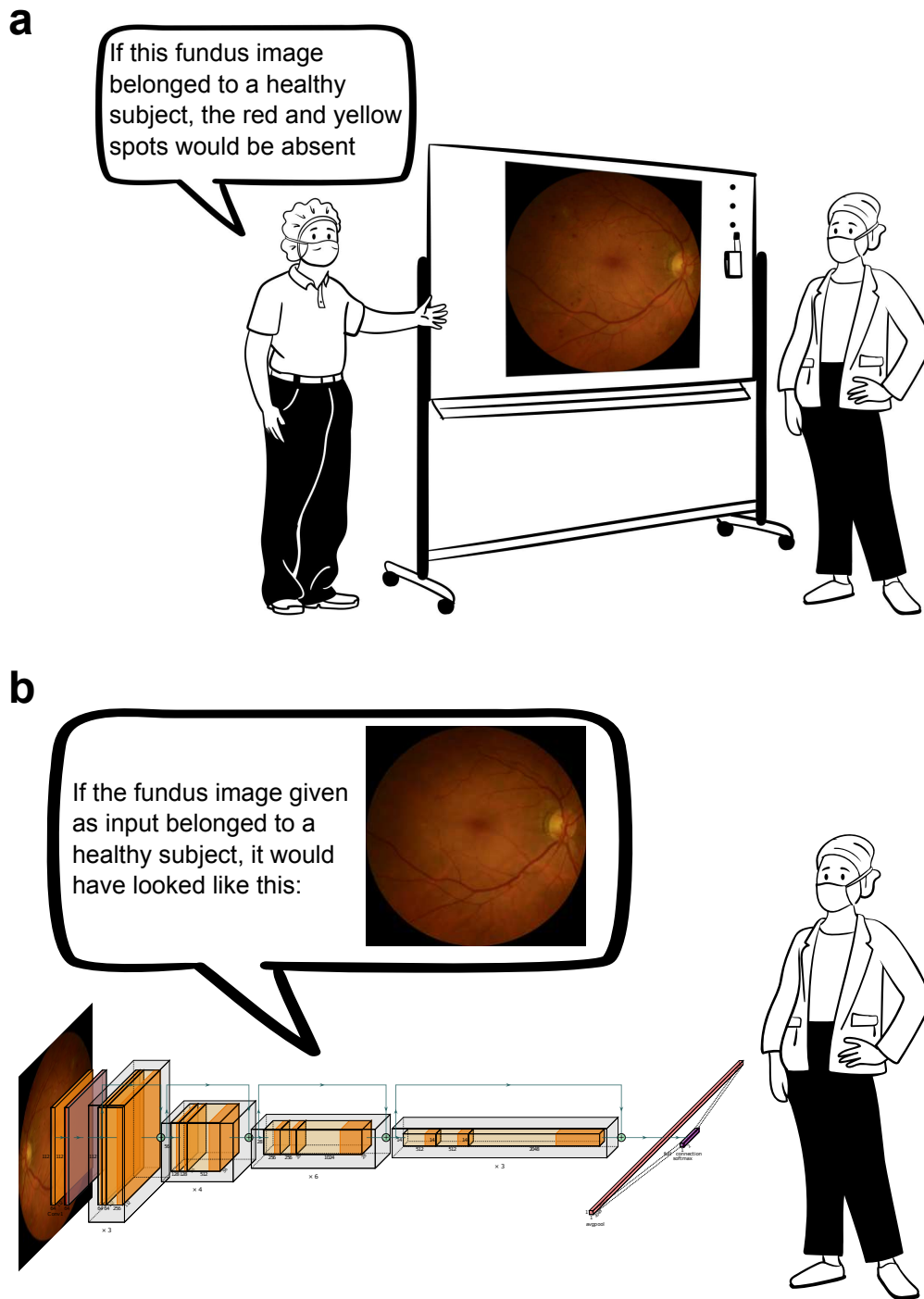


Figure 1.5: Counterfactual reasoning by humans vs counterfactual explanations by a CNN model. **a.** A clinician explains to another clinician why the color fundus photograph belongs to a subject diagnosed with Diabetic Retinopathy by providing a counterfactual reasoning. In the reasoning, he describes the features of Diabetic Retinopathy that appear on the fundus image. **b.** Explanation provided by a CNN model. The model also receives as input the same image as in a. and produces a visual counterfactual explanation. It shows how the image would have looked if it belonged to a healthy subject. To achieve this, the model removes the features relevant to Diabetic Retinopathy on the input image.

[46]. Humans explain their actions and decisions on a regular basis in order to gain common understanding and meaning, to resolve any contradictions or inconsistencies and to influence other's actions and beliefs [47]. The action or decision of interest is the explanandum, the person providing the explanation is the explainer and the one who receives it is the explainee. Typically human explanation involves two stages: the first stage occurs cognitively where the explainer uses his/her mental model of the world and chooses the most relevant set of causes that led to the event. When humans identify an object to a particular category, they think of the intrinsic properties of the object as the cause which led to their decision [47, 48]. For instance, a doctor might explain a patient's condition by indicating the presence of lesions on the patient's medical scan. Here, the lesions constitute the properties of the disease. More often, however, humans derive the causes for a decision or event through contrastive or counterfactual thoughts [47, 49, 50]. People tend to explain the cause of an event by imagining an alternative event that did not occur. For example, in a clinical setting, a doctor might explain why this patient was diagnosed with a certain condition by mentioning how his/her data would have been different if he/she were healthy (Fig. 1.5). The second stage in the process of explaining is a social process where the explainer communicates all of the generated information from the cognitive process to the explainee [47]. Typically, the social process occurs through natural language conversations and involves several exchanges of dialogues between explainer and explainee till they arrive at a common understanding.

These qualities of human explanations provide a framework for explaining decisions of DNN models. We investigated two explainability methods that are closely connected to human explanation processes for DNN models applied to various classification tasks in ophthalmology. In one method, the DNN models highlight regions that played an important role in the prediction of a class (Fig. 1.4). This is roughly equivalent to the specification of the inherent properties of the class by humans as the DNN typically localizes the regions where it found the properties corresponding to the class. The second method is based on generation of counterfactual examples which change the model's decision to one of the contrastive classes (Fig. 1.5). This method is in principle similar to counterfactual reasoning by humans. Our methods are focused on Convolutional Neural Network models specifically trained for visual tasks, devoid of any natural language elements. Consequently, the outcome of the explanation process is exclusively visual and presented in the form of saliency maps, image patches, or counterfactual explanations. Hence, in the context of DNNs the social process unfolds through visual cues.

Independent of human behaviour, DNN explanations can also be classified based on their internal mechanisms. One class of methods called "inherently interpretable" models introduce modifications to the architecture such that the final decision of the network is more interpretable. The explanations from these models are in the form of saliency maps [41, 51], prototypes [52], concepts [53] or local patches of the images which maximally contributed to the final decision of the network [41, 53]. Another class of methods retain the highly complex model architectures and generate explanations in a post-hoc manner. The explanations generated from these methods could be saliency maps [54, 55] or counterfactual explanations [56, 57]. In this work, we investigate an inherently interpretable model called BagNet and a post-hoc explanation method that generates counterfactual explanations.

## 1.4 Outline

The rest of the Thesis is organized as follows. In Chapter 2 we will introduce the mathematical and technical aspects of Convolutional Neural Networks, describe the various loss functions used for their training and the metrics used for their evaluation. Here, we also discuss the different CNN architectures used throughout this work. We follow this up with a description of the methods used to achieve explainability from the Convolutional Neural Network models in Chapter 3. In this chapter, we describe the methods and working principles of an inherently interpretable model architecture called BagNets [41]. Then we define counterfactual explanations in the context of CNN decisions. We present two methods used to generate counterfactual explanations: one which relies solely on adversarially robust classifiers [58] and the other which uses both robust classifiers and generative diffusion models [59]. Following this, we delve into applications in ophthalmology. In Chapter 4, we present explanations using BagNets for gender classification from retinal fundus images. In Chapters 5 and 6, we focus on counterfactual explanations for diagnostic tasks in ophthalmology. In Chapter 5, we investigate a method for generating counterfactual explanations using adversarially robust classifiers [56] for the task of detecting diabetic retinopathy from retinal fundus images. The visual quality of the counterfactual explanations generated in Chapter 5 can be improved by using diffusion models as shown in [57]. We investigate the effectiveness of this method in generating realistic counterfactuals for medical tasks such as detecting diabetic retinopathy from retinal fundus images and detecting various retinal disorders from OCT scans in Chapter 6. Finally, in Chapter 7, we conclude this thesis and discuss future directions of this research work.

## Chapter 2

# Theoretical and mathematical background

Since this work focuses on generating explanations for various medical classification tasks in ophthalmology, we will first provide a short description of how CNNs are generally used for image classification tasks. Here, we will briefly describe the structure of CNN models, the significant mathematical operations that drive the network, the loss functions and the evaluation metrics for testing the model. Finally, we describe the core features of two CNNs architectures that we used in this work namely the Inception-v3 model and the ResNet model.

### 2.1 Supervised classification with CNNs

We will consider  $K$ -class supervised classification tasks with data sets consisting of images  $x \in \mathbb{R}^d$  and associated labels  $\hat{y} \in \{1, \dots, K\}$ . The CNN models learns an approximation of the mapping between the images and their labels. To achieve this, the CNN first maps the images to a real-valued  $K$ -length vector through a parameterized function  $f_\phi(x) : \mathbb{R}^d \rightarrow \mathbb{R}^K$  where  $\phi$  is the set of parameters that are estimated during the training phase. The values in this vector are commonly referred to as “logits”. The final classification output  $y$  of the CNN model is the class which is assigned the highest logit value:

$$\hat{y}_i = \arg \max_{c \in \{1, \dots, K\}} f_\phi(x_i)_c \quad (2.1)$$

The input image passes through several layers in a CNN model and undergoes various transformations before it is mapped into the logit vector. The function  $f_\phi$  is in fact a chain function of the functions applied across the different layers of the CNN model [21]. Let  $f_l$  denote the function at layer  $l$ , then if the CNN has  $N$  layers then  $f_\phi(x) = f_N(f_{N-1}(\dots f_2(f_1(f_0(x))))))$ . Depending on the type of the layer, the function applied to the input varies. The two main types of layers in a CNN are the convolutional layers and the pooling layers. The convolutional layer applies a convolution to its two-dimensional input  $I$ :

$$F(i, j) = \sum_m \sum_n I(m, n) Q(i - m, j - n) \quad (2.2)$$

where  $Q$  is the convolution kernel and the values in the kernel belong to the set of learnable parameters  $\phi$ . In practice, a single convolutional layer consists of  $M$  convolutional kernels and hence, the output of the layer has  $M$  different channels corresponding to each kernel. A convolutional layer is typically followed by a non-linear activation function such as the Rectified Linear Units (ReLU) [60] which is a piecewise linear function:

$$\text{ReLU}(F(m, n)) = \max\{0, F(m, n)\} \quad (2.3)$$

This is followed by a pooling layer which combines the values in a particular local neighborhood by a summary statistic. One type of pooling is the max pooling layer which reduces the size of the feature maps by replacing the values of a  $c \times c$  local patch by the maximum value in this region. For example, a  $2 \times 2$  max pooling layer halves the width and height of the input feature channels.

Typically, these three operations of convolutions, activations and pooling is repeated several times before the last layer. At the penultimate layer, the two dimensional feature maps are flattened to obtain a single high-dimensional vector  $h$  with size  $H$  ranging from 512 to 2048.

This vector  $h$  is then passed through a fully-connected layer which results in the desired  $K$ -dimensional vector of logits:

$$f_\phi(x) = W^T h + b \quad (2.4)$$

where  $W \in \mathbb{R}^{H \times K}$  is a weight matrix and  $b \in \mathbb{R}^K$  is the bias vector both of which are parameters of the fully connected layer. The logits are then converted to class probabilities by applying different activation functions for the binary and multi-class cases. The sigmoid function [21] is used when the classification task is binary and is given by:

$$p_\phi(y_i = 1|x_i) = \sigma(f_\phi(x_i)) \quad (2.5)$$

$$= \frac{1}{1 + \exp(-f_\phi(x_i))} \quad (2.6)$$

On the other hand, the softmax function [21] is used for multi-class classification tasks:

$$p_\phi(y_i = c|x_i) = \text{softmax}(f_\phi(x_i))_c \quad (2.7)$$

$$= \frac{\exp(f_\phi(x_i)_c)}{\sum_{j=1}^K \exp(f_\phi(x_i)_j)} \quad (2.8)$$

## 2.2 Training and loss functions

During training, the parameters of the CNN model are fit such that it learns the mapping between the images and their true labels provided in the data set. In order to obtain the best set of parameters from all possible model configurations possible, a cost function or a loss function which minimizes the error rates for the fitting task is chosen and optimized. Often, the chosen loss function is not the same as the performance measure that is used to evaluate the model on the test set but a function that is expected to improve the performance measure. The overall CNN loss function is non-convex in nature due to the non-linearities which makes it difficult to estimate the global minima. Hence, one usually finds a local minima with a very low value of the loss function.

To achieve this, the Stochastic Gradient Descent (SGD) [21] algorithm is used which updates the weights by moving in the direction of gradients of the loss function with respect to the model parameters. The updates occur in a step-by-step fashion with each step performing computations on a mini batch of examples across the entire dataset. The rate of convergence can be tuned using the learning rate hyperparameter. It is also possible to vary the learning rates with a schedule across different epochs during training. Besides this, other update methods based on momentum [21] and Nesterov momentum [61] could also be used to speed up convergence and obtain better convergence guarantees.

CNNs like most modern deep learning models are trained using maximum likelihood principle [21]. This means that the loss function is simply the negative log-likelihood, equivalently described as the cross-entropy between the training data and the model distribution:

$$J(\phi) = -\mathbb{E}_{x,y \sim p_{data}} \log p_\phi(y|x) \quad (2.9)$$

Therefore, in binary classification tasks where the sigmoid function (Eqn (2.6)) is used to estimate class probabilities, the negative log likelihood function reduces to:

$$J(\phi) = -\log p_\phi(y|x) \quad (2.10)$$

$$= -\log \sigma((2y-1)f_\phi(x)) \quad (2.11)$$

For multiclass tasks, when the softmax function (Eqn (2.8)) is used, the loss function for a single data point with ground truth label  $c$  reduces to:

$$J(\phi)_c = \log \text{softmax}(f_\phi(x))_c \quad (2.12)$$

$$= f_\phi(x)_c - \log \sum_j \exp(f_\phi(x)_j) \quad (2.13)$$

## 2.3 Testing and performance measures

During the testing phase, the model's performance is evaluated on held-out data with examples previously unseen by the model. The performance measure generally depends on the task carried out by the model.

For classification tasks, accuracy [62] is the most commonly used performance measure. It is defined in terms of the true labels  $\hat{y}$  and the predicted labels  $y$  as:

$$\text{accuracy}(\hat{y}, y) = \frac{1}{N} \sum_{i=0}^{N-1} \mathbb{I}(\hat{y}_i = y_i) \quad (2.14)$$

where  $\mathbb{I}$  is the indicator function and  $N$  denotes number of samples in the test set.

For classification in medical tasks where class imbalance in data sets is highly prevalent, other performance measures such as the Receiver Operating Characteristic (ROC) curve and balanced accuracy are also widely used.

Some of these performance measures rely on values in a matrix with rows corresponding to ground truth labels and columns to predicted values. This matrix is called the confusion matrix [63] which we will denote by  $C$ . Each element  $C_{ij}$  of a confusion matrix denotes the number of observations with ground truth label  $i$  and model prediction  $j$ . As an example, for a binary classification task, the entry in  $C_{00}$  is the True Negatives (TN),  $C_{11}$  is True Positives (TP),  $C_{01}$  is False Positives (FP) and  $C_{10}$  is the False Negatives (FN).

A Receiver Operating Characteristic [62, 63] is a two-dimensional curve with the true positive rate on the y-axis and the false positive rate on the x-axis. The different points on the ROC graph are obtained by varying the decision threshold of the model. The Area Under the ROC curve (AUC) [62, 63] is a quantitative measure derived from the ROC curve which is used to compare performances of different models.

For binary classification, balanced accuracy is defined based on entries of the confusion matrix [64]:

$$\text{balanced-accuracy} = \frac{1}{2} \left( \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad (2.15)$$

For multi-class classification with  $k$  classes, the class balanced accuracy is defined as follows:

$$\text{balanced-accuracy} = \frac{1}{k} \sum_i^k \frac{C_{ii}}{N_i} \quad (2.16)$$

where  $N_i$  is the number of samples with ground truth label  $i$  [64].

Another popular metric for evaluating medical grading tasks, especially Diabetic retinopathy grading from retinal fundus images is the Cohen's Kappa metric [65]. The kappa score is intended to compare labels by different annotators and is a number between -1 and 1. Values  $\leq 0$  as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80

as substantial, and 0.81 – 1.00 as almost perfect agreement [66]. Mathematically, Cohen’s Kappa is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (2.17)$$

where  $p_o$  is the actual observed agreement and  $p_e$  is the chance agreement.

The Kappa score has two variants: linearly weighted and quadratic weighted. As the disease grading task is ordinal in nature, the quadratic variant is more suitable than the linear variant for evaluating our models. In order to compute the quadratic weighted variant, first a weight matrix  $W_k$  is defined where each element is given by:

$$W_k(i, j) = \frac{(i - j)^2}{(K - 1)^2} \quad (2.18)$$

Then, a matrix of expected outcomes  $E$  is calculated assuming that there is no correlation between the grades assigned by the clinicians in the data set and grades predicted by the CNN model. This is equivalent to the outer product between the histogram vector of ground truth grades and the histogram vector of predicted grades, normalized such that both  $E$  and the confusion matrix  $C$  have the same sum. From these matrices  $W_k$  and  $E$ , the quadratic weighted Kappa  $\kappa$  can be calculated as:

$$\kappa = 1 - \frac{\sum_{i,j} w_{ij} C_{ij}}{\sum_{i,j} w_{ij} E_{ij}}. \quad (2.19)$$

In this work, we use the accuracy, Area under Receiver Operating Characteristic and the balanced accuracy as performance measures for binary classification tasks. For multi-class classification tasks, we use accuracy, balanced accuracy and the Cohen’s quadratic Kappa score as the performance metric where applicable.

## 2.4 CNN architectures

We used different CNN architectures for the various classification tasks with ophthalmology data. The Inception-v3 [5] architecture served as a baseline model for detecting gender from retinal fundus images owing to the high AUC it achieved on this task [32]. The explainability of this task was studied with a BagNet [41] model which is a variant of the ResNet [4] model. Other classification tasks in this study include detection of referable Diabetic Retinopathy from retinal fundus images, grading Diabetic Retinopathy stages and categorization of OCT scans into the classes: normal, choroidal neovascularization, drusen and diabetic macular edema. For all these tasks, we used the ResNet-50 architecture. Below we describe the main features of the Inception-v3 and ResNet architectures.

### 2.4.1 Inceptionv3

The Inception or the GoogLeNet model [19] from Google won the ImageNet Large Scale Visual Recognition challenge (ILSVRC) [67] in the year 2014. The GoogLeNet architecture introduced an Inception module which significantly reduced the size and computational costs of deep networks. For example, this model had 15 times fewer parameters than the AlexNet model. The inception modules consisted of parallel paths of computation of various filter sizes including  $1 \times 1$ ,  $3 \times 3$  and  $5 \times 5$  filters. It also contained a path with a pooling layer. Furthermore, the module used bottleneck layers of  $1 \times 1$  convolutions in order to reduce the number of channels and subsequently the number of parameters in the layers that followed the bottleneck layers. Additionally, the network eliminated a large number of parameters by using an average pooling instead of fully connected layer at the top. The Inception-v3 model [5] was a follow-up of the original Inception model. This model further improved the computational efficiency by introducing modifications to the original Inception module such as factorizing convolutions into smaller convolutions and factorizing convolutions to asymmetric convolutions.

## 2.4.2 ResNet

In 2015, deep Residual Network, popularly known as ResNet [4], won the ILSVRC [67] challenge. This architecture divided the network into several blocks called the residual blocks  $R$  and introduced skip connections between these blocks. Each residual block is composed on convolutional and pooling layers. The residual block takes an input  $i$  and learns the residue  $R(i)$ . The skip connection concatenates the input with the residue to pass on  $i + R(i)$  as input to the next block. The skip connections have been found to be effective in addressing the drawback of vanishing gradients in several deep architectures. The architecture also relies on a heavy use of batch normalization. Similar to Inception architectures, this network also eliminates the fully connected layers at the top. ResNets are the most commonly used networks and default choice of CNNs in most vision applications.



## Chapter 3

# Explaining CNN decisions

In this work, we investigated the applicability of two different explainability approaches to various classification tasks in ophthalmology using retinal fundus and optical coherence tomography images. One class of methods uses specialized model architectures called BagNets [41] which are adapted to provide explanations in addition to decision probabilities. This method is discussed in detail in Section 3.1. This section is based on the methods used in [68]. Another class of methods rely on the generation of counterfactual explanations [56] that visualize the features that the classifier found to be most important for a given target class by introducing minimal changes on an input image to alter the decision of the classifier to the target class. We will describe two methods to generate counterfactual explanations in Section 3.2. These sections are based on the methods of [69] and [70].

### 3.1 BagNet based explanations

BagNets[41] are a bag-of-features image classification model which view an image as a collection of local patches or visual words. The final decision of the model is then obtained by aggregating the occurrence of patches which contain features that are essential to the classification task. The BagNet model does not require an explicit splitting of images into patches. Instead it modifies the ResNet[4] architecture to restrict the field of view such that the penultimate layer implicitly outputs the feature representation of fixed-size local patches on the image. Concretely, the  $3 \times 3$  filters in some of the convolutional layers are replaced by  $1 \times 1$  filters so that the receptive field size at the final layer is  $q \times q$  where  $q \in \{9, 17, 33\}$ . For example, when only 5 convolutional layers consist of  $3 \times 3$  filters across all blocks in the ResNet architecture, the receptive field at the penultimate layer is  $33 \times 33$ . The penultimate layer is 2048 dimensional, hence with a stride of 8 between  $33 \times 33$  patches there are a total of  $24 \times 24$  feature vectors corresponding to each patch. At the final layer, these features are average pooled to obtain a single 2048 dimensional vector for the complete image which is then passed to a linear dense layer to produce the final classification logits.

Since the operations between the penultimate layer until the final logits in a BagNet architecture are linear, they can be readily swapped without changing final output. When the BagNet is used for explaining decisions, the 2048-dimensional feature vectors of all patches can be individually passed through the dense layer to produce a logits for each patch which determines the weight of this patch in the final classification decision. The weights of all the patches can be aggregated to final classification decision. This swapping of operations allows to generate visualizations in the form of heat maps which highlight the contribution of each local patch of size  $q$  to the final decision. Further analysis of these patches and their logit values through visualization methods such as t-Stochastic Neighborhood Embedding (t-SNE) and density plots can further enable to globally understand the important features for each class over the entire dataset.

## 3.2 Counterfactual Explanations

Visual counterfactual explanations (VCEs) are minimal, realistic and high-confidence changes to an image  $x_0$  by which a classifier’s prediction can be altered to a desired target class [56]. They show what features are important for the classifier to change the decision to a particular class, and hence provide insights into what is learned by the classifier.

In this section we will describe two methods for generating visual counterfactual explanations which rely on using models that are robust to adversarial attacks. The first method solely uses gradients of an adversarially robust classifier with respect to the input image to generate counterfactuals. Since the generative capabilities of a classifier are typically limited and it cannot by itself generate realistic counterfactuals, the second method relies on a diffusion model [71] to achieve realism. In order to generate diffusion counterfactuals, the reverse diffusion process is modified such that classifier gradients contribute to this process and guide the diffusion model towards producing counterfactuals in the desired class [57].

We will first discuss plain and adversarially robust classifiers in Section 3.2.1 and then introduce the Sparse Visual Counterfactual Explanations (SVCEs) in Section 3.2.2 We will briefly discuss about diffusion models in Section 3.2.3 before we introduce Diffusion Visual Counterfactual Explanations (DVCEs) in Section 3.2.4.

### 3.2.1 Plain and adversarially robust classifiers

The changes that are introduced on a input image to generate a counterfactual explanation are based on the gradients of the classifier with respect to the input image or pseudo reconstructions of the input image. A plain classifier does not have gradients that are perceptually aligned with the features of a particular class and could result in counterfactual explanations that look visually similar to the original image when the changes are constrained to be minimal. In contrast, the gradients of adversarially robust models have strong generative properties and are more effective in generating meaningful features for a target class [72, 57] despite being subjected to a constraint for producing minimal changes.

This property of adversarially robust models can be attributed to their training procedures which expose them to adversarial attacks. Consider a  $K$ -class classifier  $f_\phi$  with parameters  $\phi$ , logits  $f_\phi(x) \in \mathbb{R}^K$  and output probabilities  $p_\phi(c|x) \in [0, 1]^K$  where  $x \in \mathbb{R}^d$  is the input to the classifier and  $c \in \{1, \dots, K\}$ . A targeted adversarial attack adds imperceptible perturbations to a starting image  $x_0$  which changes the decision of the classifier from the correct class to a target class  $k$ . More precisely, an  $\ell_p$  targeted adversarial attack for  $f_\phi$  at  $x_0$  produces a sample  $x$ , such that

$$\arg \max_{c \in \{1, \dots, K\}} f_\phi(x)_c = k, \quad x \in [0, 1]^d \cap B_p(x_0, \varepsilon). \quad (3.1)$$

where  $B_p(x_0, \varepsilon) := \{\hat{x} \in \mathbb{R}^d \mid \|x_0 - \hat{x}\|_p \leq \varepsilon\}$  is an  $\ell_p$  ball around the original image  $x_0$  with radius  $\varepsilon$ . One usually maximizes a surrogate loss  $L$  for this:

$$\arg \max_{x \in [0, 1]^d \cap B_p(x_0, \varepsilon)} L(f_\phi(x), k). \quad (3.2)$$

To defend the classifier  $f_\phi$  empirically against such attacks, one can perform adversarial training. A well-known and commonly used algorithm for this is TRADES [58]. Its loss function incorporates a term for the adversarial examples in addition to the standard cross-entropy loss:

$$\frac{1}{n} \sum_{i=1}^n \left[ -\log(p_\phi(y_i|x_i)) + \beta \max_{x \in B_2(x_i, \varepsilon)} D_{KL}(p_\phi(\cdot|x) \parallel p_\phi(\cdot|x_i)) \right], \quad (3.3)$$

where  $D_{KL}$  denotes the Kullback-Leibler divergence and  $\beta$  controls the trade-off between adversarial and plain training schemes. In our experiments, we set  $\beta$  to 6 [73, 58]. This process results in a classifier  $f_\psi$  which is robust to adversarial perturbations. Plain classifiers  $f_\phi$ , on the other hand, are not robust to adversarial attacks and corresponds to training with only the cross-entropy loss i.e.  $\beta$  is 0.

### 3.2.2 Sparse Visual Counterfactuals

Generating Sparse Visual Counterfactuals (SVCs) requires an adversarially robust classifier [56] or at least an ensemble of plain and adversarially robust classifiers [69]. Sparse counterfactuals are computationally similar to adversarial examples but conceptually different from them due to the fact that sparse counterfactuals show meaningful changes that are relevant to the target class instead of the imperceptible noise added to original examples.

For generating sparse counterfactuals, we used the log probability of the target class as a surrogate loss function (Eqn. (3.2)):

$$L(f_\psi(x), y) = -\log p_{f_\psi}(y|x) \quad (3.4)$$

The sparsity and degree of realism of the generated counterfactuals can be controlled by changing the norm used for defining the constrained set  $B_p(x_0, \varepsilon)$ . Depending on the norm, either the adaptive projected gradient descent (APGD) [74] and the Frank-Wolfe [75, 76] based schemes can be used as optimizers. APGD requires projections onto  $\ell_p$ -balls which are available in closed form for  $\ell_2$  and  $\ell_\infty$  or can be computed efficiently for  $\ell_1$  [77]. However, for  $p \notin \{1, 2, \infty\}$ , there is no such projection available and the Auto-Frank-Wolfe (AFW) algorithm [56] was used to solve the optimization and generate sparse counterfactuals.

### 3.2.3 Diffusion Models

Diffusion models are generative image models that yield high-quality and realistic images as a result of two processes [59, 71]: forward diffusion and reverse diffusion. Forward diffusion is a Markov chain that gradually adds Gaussian noise to a starting image  $x_0$ :

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbb{I}), \quad (3.5)$$

where  $t \in \{1, \dots, T\}$ ,  $\beta_t$  denotes a noise schedule such that  $q(x_T|x_0) \approx \mathcal{N}(x_T; 0, \mathbb{I})$ . Given  $x_0$ , the noisy images at any time step  $t$  can be also expressed in closed form:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, \mathbb{I}), \quad (3.6)$$

where  $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$ .

Then, in the reverse diffusion process, the posterior  $q(x_{t-1}|x_t, x_0)$ , when conditioned on  $x_0$ , can be estimated using the Bayes Theorem [78]. The unconditioned posterior  $q(x_{t-1}|x_t)$  is, however, intractable and has to be approximated by a parameterized distribution  $p_\theta(x_{t-1}|x_t)$ :

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (3.7)$$

The mean and diagonal covariance of this distribution are predicted by DNNs denoted by  $\mu_\theta(x_t, t)$  and  $\Sigma_\theta(x_t, t)$ , respectively. Briefly, these models are trained by optimizing a simplified loss function derived from the Variational Lower Bound (VLB) of the negative log likelihood  $-\log p_\theta(x_0)$ . The simplification involves learning the residual noise  $\epsilon_\theta(x_t, t)$  at each time step and then expressing the mean  $\mu_\theta(x_t, t)$  in terms of  $\epsilon_\theta(x_t, t)$ .  $\Sigma_\theta(x_t, t)$  is modeled as an interpolation between  $\beta_t$  and  $\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$  using a vector  $\mathbf{v}$  that is output by the DNN. For further details about the loss functions and training, see [71]. Using  $\mu_\theta(x_t, t)$  and  $\Sigma_\theta(x_t, t)$ , one can generate images from the data distribution  $p(x)$  by starting with a sample from the standard normal distribution and iteratively reconstructing less noisy images at previous time steps from the current noisy image at time step  $t$ .

The sampling procedure from Eqn. (3.7) results in unconditional samples from  $p(x)$  whereas counterfactuals must belong to a specified target class, thus requiring conditional sampling from  $p(x|y)$ . Classifiers can be used to drive diffusion models towards producing realistic images that belong to a desired class [71, 79, 80]. More specifically, the gradients of a classifier with respect to the image shift the mean of the reverse transitions (Eqn. (3.7)) to guide the diffusion model in the right direction.

### 3.2.4 Diffusion Visual Counterfactuals

Here, we describe how to produce realistic Diffusion Visual Counterfactuals (DVCs). Following [57], we combined an unconditionally trained diffusion model  $p_\theta$  as described in Section 3.2.3 with an independently trained classifier  $f_\phi$  (see Section 3.2.1) so that the diffusion model can generate class-conditional samples.

In general, classifiers used in conjunction with diffusion models for conditional sampling are noise-aware, i.e., they are trained on noisy images which occur at various time steps in the diffusion process [71]. Consequently, the input to these classifiers includes the time step  $t \in \{1, \dots, T\}$  at which the image  $x_t \in \mathbb{R}^d$  occurred. Since our classifiers are not noise-aware or time-dependent, that is to say the inputs are only the images  $x \in \mathbb{R}^d$ , we use an estimation of  $x_0$  from a given noisy sample  $x_t$  as input to the classifier. We denote this estimated  $x_0$  by  $x_{0,dn}(x_t, t)$  and following Eqn. (3.6) it can be expressed as:

$$x_{0,dn}(x_t, t) \rightarrow \frac{x_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}, \quad (3.8)$$

where  $\epsilon_\theta(x_t, t)$  can be calculated as a function of the mean  $\mu_\theta(x_t, t)$  [71]. The reverse process transitions of a diffusion model guided by an external classifier which is not noise-aware are given by:

$$p_{\theta,\phi}(x_{t-1}|x_t, y) = Z p_\theta(x_{t-1}|x_t) p_\phi(y|x_{0,dn}(x_t, t)) \quad (3.9)$$

where  $Z$  is a normalization constant. Exact sampling from this distribution is intractable, however, it can be approximated by a Gaussian distribution in a way similar to the unconditional reverse transitions (Eqn. (3.7)) but with shifted mean [71][57].

$$p_{\theta,\phi}(x_{t-1}|x_t, y) = \mathcal{N}(\mu_t, \Sigma_\theta(x_t, t)), \quad (3.10)$$

$$\mu_t = \mu_\theta(x_t, t) + \Sigma_\theta(x_t, t) \nabla_{x_t} \log p_\phi(y|x_{0,dn}(x_t, t)) \quad (3.11)$$

The shift in the mean depends on the gradients of the external classifier which guide the diffusion model to generate images in a specified target class. This, however, does not ensure that the generated image will stay close to the original image  $x_0$  in pixel space, which is one of the qualifying factors for realistic visual counterfactuals. Therefore, to obtain a counterfactual from  $p(x|y)$  that remains structurally close to the original image,  $x_0$ , we find it beneficial to add a distance regularization term to Eqn. (3.11). To maintain consistent parameters  $\lambda_c, \lambda_d$  across different images, an adaptive parameterization, as discussed in [57], is important. This adaptation changes the mean of the transition kernel to:

$$\mu_t = \mu_\theta(x_t, t) + \Sigma_\theta(x_t, t) \|\mu_\theta(x_t, t)\|_2 \Gamma_{\text{DVC}}, \quad (3.12)$$

$$\Gamma_{\text{DVC}} = \lambda_c \frac{\nabla_{x_t} \log p_\phi(y|x_{0,dn}(x_t, t))}{\|\nabla_{x_t} \log p_\phi(y|x_{0,dn}(x_t, t))\|_2} - \lambda_d \frac{\nabla_{x_t} d(x_0, x_{0,dn}(x_t, t))}{\|\nabla_{x_t} d(x_0, x_{0,dn}(x_t, t))\|_2} \quad (3.13)$$

As a further measure to avoid generating images that deviate too much from the original, we start the reverse of the diffusion process from the noisy image at step  $\frac{T}{2}$  instead of the completely distorted version of the image at the last step  $T$  [57].

In Eqn. (3.13), the plain model  $f_\phi$  can also be replaced by the adversarially robust model  $f_\psi$ . While adversarially robust models have stronger generative properties, they suffer from a considerable drop in accuracy compared to plain models. Hence, it would be advantageous to explain a plain model with better performance while also utilizing the stronger gradients of the robust model. To achieve this, we project the gradients of the adversarially robust model  $\nabla_{x_t} \log p_\psi(y|x_{0,dn}(x_t, t))$  onto a cone centered around the gradients of the plain model  $\nabla_{x_t} \log p_\phi(y|x_{0,dn}(x_t, t))$ . This procedure is called cone projection [57] and it is done by changing  $\Gamma_{\text{DVC}}$  in Eqn. (3.13) to

$$\Gamma_{\text{DVC}} = \lambda_c \frac{\Gamma_{\text{cone}}}{\|\Gamma_{\text{cone}}\|_2} - \lambda_d \frac{\nabla_{x_t} d(x_0, x_{0,dn}(x_t, t))}{\|\nabla_{x_t} d(x_0, x_{0,dn}(x_t, t))\|_2}, \quad (3.14)$$

where

$$\Gamma_{\text{cone}} = P_{\text{cone}(\alpha, \nabla_{x_t} \log p_\phi(y|x_{0,dn}(x_t, t))} [\nabla_{x_t} \log p_\psi(y|x_{0,dn}(x_t, t))], \quad (3.15)$$

$\lambda_c$  and  $\lambda_d$  are positive constants and  $\alpha$  is the angle of the cone, which we set to  $30^\circ$  following [57].

## Chapter 4

# Interpretable gender classification from retinal fundus images using BagNets

Author	Author position	Scientific ideas %	Data %	Analysis and interpretation %	Paper writing %
Indu Ilanchezian	1	20	80	60	60
Dmitry Kobak	2	20	0	10	10
Hanna Faber	3	0	10	5	0
Focke Ziemssen	4	0	10	5	0
Philipp Berens	5	30	0	10	10
Murat Seçkin Ayhan	6	30	0	10	20
<b>Publication status:</b>	Published in MICCAI 2021				

### 4.1 Abstract

Deep neural networks (DNNs) are able to predict a person’s gender from retinal fundus images with high accuracy, even though this task is usually considered hardly possible by ophthalmologists. Therefore, it has been an open question which features allow reliable discrimination between male and female fundus images. To study this question, we used a particular DNN architecture called BagNet, which extracts local features from small image patches and then averages the class evidence across all patches. The BagNet performed on par with the more sophisticated Inception-v3 model, showing that the gender information can be read out from local features alone. BagNets also naturally provide saliency maps, which we used to highlight the most informative patches in fundus images. We found that most evidence was provided by patches from the optic disc and the macula, with patches from the optic disc providing mostly male and patches from the macula providing mostly female evidence. Although further research is needed to clarify the exact nature of this evidence, our results suggest that there are localized structural differences in fundus images between genders. Overall, we believe that BagNets may provide a compelling alternative to the standard DNN architectures also in other medical image analysis tasks, as they do not require post-hoc explainability methods.

### 4.2 Introduction

In recent years, deep neural networks (DNNs) have achieved physician-level accuracy in various image-based medical tasks, e.g. in radiology [81], dermatology [82], pathology [83] and ophthalmology [29, 34]. Moreover, in some cases DNNs have been shown to have good performance in tasks that are not straightforward for physicians: for example, they can accurately predict the gender from retinal images [32]. As this task is typically not clinically relevant, ophthalmologists are not

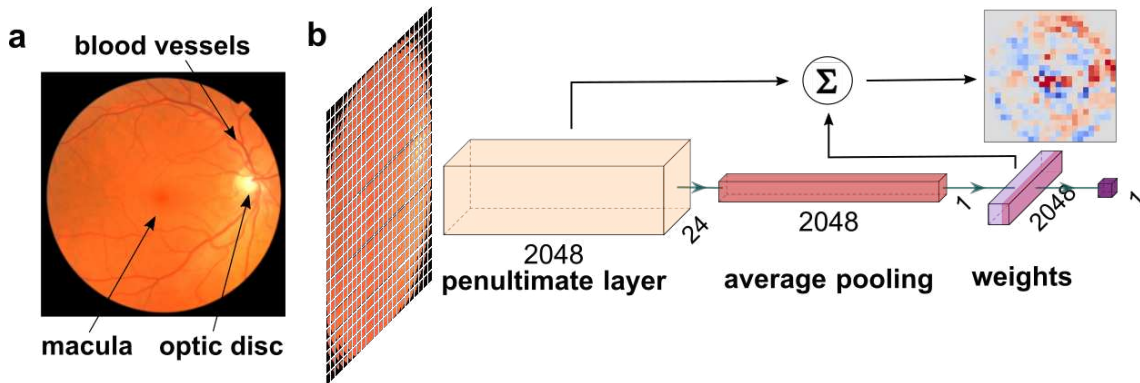


Figure 4.1: A sketch of the gender prediction via BagNet33. (a) Example fundus image from the UK Biobank. The optic disc is the bright spot on the right, the macula is the slightly darker spot in the middle and the blood vessels are extending from the optic disc in darker red. (b) The BagNet33 extracts 2048-dimensional feature vectors from  $33 \times 33$  patches and stores them in the penultimate layer. Via spatial average pooling and a linear classifier, it then forms the final predictions for the gender. The same linear classifier can be applied directly to the feature representation in the penultimate layer to compute the local evidence, which can be visualized as a saliency map. Plotted with PlotNeuralNet [20].

explicitly trained for it. Nevertheless, the comparably poor performance of ophthalmologists at this task suggests that gender differences in fundus images are not obvious or salient. Even though saliency maps used by [32] and follow-up studies [84, 85] have tentatively pointed at the optic disc, the macula, and retinal blood vessels as candidate regions for gender-related anatomical differences in fundus images, conclusive evidence is still lacking. Therefore the high gender prediction performance of DNNs has created lots of interest in the medical imaging community as one hope for DNNs is to unravel biomarkers that are not easily found by humans. Here, we performed a proof of principle study to make progress on the question of how DNNs are able to detect gender differences in retinal fundus. Our contribution is twofold: we (1) introduced BagNets [41] — a ‘local’ variant of the ResNet50 architecture [4] — as an interpretable-by-design architecture for image analysis in ophthalmology and (2) used them to narrow down the hypothesis space for question at hand.

We trained the BagNets on a large collection of retinal fundus images obtained from the UK Biobank [86] (Fig. 4.1a). BagNets use a linear classifier on features extracted from image patches to compute local evidence for each class, which is then averaged over space to form the final prediction, without considering any global relationships. Thus, BagNets resemble ‘bag-of-features’ models popular before deep learning [87]. Despite this simple bag-of-features approach, the BagNet performed on par with an Inception-v3 network in terms of gender prediction accuracy, indicating that gender can be determined from the local characteristics of the fundus image. Also, the BagNet architecture naturally allowed to construct saliency maps to highlight the most informative regions for gender prediction in the retina (Fig. 4.1b). We found that the macula contained most distinctive female patches, while the optic disc contained male ones. In addition, we showed that the decision of the BagNet was not simply caused by some exclusively female or male patches in the images, but rather by a change in both frequency and the degree of ‘femaleness’ or ‘maleness’ of individual patches. Overall, we argue that BagNets can be useful in medical imaging applications including both disease diagnosis and biomarker discovery, thanks to interpretability provided by their local architecture. Our code is available at <https://github.com/berenslab/genderBagNets>.

### 4.3 Related Work

Previous work on gender prediction from fundus images have used either standard DNN architectures or simple logistic regression on top of expert-defined features. For example, [32] trained Inception-v3 networks on the UK Biobank dataset to predict cardiovascular risk factors from fundus images and found that DNNs were also capable of predicting the patient’s gender (AUC = 0.97). A similar network was used by [84]. In both studies, the authors computed post-hoc saliency maps to study the features driving the network’s decisions. In a sample of 100 attention maps, [32] found

Table 4.1: Gender prediction performances of DNNs

	TRAINING		VALIDATION		TEST		CLINICAL	
	ACC.	AUC	ACC.	AUC	ACC.	AUC	ACC.	AUC
InceptionV3	92.99	0.98	83.87	0.92	82.97	0.91	62.07	0.78
BagNet33	93.44	0.98	85.48	0.93	85.26	0.93	72.41	0.70
BagNet17	86.66	0.94	82.30	0.90	82.11	0.90	37.93	0.51
BagNet9	82.41	0.92	79.95	0.89	80.57	0.90	41.38	0.45

that the optic disc, vessels, and other nonspecific parts of the images were frequently highlighted. However, this seems to be the case for almost all the dependent variables and it is very hard to derive testable hypotheses for gender specific differences. Likewise, [84] manually inspected a sample of occlusion maps and concluded that DNNs may use geometrical properties of the blood vessels at the optic disc for predicting gender. More recently, [85] demonstrated that DNNs can predict gender not only from retinal fundus images but also from OCT scans, where the foveal pit region seemed most informative based on gradient-based saliency maps. Taking a different approach, [88] used expert-defined image features in a simple logistic regression model. Although the performance of their model was worse (AUC = 0.78), they found various color-intensity-based metrics and the angle between certain retinal arteries to be significant predictors, but most effect sizes were small.

BagNets provide a compromise between linear classifiers operating on expert-defined features [88] and high-performing DNNs [32, 84, 85], which require complex post-hoc processing for interpretability [89, 90]. In BagNets, a saliency map is also straightforward to compute by design, and it has been shown to provide more information about the location of class evidence than auxiliary interpretability methods [41]. Such native evidence-based maps returned by BagNets are interpretable as is, while standard saliency maps require fine-tuning and post-processing for compelling visualizations [90]. Thanks to these benefits, BagNets have also been used in the context of histopathological microscopy [91].

## 4.4 Methods

### 4.4.1 Data and preprocessing

The UK Biobank [86] offers a large-scale and multi-modal repository of health-related data from the UK. From this, we obtained records of over 84,000 subjects with 174,465 fundus images from both eyes and multiple visits per participant. Male and female subjects constituted 46% and 54% of the data, respectively. As a substantial fraction of the images were not gradable due to image quality issues (artefacts, high contrast, or oversaturation), we used the EyeQual networks [33] to filter out poor images. 47,939 images (47% male, 53% female) passed the quality check by the EyeQual ensemble. We partitioned them into the training, validation and test sets with 75%, 10% and 15% of subjects, respectively, making sure that all images from each subject were allocated to the same set.

Additionally, we obtained 29 fundus images from patients (11 male, 18 female, all older than 47 years) at the University Eye Hospital with permission of the Institutional Ethics Board. We used these additional images as an independent test set. For all images, we applied a circular mask to capture the 95% central area and to remove camera artifacts at the borders.

### 4.4.2 Network architecture and training

We used BagNets [41] (Fig. 4.1b) and standard Inception-v3 [5] network as implemented in Keras [92]. In a BagNet, neurons in the final layer have a receptive field restricted to  $q \times q$  pixels, where we used  $q \in \{9, 17, 33\}$ . The convolutional stack in the network extracts a 2048-dimensional feature vector for each  $q \times q$  image patch. Patches were implicitly defined, with a stride for convolutions of 8 pixels for  $q = 33$ . Therefore local features were extracted for each patch on a  $24 \times 24$  grid (Fig. 4.1b). A linear classifier combined these 2048 features to obtain the local class evidence which was then averaged across all image patches (average pooling layer).

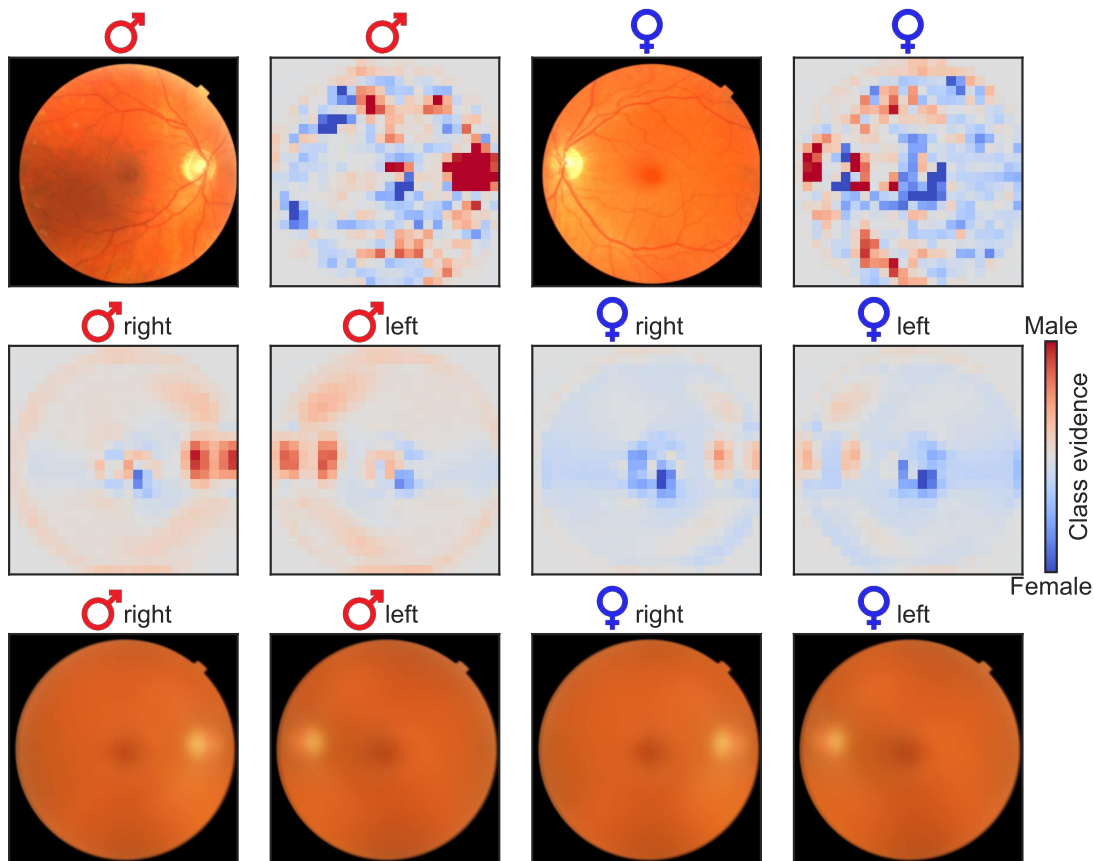


Figure 4.2: Saliency maps obtained by BagNet showing class evidence for each of the image patches on a  $24 \times 24$  grid. Top row shows exemplary test images along with their saliency maps. Middle row shows the average saliency maps for correctly classified male and female patients. Bottom row shows the average fundus images corresponding to the middle row.

All networks had been pretrained on ImageNet [67] by their respective developers. For our binary classification problem, we replaced the 1000-way softmax output layer with a single logistic output neuron (Fig. 4.1b). We initially trained only the output layer using the fundus images for 10 epochs. This was followed by fine-tuning all layers for 100 epochs. We used stochastic gradient descent (SGD) with the learning rate set to 0.01 and the batch size to 16. We used data augmentation via random rotations and flips, width and height shifts, random brightness, and random zooming operations. We picked the best epoch from the [95, 100] range based on the validation performance. We evaluated the final performance on both the test set and the data from the University Eye Hospital.

#### 4.4.3 Generation of saliency maps

To compute saliency maps, we applied the weights  $\mathbf{w}$  in the final classification layer of BagNet33 to the feature vectors, e.g.  $\mathbf{x}$ , in its penultimate layer (Fig. 4.1b), yielding the local evidence (logits) for each patch via  $\mathbf{w} \cdot \mathbf{x} = \sum_i w_i x_i$ . We clipped the resulting values to  $[-75, 75]$  for visualization purposes. The resulting saliency maps were  $24 \times 24$  (Fig. 4.2).

#### 4.4.4 Embedding of image patches

To explore which image patches were informative for classification, we used t-Stochastic Neighborhood Embeddings (t-SNE) [93], a non-linear dimensionality reduction method. To embed the feature representations of  $>1,000,000$  image patches extracted from the fundus images, we used FIt-SNE implementation [94] with uniform affinity kernel in the high-dimensional space across 15 nearest neighbours. We used PCA initialization to better preserve the global structure of the data



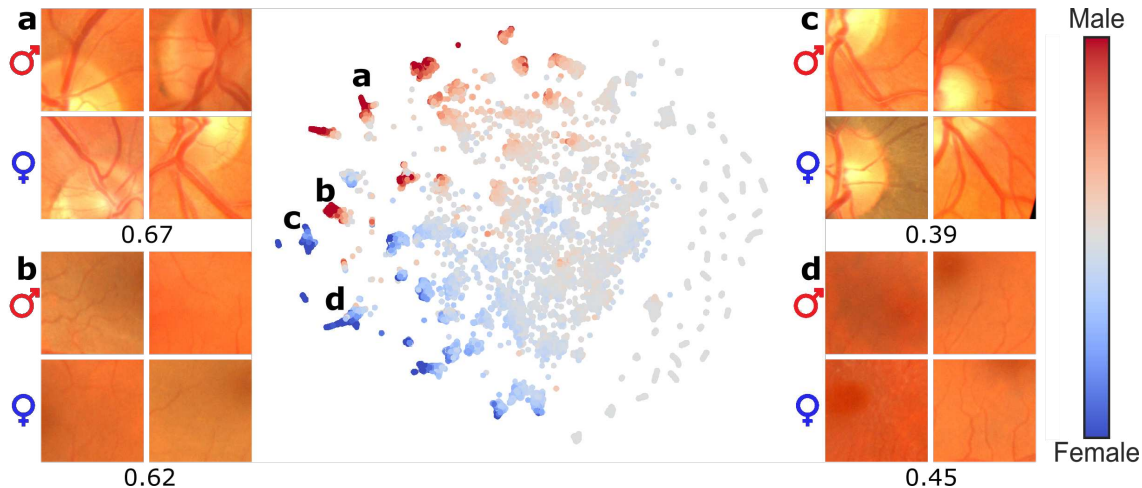


Figure 4.3: Visualization of image patches and associated class evidence via t-SNE. 213,696 patches extracted from 371 correctly classified test images (using training set images yielded a similar embedding; not shown). Four patches with high evidence (two male, two female) are shown from each of the four highlighted clusters. The fraction of male patches in each of these clusters is given below the corresponding exemplary patches. The colors show the logit class evidence. Note that the color does not indicate the correct label of each patch.

and improve the reproducibility [95]. We used a heavy-tailed kernel  $k(d) = 1/(1 + d^2/\alpha)^\alpha$  with  $\alpha = 0.5$  to emphasize cluster structure [96].

## 4.5 Results

We trained BagNets with three different receptive field sizes to predict patient’s gender from retinal fundus images based on the UK Biobank data. We evaluated their performances using prediction accuracy and the Area Under the Receiver Operating Characteristic curve (AUC) and compared to an Inception-v3 network (Table 4.1). BagNet33 and Inception-v3 performed on par with each other, while BagNet17 and BagNet9 performed worse. BagNet33 and Inception-v3 also generalized better to a new clinical dataset, albeit with a substantial drop in performance. Together, this suggests that the  $33 \times 33$  patches captured the relevant information for gender prediction better than smaller patches. Thus, for the remainder of the paper, we will focus our analysis on the BagNet33 (referring to it simply as BagNet).

We inspected saliency maps for gender prediction computed by evaluating the classifier on each feature representation in the penultimate layer (Fig. 4.2, top). In a typical male example, we found that the optic disc provided high evidence for the male class, along with more scattered evidence around the major blood vessels. For a typical female example, high evidence was found for the female class in the macula. Averaging the saliency maps across all correctly classified male/female test images confirmed that the BagNet relied on the optic disc and the blood vessels to identify male images and on the macula to identify female ones (Fig. 4.2, middle).

Interestingly, the individual and the average saliency maps also showed that the optic disc patches tended to always provide male evidence, to some extent even in correctly classified female images. Similarly, the macula patches tended to provide female evidence, even in correctly classified male images. The BagNet could nevertheless achieve high classification performance after averaging the class evidence across all patches.

As a sanity check, we show the averaged fundus images across all correctly classified male/female images in the bottom row of Fig. 4.2. These average images are nearly identical across genders, demonstrating that it is not the location, the size, or the shape of the optic disc or macula that drive the BagNet predictions.

To further explore the structure of local image features informative about gender, we embedded the 2048-dimensional feature representation of each image patch into 2D using t-SNE and colored

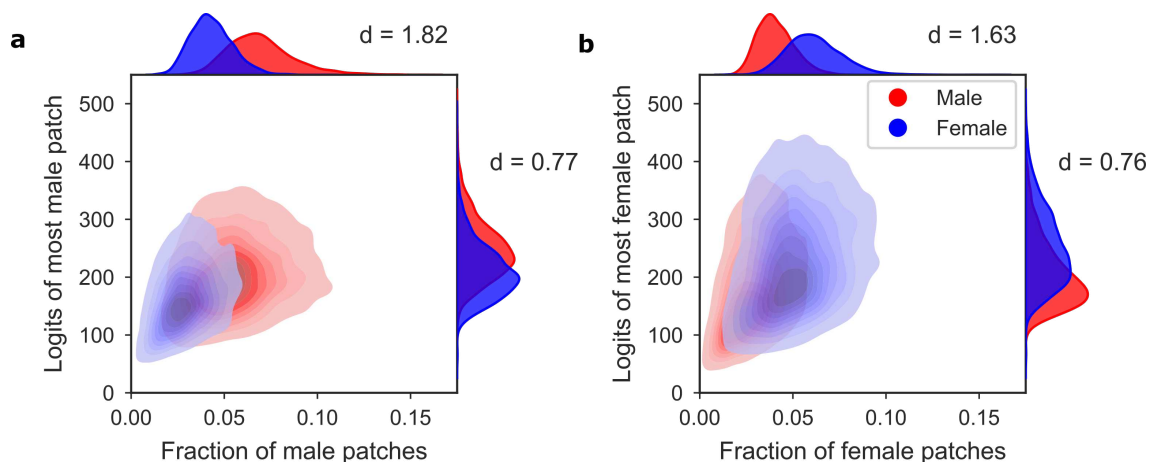


Figure 4.4: Two factors determine the gender predictions of the BagNet: the maximal strength of evidence and the frequency of strong evidence. **(a)** Kernel density estimate of all male (red) and female (blue) test set images. Horizontal axis: fraction of male patches, defined as having logit values above 50. Vertical axis: the absolute logit value of the most male patch. **(b)** The same for patches providing female evidence (logit values below  $-50$ ).

them by the provided class evidence (Fig. 4.3). We found that most image patches provided only weak evidence for either class, but some distinct clusters of patches had consistently high logits. We further explored these clusters and found that they consistently showed the optic disk with blood vessels (**a** and **c**) or the macula (**b** and **d**), in line with the saliency maps computed above (Fig. 4.2). However, even though the clusters **a** and **b** consistently provided evidence for the male class, patches in these clusters occurred in true female and male fundus images alike (67% and 62% patches from male images, respectively). Similarly, clusters **c** and **d** provided evidence for the female class but yet came from male and female fundus images (39% and 45% patches from male images, respectively).

This raised the question of whether the BagNet’s decisions were mostly driven by (i) male/female images having individual patches with stronger male/female evidence; or (ii) male/female images having a larger number of patches with male/female evidence (Fig. 4.4). We found that both factors played a role in determining the final gender predictions, but the fraction of male/female patches seemed to be a stronger factor: Cohen’s  $d = 1.82$  and  $d = 1.63$  for the difference in fraction of male (logit value  $>50$ ) and female (logit value  $< -50$ ) patches between genders, vs.  $d = 0.77$  and  $d = 0.76$  for the difference in the logit value of the most male and the most female patch. Thus, female images contained more patches providing strong female class evidence, and vice versa for male fundus images.

## 4.6 Discussion

In summary, we argued that the BagNet architecture is particularly suitable for medical image analysis, thanks to its built-in interpretability. Here we used BagNets to investigate the high accuracy of DNNs in gender prediction from retinal fundus images. BagNet33 achieved a performance similar to Inception-v3 despite having a much simpler architecture and using only local image features for prediction. This suggested that local features are sufficient for gender prediction and the global arrangement of these features is not essential for this task.

In BagNets, saliency maps can be readily computed without auxiliary gradient-based methods or layer-wise relevance propagation [89]. We used the native saliency maps of BagNets and a two-dimensional t-SNE embedding of image patches to identify the most informative regions for the gender prediction task in fundus images. This allowed us to go beyond the previous reports [32, 84] and for the first time to provide conclusive evidence that the optic disk region contains features used to inform a male prediction and the macula region for a female prediction. We found that both the frequency of informative male/female patches and — albeit to a lesser degree — the strength of the most informative male/female patches were important factors for gender prediction

by BagNets.

It is, however, not the case that the optic disc in males is substantially larger than in females, as can be seen in the average fundus images shown in Fig. 4.2. The relative optic disc and macula sizes, shapes, brightness levels, etc. seem all to be roughly the same for both genders. Instead, our results suggest *structural but localized* differences in the male and female retinas, mainly within the optic disc and macula regions. This is supported by the previous findings showing that the retinal nerve fibre layer in the optic disk is slightly thicker in females [97] and that the macula is slightly thinner [85] and wider [98] in females. However, these previously reported gender differences have small to moderate effect sizes (Cohen’s  $d = 0.11$ ,  $d = 0.52$ , and  $d = 0.17$  respectively for the comparisons referenced above; computed here based on reported means and standard deviations) and it is unclear if they alone can explain the BagNet performance.

Therefore, future work is needed to understand what exactly it is that allows the network to assign high male evidence to the optic disc patches from male patients and high female evidence to the optic disc patches from female patients. In this sense, the results presented here do not provide the final solution to the gender prediction mystery. Nevertheless, we believe that our results make a step in the right direction as they demonstrate structural but localized gender differences and reduce the problem complexity down to specific small patches of the fundus image that can be further analyzed separately.

We believe that BagNets may also be more widely applicable for clinically relevant diagnostic tasks involving medical images in ophthalmology and beyond, provided that they are coupled with reliable uncertainty estimation [99]. In many cases, pathologies often manifest in localized regions, which can be readily picked up by BagNets. For example, BagNets could be used to further explore clinically relevant changes underlying progressive diseases such as diabetic retinopathy. The interpretable architecture of BagNets may increase the trust of clinicians and patients, which is a critical issue for adoption of deep learning algorithms in medical practice [42].

## Acknowledgements

We thank Wieland Brendel for his support with BagNets. This research was supported by the German Ministry of Science and Education (BMBF, 01GQ1601 and 01IS18039A) and the German Science Foundation (BE5601/4-2 and EXC 2064, project number 390727645). Hanna Faber received research funding from the Junior Clinician Scientist Program of the Faculty of Medicine, Eberhard Karls University of Tübingen, Germany (application number 463–0–0). Additional funding was provided by Novartis AG through a research grant. The funding bodies did not have any influence in the study planning and design. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Indu Ilanchezian.

## Chapter 5

# Visual explanations for the detection of diabetic retinopathy from retinal fundus images

Author	Author position	Scientific ideas %	Data %	Analysis & interpretation %	Paper writing %
Valentyn Boreiko*	1	15	35	40	40
Indu Ilanchezian*	2	15	40	30	20
Murat Seçkin Ayhan	3	15	0	5	10
Sarah Müller	4	0	20	0	0
Lisa M. Koch	5	15	0	5	5
Hanna Faber	6	0	5	5	5
Philipp Berens	7	20	0	5	10
Matthias Hein	8	20	0	5	10
<b>Publication status:</b>	Published in MICCAI 2022				

### 5.1 Abstract

In medical image classification tasks like the detection of diabetic retinopathy from retinal fundus images, it is highly desirable to get visual explanations for the decisions of black-box deep neural networks (DNNs). However, gradient-based saliency methods often fail to highlight the diseased image regions reliably. On the other hand, adversarially robust models have more interpretable gradients than plain models but suffer typically from a significant drop in accuracy, which is unacceptable for clinical practice. Here, we show that one can get the best of both worlds by ensembling a plain and an adversarially robust model: maintaining high accuracy but having improved visual explanations. Also, our ensemble produces meaningful visual counterfactuals which are complementary to existing saliency-based techniques. Code is available under [https://github.com/valentyn1boreiko/Fundus\\_VCEs](https://github.com/valentyn1boreiko/Fundus_VCEs).

### 5.2 Introduction

In many medical domains, deep learning systems have been shown to perform close to or even better than domain experts in detecting disease from images [100]. For clinicians and patients to trust such systems in practice, they need to be interpretable [101, 42]. Current techniques for interpreting model decisions, however, have critical shortcomings. For instance, post-hoc interpretability techniques such as saliency maps are often used to generate explanations for a classifier’s decision. These have been evaluated for clinical relevance, e.g. in ophthalmology [102, 103, 104], with some methods producing more meaningful visualizations than others. As DNNs can rely on spurious features and are not necessarily learning all class-relevant features [105, 106], saliency maps may

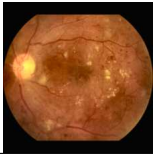


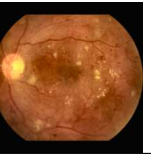


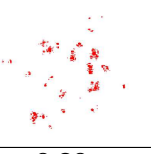
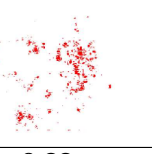

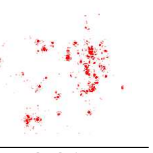

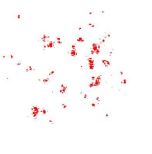



Model	Orig.(GT:DR)	T-GBP	T-IG	$l_{1.5}$ -VCE, $\epsilon=30$ →DR: 1.00	T-VSM
Plain	DR:1.00 	<b>0.10</b> 	<b>0.12</b> 		<b>0.10</b> 
Robust	DR:0.99 	<b>0.19</b> 	<b>0.20</b> 	→DR: 1.00 	<b>0.20</b> 
Ensemble (proposed)	DR:1.00 	<b>0.20</b> 	<b>0.22</b> 	→DR: 1.00 	<b>0.21</b> 

Figure 5.1: Visual explanations of decisions are better for robust and ensemble models than for plain models, as shown by intersection over union (IoU) between saliency maps (P) and ground truth (GT) masks ( $\mathbf{IoU}(P, GT) := \frac{|P \cap GT|}{|P \cup GT|}$ ) (in bold). We show an image correctly classified as DR (left), post-hoc explanations for the decision using thresholded Guided Backprop (T-GBP), Integrated Gradients (T-IG) and visual counterfactual examples (VCEs) for enhancing the classifiers’ confidence into DR as well as the corresponding saliency map: thresholded VCE Saliency Map (T-VSM). Numerical evaluation of these maps in comparison to the ground truth segmentation can be found in Tab. 5.2.

also have limited usefulness in clinical settings [107, 104]: for standard classifiers they sometimes just highlight high-frequency components of an image [103]. Especially for healthy cases, these are often hard to interpret during screening for timely intervention.

Interestingly, models trained to provide inherent robustness against adversarial attacks [108, 74], have also been shown to yield better saliency maps [109, 110]. Also, these robust models allow to generate visual counterfactual explanations (VCEs) [111, 56], an alternative image-wise interpretability technique that shows the minimal changes necessary to maximize the confidence of the classifier in a desired class (Fig. 5.1). But, the gain of these models in adversarial robustness comes at the price of a loss in accuracy [112, 113] which is unacceptable especially in medical applications. Thus, adversarially robust models have not seen widespread use in practice.

Here we show that an ensemble of a plain and an adversarially robust model yields improved saliency maps and allows for the computation of VCEs to further explore the basis of the model’s decision. Further, it achieves almost the same accuracy as the plain model. We demonstrate this new approach to explainability for medical image classifiers for the case of diabetic retinopathy (DR) detection from retinal fundus images and propose a new type of the saliency map.

## 5.3 Methods

### 5.3.1 Datasets

We used three publicly available datasets of retinal fundus images for which DR grades were available: the Kaggle DR detection challenge data [114] for method development and main results, the Messidor dataset [115] for additional external validation, and a portion of the Indian Diabetic Retinopathy Image Dataset (IDRiD) [116] for quantitative evaluation of visual explanations, as these data additionally had DR lesion annotations at pixel level. We pre-processed the images using contrast limited adaptive histogram equalization (CLAHE) [117], and by tightly cropping the circular mask of the retinal fundus, which was detected by iterative least-squares fitting of a

circular shape to image edges. For the Kaggle dataset, we filtered out poor quality images using an ensemble of EfficientNets [118] trained on the ISBI2020 challenge dataset<sup>1</sup>. This quality filtering model achieved 87.50% accuracy for image gradability. After quality filtering, the resulting dataset contained 45,923 images (at a final resolution of  $224 \times 224$  pixels): 33,783 in class ‘no DR’, 3,598 in ‘mild DR’, 6,765 in ‘moderate DR’, 1,186 in ‘severe DR’ and 591 in ‘proliferative DR’. The Messidor dataset contained 1200 retinal fundus images, and the IDRiD 81 images along with annotations for microaneurysms, haemorrhages, hard and soft exudates. We combined the annotations of these lesion types to obtain a single ground truth mask.

### 5.3.2 Plain, robust and ensemble models

As mild DR is a transitional stage between no DR and moderate-to-advanced stages of DR [119], these images lead to high uncertainty in decisions of both DNNs and clinicians [120]. Therefore, to obtain a clear separation of ‘no DR’ and DR classes, we excluded the ‘mild DR’ cases. We then trained binary classifiers  $f_\gamma : \mathbb{R}^d \rightarrow \mathbb{R}^2$  to predict whether a fundus image  $x$  was in the ‘no DR’ class or belonged to moderate-to-advanced stages of DR, with  $p_\gamma(y=1|x)$  indicating the predicted probability of disease. We used 75% of the Kaggle data for training, 15% for validation, 4% for temperature scaling [121] and 6% for testing.

For the plain model  $f_\phi$  we used a ResNet-50 [122] which was trained with cross-entropy loss. We used batch size of 128, with oversampling of the DR cases to account for class imbalance. We first trained the model for 500 epochs with learning rate of 0.01 and a cosine learning rate schedule. This model was further fine-tuned for 3 epochs with a cyclic triangle schedule for one cycle. We chose the model with the best balanced accuracy on the validation set.

The robust model  $f_\psi$  used the same architecture but was trained using TRADES [113] for  $\ell_2$ -adversarial robustness, where one minimizes for the given training set  $(x_i, y_i)_{i=1}^n$  the objective:

$$\frac{1}{n} \sum_{i=1}^n [-\log(p_\psi(y_i|x_i)) + \beta \max_{x \in B_2(x_i, \epsilon)} D_{KL}(p_\psi(\cdot|x) || p_\psi(\cdot|x_i))], \quad (5.1)$$

where  $D_{KL}$  denotes the Kullback-Leibler divergence,  $p_\psi(\cdot|x)$  is the predicted probability distribution over the classes at  $x$ ,  $\beta$  controls the trade-off between adversarial and plain training schemes, and  $B_p(x, \epsilon) := \{\hat{x} \in \mathbb{R}^d | \|x - \hat{x}\|_p \leq \epsilon\}$ . For training we used  $p = 2$  and  $\epsilon = 0.25$  and set  $\beta = 6$ .

In our experience, tuning  $\beta$  down during training can increase accuracy but negatively affects interpretability. Hence, we built the following ensemble of plain and robust models, which preserves both accuracy and interpretable gradients for the given  $\beta$ :

$$p_{\text{ensemble}}(k|x) := \frac{1}{2}[p_\phi(k|x) + p_\psi(k|x)], \quad k = 0, 1. \quad (5.2)$$

As saliency methods often require logits  $f$  instead of probabilities, we defined logits for the ensemble as  $f_k := \log(p_{\text{ensemble}}(k|x))$ . All models are calibrated via temperature scaling by minimizing the expected calibration error [121].

Experiments were done on an Nvidia Tesla V100 GPU with 32GB RAM, using PyTorch. Code for pre-processing and training as well as the trained models will be available upon acceptance.

### 5.3.3 Generating visual counterfactual explanations (VCEs)

Following [56], a VCE  $\tilde{x}$  should have high probability  $p_\gamma(k|\tilde{x})$  in a chosen class  $k$  (“validity”). It should be similar to the starting image  $x_0$  (“sparsity”) and close to the data manifold (“realism”). For generating an  $\ell_p$ -VCE  $\tilde{x}$  for a classifier  $\gamma \in \{\phi, \psi, \text{ensemble}\}$  we solved

$$\tilde{x} = \arg \max_{x \in B_p(x_0, \epsilon) \cap [0, 1]^d \cap \mathcal{M}} \log(p_\gamma(k|x)) \quad (5.3)$$

where  $\mathcal{M}$  is the mask for the region of the eye obtained by our pre-processing. The formulation of VCEs suggests that some “robustness” is required as Eq. 5.3 is similar to the formulation of adversarial examples [56]. Compared to saliency maps the advantage of VCE is that the generated

<sup>1</sup><https://isbi.deeppdr.org/challenge2.html>

Table 5.1: Evaluation of plain and robust classifier and their ensemble in terms of standard, balanced and  $\ell_2$ -robust accuracy. The ensemble maintains the accuracy but gains sufficient robustness required for better interpretability (see Tab. 5.2).

	Kaggle			Messidor		
	acc.	bal. acc.	rob. acc.	acc.	bal. acc.	rob. acc.
Plain	89.5	85.8	15.2	89.5	89.5	20.6
Robust	78.4	71.6	66.6	66.1	66.5	60.9
Ensemble	89.7	85.2	19.4	87.9	87.9	24.4

images are purely based on the behavior of the classifier. We used adaptive projected gradient descent (APGD) [74] and Frank-Wolfe [75, 76] based schemes as optimizers. APGD requires projections onto  $\ell_p$ -balls which are available in closed form for  $\ell_2$  and  $\ell_\infty$  or can be computed efficiently for  $\ell_1$  [77]. However, for  $p \notin \{1, 2, \infty\}$ , there is no such projection available and thus we used for the generation of  $\ell_p$ -VCEs the Auto-Frank-Wolfe scheme of [56].

### 5.3.4 Saliency maps

We used Guided Backprop (GBP) [54] and Integrated Gradients (IG) [55] from a public repository [123] to generate saliency maps for the models’ decisions. GBP and IG are among the best saliency techniques for DR detection [102, 103]. Based on our VCEs, we also introduced the VCE Saliency Map (VSM) as the difference between VCE and the original image. For all saliency methods, we used absolute saliency values summed over color channels in order to better cover salient regions [103]. Then, saliency scores were normalized to  $[0, 1]$  via min-max normalization and thresholded at the  $\tau$ -quantile for sparsity. The threshold  $\tau$  was optimized for each method on 40 out of 81 images in the IDRiD dataset by computing the intersection over union (IoU) with respect to the pixel-wise annotation of DR lesions. This yielded  $\tau = 0.98$  for GBP,  $\tau = 0.96$  for both IG and VSM. For the VSMs we additionally optimized over the norm  $p \in \{1.5, 2, 4\}$  and different  $\epsilon$  per norm and found  $p = 1.5$ ,  $\epsilon = 30$  to be the best.

### 5.3.5 Model evaluation

We evaluated the performance of models on the Kaggle test set and Messidor images using accuracy (acc.), and balanced accuracy (bal. acc., mean of TPR and TNR). Additionally, we reported  $\ell_2$ -robust accuracy (rob. acc.) for a perturbation budget of  $\epsilon = 0.1$  which we evaluated using 9 restarts of 100 iterations of APGD [74] maximizing the confidence in the wrong class. The robust accuracy is the fraction of test inputs where the decision could not be changed by the attack.

For a quantitative evaluation of our visual explanations, we used the 41 images on which  $\tau$  had not been optimized from the IDRiD dataset. Tab. 5.2 shows the mean IoU for all models and saliency techniques (including T-VSMs for different  $p$ -norms) with the pixel-level DR lesion annotations.

This evaluation indicates that the saliency maps derived from VCEs are on par with state-of-the-art techniques, such as GBP and IG. However, VCEs go beyond those techniques as they can be used to generate images and even animations that illustrate how an image would have to change to affect the prediction of the classifier.

## 5.4 Results

First, we analyzed the properties of the plain and robust classifiers, and the ensemble introduced in Eq. 5.2. Then, we explored VCEs as an alternative for explaining classifier decisions and studied the sparsity-realism trade-off for VCEs. Finally, we show the effect of different perturbation budgets on VCEs.

### 5.4.1 Ensembling plain and adversarially trained DNNs

We found that the plain model achieved good standard and balanced accuracy for classifying DR from fundus images (Tab. 5.1), but with comparably low robust accuracy (see Sec. 5.3.5). In

Table 5.2: Evaluation of saliency maps and T-VSMs on IDRiD. The IoU-score of the ensemble is higher than for the plain model for all interpretability methods including VCEs (higher is better, mean  $\pm$  std).

	GBP	IG	$\ell_{1.5}, \epsilon = 30$	$\ell_2, \epsilon = 6$	$\ell_4, \epsilon = 0.2$
Plain	$0.09 \pm 0.03$	$0.08 \pm 0.03$	$0.07 \pm 0.03$	$0.07 \pm 0.03$	$0.07 \pm 0.03$
Robust	$0.15 \pm 0.06$	$0.14 \pm 0.06$	$0.13 \pm 0.06$	$0.12 \pm 0.06$	$0.12 \pm 0.05$
Ensemble	$0.15 \pm 0.06$	$0.14 \pm 0.06$	$0.13 \pm 0.06$	$0.12 \pm 0.06$	$0.12 \pm 0.05$

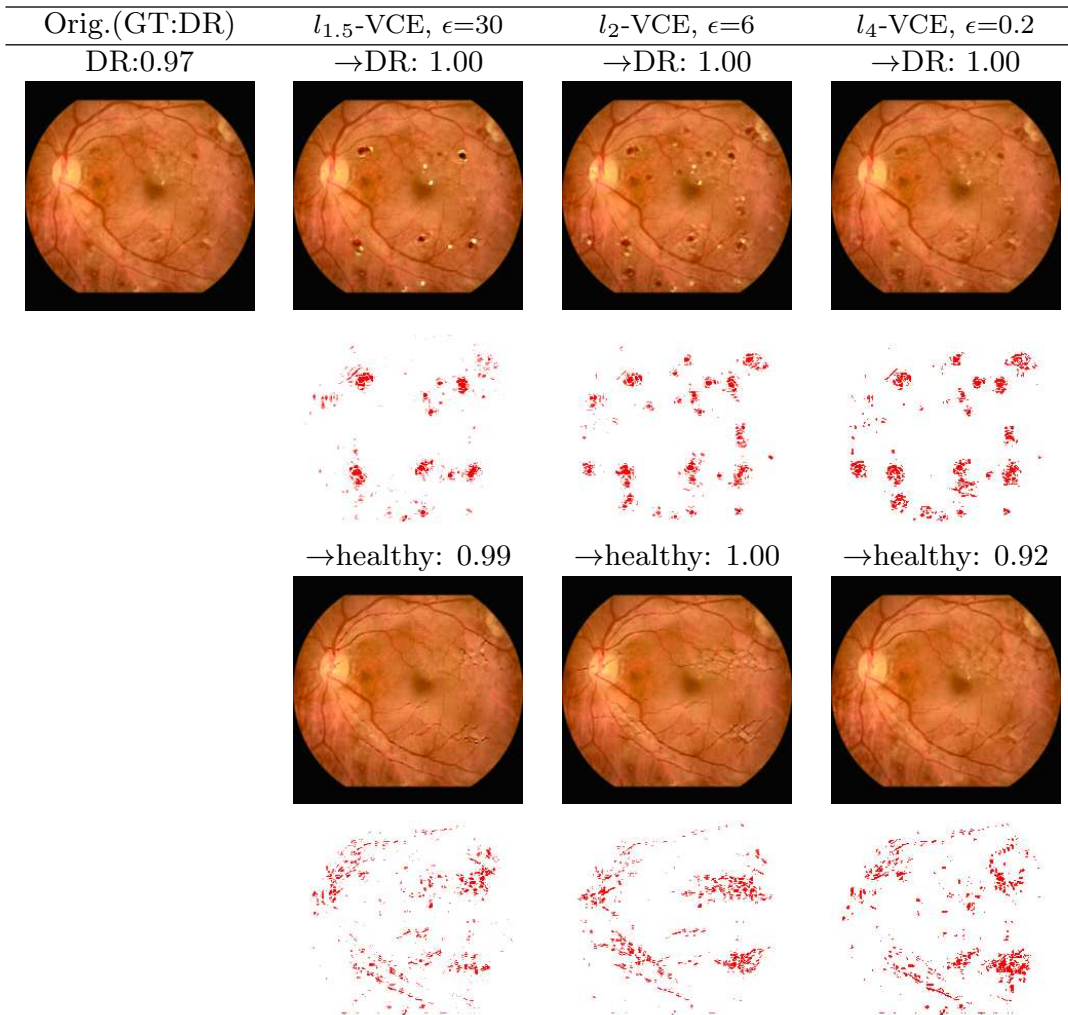


Figure 5.2: VCEs for the ensemble with varying degree of sparsity:  $p \in \{1.5, 2, 4\}$ . For a correctly classified DR image, we show VCEs when transformed further into the DR or the healthy class. Below VCEs, T-VSMs are shown. The VCE radius was adapted to the sparsity condition. In addition, the confidence of the classifier is reported above the image.

contrast, the robust classifier achieved high robust accuracy, but suffered a large drop in accuracy of more than 10-20%. Interestingly, and in line with the literature [109, 110], the saliency maps of the robust model were much better than those of the plain model (Tab. 5.2, Fig. 5.1) for both of the tested saliency methods, Guided Backprop (GBP) and Integrated Gradients (IG). In fact, the saliency maps of the plain classifier were of rather low quality, focusing on less prominent disease-related regions of the image (Fig. 5.1).

We found that an ensemble of the plain and robust models (Eq. 5.2) combined their advantages: It had about equal standard and improved robust accuracy compared to the plain model (Tab. 5.1) and its saliency maps were as good as those of the robust model (Tab. 5.2, Fig. 5.1).












Orig.(GT:DR)	$l_4$ -VCE, $\epsilon=0.1$	$l_4$ -VCE, $\epsilon=0.2$	$l_4$ -VCE, $\epsilon=0.3$	$l_4$ -VCE, $\epsilon=0.4$
DR:0.95	→DR: 1.00	→DR: 1.00	→DR: 1.00	→DR: 1.00
				
	→healthy: 0.74	→healthy: 0.95	→healthy: 0.99	→healthy: 1.00
				

Figure 5.3: VCEs show increasingly strong modification for different radii. For one correctly classified DR image, we show for the ensemble the  $l_4$ -VCEs for  $\epsilon \in \{0.1, 0.2, 0.3, 0.4\}$  when transforming into the DR and healthy class, respectively.

#### 5.4.2 VCEs as an alternative to saliency maps

We next explored VCEs (Eq. 5.3) as an alternative for explaining classifier decisions. The properties of the VCEs depend on the chosen model for the perturbation, which in this paper was always an  $l_p$ -ball, and the perturbation budget in form of the radius of  $l_p$ -ball. Small values of  $p$  always lead to sparse changes whereas for larger  $p$  one can realize much more outspread changes affecting larger parts of the image. As discussed in Sec. 5.3.4 we chose  $l_{1.5}$ -VCEs of radius  $\epsilon = 30$  as they produced the best quality of T-VSMs. We found that the robust model and the ensemble allowed for the computation of realistic VCE (Eq. 5.3, Fig. 5.1). T-VSMs (see Sec. 5.3.4) also provided good explanations for the classifiers’ decision (Tab. 5.2), highlighting exudates and haemorrhages. In contrast, the VCE of the plain model was not very meaningful as its main changes were only vaguely related to the diseased regions.

#### 5.4.3 Sparsity versus Realism of VCEs

We then analyzed the effect of different perturbation models in terms of different  $l_p$ -balls (Fig. 5.2). We first studied the VCEs for enhancing the correct decision for a DR image. We found that the changes of  $l_{1.5}$ -perturbation model were sparser and thus looked more cartoon-like than for  $l_4$ . The VCEs of the  $l_4$  model appeared much more natural although they even introduced new diseased regions not present in the original image. Thus the classifier seems to have picked up certain disease signs very well and can integrate even new disease patterns in a natural fashion into fundus images. We next studied the VCE for changing the decision of the classifier to ‘no DR’. Here, all  $l_p$ -perturbation models attempted to “smooth out” the main lesions as well as the exudates. This provides complementary evidence that the classifier picked up the right disease signal in the data. Note that the artefact around the optic nerve was not changed in the VCE, showing that the classifier has correctly identified it as a feature which is not discriminatory for the disease decision. Not all VCEs, however, provided by our method are perfectly realistic: for example, the algorithm often tried to cover lesions with vessels when creating a VCE turning a diseased image into a healthy one. Further failure cases are shown in Appendix A.1 and A.2.

#### 5.4.4 VCEs for different budgets

Finally, we investigated how the VCEs changed with increasing budget parameterized with  $\epsilon$  (Fig. 5.3). We found that an increasing number of new lesions were introduced for both the sparse  $l_{1.5}$ -VCE as well as the realistic  $l_4$ -VCE, when increasing the budget for more DR evidence. Here, the difference between the two models — that  $l_4$ -VCEs appeared more realistic — became even more clear. When generating VCEs for turning the diseased image into an healthy one, also increasingly large regions of lesions were covered, e.g. through artificial vessels. Such VCE with

different budgets could be useful to generate gradual changes in either directions, providing good intuitions for a classifiers decision.

## 5.5 Discussion

We showed that the ensemble of plain and robust models can preserve accuracy of plain models, yet provide better visual explanations. In agreement with the literature [109, 110], the resulting saliency maps highlight clinically relevant lesions more reliably. Therefore, the explanations obtained for diseased images are often satisfying, while those for healthy images are less so — showing the absence of lesions is difficult in this framework. The ensemble model allowed us to compute also realistic VCEs [56], to yield interpretable explanations of the classifier’s decision, pinpointing the features in the image the classifier picks up on.

In related work, iterative augmentation of saliency maps has been used to improve saliency-based visual explanations [124]. Also, VCEs have been generated using GANs [125] (no models/code is available) but the advantage of our VCE is that they depend only on the classifier and thus there is no danger that the prior of the GAN “hides” undesired behavior of the classifier. Finally, models interpretable-by-design such as BagNets [68] have been advocated for medical imaging tasks [126]. As many high-performing DNNs do not fall into this category, we view our work as complementary.

We believe realistic VCEs and derived T-VSMs will be a useful tool to better understand the behavior of DNN-based classifiers in medical imaging, in particular when gradually morphing an image from one class to the other which is the main complementary strength of VCEs compared to saliency maps. As the sparseness and the degree of changes allowed can be precisely controlled, it is straightforward to yield more or less natural VCEs. Even extreme and therefore less natural VCEs can be useful, as they provide a “cartoon” version of what the classifier believes the disease looks like.

## Acknowledgement

We acknowledge support by the German Ministry of Science and Education (BMBF, 01GQ1601 and 01IS18039A) and the German Science Foundation (BE5601/8-1 and EXC 2064, project number 390727645). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting I.I.

## Chapter 6

# Generating Realistic Counterfactuals for Retinal Fundus and OCT Images using Diffusion Models

Author	Author position	Scientific ideas %	Data %	Analysis & interpretation %	Paper writing %
Indu Ilanchezian*	1	15	40	40	40
Valentyn Boreiko*	2	15	30	20	20
Laura K�uhlewein	3	0	10	10	0
Ziwei Huang	4	0	20	0	0
Murat Seękin Ayhan	5	10	0	5	10
Matthias Hein	6	15	0	5	10
Lisa M. Koch	7	20	0	15	10
Philipp Berens	8	25	0	5	10
<b>Publication status:</b>	Under Review at Medical Image Analysis				

### 6.1 Abstract

Counterfactual reasoning is often used in clinical settings to explain decisions or weigh alternatives. Therefore, for imaging based specialties such as ophthalmology, it would be beneficial to be able to create counterfactual images, illustrating answers to questions like "If the subject had had diabetic retinopathy, how would the fundus image have looked?". Here, we demonstrate that using a diffusion model in combination with an adversarially robust classifier trained on retinal disease classification tasks enables the generation of highly realistic counterfactuals of retinal fundus images and optical coherence tomography (OCT) B-scans. The key to the realism of counterfactuals is that these classifiers encode salient features indicative for each disease class and can steer the diffusion model to depict disease signs or remove disease-related lesions in a realistic way. In a user study, domain experts also found the counterfactuals generated using our method significantly more realistic than counterfactuals generated from a previous method, and even indistinguishable from real images.

### 6.2 Introduction

Humans naturally use counterfactual thoughts, deliberations and statements to reason about the causal structure of the world, understand the past and prepare for the future [49]. For example, counterfactuals are used in medicine to explain decisions or weigh alternatives: "If we had treated the patient with drug X, she might have experienced severe side effects." [127]. In a similar way,

when medical images are available for diagnosis, it might be useful to create counterfactual images that visualize the answer to the question: "For a given subject who we believe is healthy, how would the imaging data have looked for the same subject to be identified as the diseased class?"

In ophthalmology, for instance, clinicians regularly use imaging modalities such as retinal fundus photography and Optical Coherence Tomography (OCT) for diagnosing sight-threatening diseases like Diabetic Retinopathy (DR) and Age-Related Macular Degeneration (AMD). Counterfactual images as described above can be generated from Deep Neural Networks (DNNs) that are trained to detect the presence of these diseases. In this context, counterfactuals are artificially generated images that contain minimal, realistic, meaningful and high-confidence changes to an input image such that the DNN classifier alters its decision to a desired target class [56]. For them to look realistic and meaningful, the models used to create them need to have outstanding generative abilities. The resulting images can then also be viewed as explanations of the DNN's decisions as they enable the user of a DNN model to visualise the features that the classifier relies on for detecting the disease [69, 128].

Previously, different strategies for generating counterfactuals have been proposed [129, 56, 69, 128, 130, 131, 132, 133]. For example, DNN-based counterfactuals can be generated by iteratively superimposing the input image with the gradients of an adversarially robust classifier, which has more informative gradients than a plain model [56, 69, 72]. While these so-called sparse counterfactuals show meaningful features, they appear to modify the original image in unexpected and unnatural ways. On fundus images, they cover lesions with unnatural blood vessels in order to generate healthy counterfactuals [69]. In a similar vein, when StyleGANs are used to generate counterfactual retinal fundus images for a Diabetic Macular Edema classifier [129], the counterfactuals generated with this procedure begin to show features relevant to target class even before the decision of the classifier changes, despite their highly realistic appearances. Counterfactuals of OCT scans have also been generated using GANs to study retinal aging, but domain experts were easily able to identify the generated images [131], suggesting that they are not sufficiently realistic. Finally, in other medical domains such as brain tumor detection from MRI images and chest X-ray interpretation, counterfactuals based on diffusion models have been used to generate healthy counterfactuals from diseased images [132, 133], but not for generating images showing a disease from healthy ones.

Here, we show that we can generate realistic counterfactual generation within the context of retinal disease detection from two ophthalmic imaging modalities, funduscopy and OCT, by relying on deep generative models known as diffusion models [134]. These diffusion models have been shown to outperform GANs in realistic image generation, while also overcoming their drawbacks by producing diverse samples and covering a broad range of the image distribution in tandem with a stable training process [59, 71]. We use classifiers trained to detect several eye diseases from retinal images and then show how to combine these with a generative diffusion model to result in realistic counterfactual retinal images that explain classifiers' decisions in both directions: from healthy to diseased and vice versa. Importantly, we show that domain specialists – ophthalmologists and AI experts – view the resulting images as realistic when probed in an odd-one-out task. This indicates that our methods generates images that fulfill the criteria for counterfactual images to be used in medical reasoning as outlined above.

## 6.3 Methods

We first describe the ophthalmic imaging datasets used in this study and then review the relevant methods for the generation of counterfactuals for such images. Lastly, we describe our design of a user study in order to evaluate the clinical relevance of counterfactuals.

### 6.3.1 Datasets

We used retinal image data sets from two common ophthalmic imaging modalities: (1) color fundus photography (CFP) and (2) Optical Coherence Tomography (OCT).

Fundus images were obtained from EyePacs Inc. through a Diabetic Retinopathy (DR) screening program<sup>1</sup>. Initially, this collection contained over 180,000 retinal fundus images from over 42,000

<sup>1</sup><https://www.eyepacs.com/blog/over-750-000-patients-screened>

Table 6.1: Summary of the retinal image collections used for model development and evaluation.

			Training	Validation	Test	
CFP	EyePacs	subjects	15,827	5,324	6,775	
		images				
			all	46,921	15,658	30,166
			healthy	38,502	12,748	24,627
			mild	3,244	1,163	2,378
			moderate	4,695	1,572	2,907
			severe	238	121	127
			proliferative	242	54	127
		Benitez	images			
			all	789	-	-
			healthy	94	-	-
			mild	6	-	-
			moderate	102	-	-
			severe+	587	-	-
		FGADR	images			
	all		1,842	-	-	
	healthy		101	-	-	
	mild		212	-	-	
	moderate		595	-	-	
		severe+	934	-	-	
OCT	Kermany	subjects	3558	712	474	
		images				
			all	71,231	14,714	10,496
			normal	34,340	6,813	4,464
			CNV	23,133	5,091	3,738
			drusen	5,393	1,221	1,288
		DME	8,365	1,589	1,006	

subjects along with meta data such as age, sex, race and blood pressure. Image quality was indicated as "Insufficient for Full Interpretation", "Adequate", "Good" or "Excellent" per image as annotated by Eyepacs Inc. Some DR labels were missing. We used "Good" and "Excellent" quality images with DR labels only, resulting in 92,745 retinal fundus images from 27,926 participants. Then, we created training, validation and test splits subject-wise (see Table 6.1). The training set was augmented with 789 images from the Benitez data set [135] and 1842 images from the FGADR data set [136] in order to strengthen the representation of diseased samples for the diffusion models. All images were cropped to square dimensions of  $224 \times 224$  pixels using a circle fitting procedure ([https://github.com/berenslab/fundus\\_circle\\_cropping/tree/v0.1.0](https://github.com/berenslab/fundus_circle_cropping/tree/v0.1.0), [137]).

For OCT B-scans, we used a data set consisting of a total of 108,309 images belonging to one of four categories [138]: normal, choroidal neovascularization (CNV), drusen and Diabetic Macular Edema (DME) (Table 6.1). In order to obtain a square center crop including the macular region, we used only the images with size  $496 \times 512$  and  $496 \times 768$  (96,441 scans). We created training, validation and test splits again subject-wise with 75% subjects in training, 15% in validation and 10% in the test set, respectively (see Table 6.1).

### 6.3.2 Generating realistic counterfactual retinal images

As mentioned above, we define visual counterfactuals as minimal, realistic and high-confidence changes to an image  $x_0$  by which a classifier's prediction can be altered to a desired target class [56]. They show what features are important for the classifier to change the decision to a particular class, and hence provide insights into what is learned by the classifier. Since the generative capabilities of a classifier are typically limited and it cannot by itself generate realistic counterfactuals, we rely on a diffusion model [71] to achieve realism. In order to generate counterfactuals, the reverse diffusion process is modified such that classifier gradients contribute to this process and guide the diffusion model towards producing counterfactuals in the desired class [57].

We will first discuss diffusion models in Section 6.3.3 and the various types of classifiers used here in

Section 6.3.4 before we introduce Diffusion Visual Counterfactuals (DVCs) in Section 6.3.5. Then, in Section 6.3.6, we briefly describe Sparse Visual Counterfactuals (SVCs) as a baseline method for counterfactual generation with retinal fundus images. Finally, in Section 6.3.7, we present the details of a user study conducted with clinicians and AI experts to evaluate the realism of generated counterfactuals.

### 6.3.3 Diffusion Models

Diffusion models are powerful generative models which can produce highly realistic images and are found to be comparable to GANs. Hence, diffusion models are an essential ingredient which ensure realism of the generated counterfactuals. See Chapter 3, Section 3.2.3 for an overview of diffusion models.

For both fundus and OCT data sets, we trained a diffusion model  $p_\theta$  for 300,000 minibatch iterations unconditionally with 1,000 time steps and a linear noise schedule for the diffusion process. The diagonal covariance  $\Sigma_\theta$  are also learned by the model during training. For the fundus data set, classes are balanced by oversampling the diseased classes to have an equal representation as that of the healthy class.

### 6.3.4 Plain and adversarially robust classifiers

It is beneficial to use adversarially robust models rather than plain models to guide the diffusion models to generate realistic and meaningful counterfactuals. For more details on plain and adversarial classifiers, see Chapter 3, Section 3.2.1.

For retinal fundus images, we trained both plain and adversarially robust classifiers in binary and multi-class settings. In the multi-class setting, the task is a 5-way classification among the classes “healthy”, “mild”, “moderate”, “severe” and “proliferative”. In the binary setting, disease onset is considered from the “moderate” class, hence, “healthy” and “mild” are grouped into the normal category and the other classes to the diseased category. For OCT scans, we trained both plain and adversarially robust classifiers in the multi-class setting to classify among the classes “healthy”, “choroidal neovascularization (CNV)”, “drusen” and “diabetic macular edema (DME)”.

All plain and robust classifiers were ResNet-50 models trained for 100 epochs with an SGD optimizer with a learning rate of 0.01 and a cosine learning rate schedule. The fundus plain classifiers were initialized with weights from ImageNet pre-trained models and the fundus robust classifier with weights from a robustly pre-trained ImageNet model [139]. All OCT classifiers were initialized with random weights. We used the cross-entropy (CE) loss as objective function for the plain model and the TRADES loss [58] with  $\varepsilon = 0.01$  for the fundus robust classifiers and  $\varepsilon = 0.5$  for the OCT classifier. For both cases, we used  $p = 2$ .

### 6.3.5 Diffusion Visual Counterfactuals

Here, we describe how to produce realistic Diffusion Visual Counterfactuals (DVCs). Following [57], we combined an unconditionally trained diffusion model  $p_\theta$  as described in Section 6.3.3 with an independently trained classifier  $f_\phi$  (see Section 6.3.4) so that the diffusion model can generate class-conditional samples. This is done by shifting the mean of the reverse transition probabilities by a value which depends on the gradients of external classifiers. More specifically, this value is the projection of the gradients of a robust classifier on a cone around the gradients of the plain classifier (Fig. 6.1). This, however, does not ensure that the generated image will stay close to the original image  $x_0$  in pixel space, which is one of the qualifying factors for realistic visual counterfactuals. Therefore, to obtain a counterfactual that remains structurally close to the original image,  $x_0$ , we find it beneficial to add a distance regularization term to the sampling process. As a further measure to avoid generating images that deviate too much from the original, we start the reverse of the diffusion process from the noisy image at step  $\frac{T}{2}$  instead of the completely distorted version of the image at the last step  $T$  [57] (Fig. 6.1). For a complete description of the generation of diffusion visual counterfactuals, see Chapter 3, Section 3.2.4.

Our code was based on <https://github.com/valentyn1boreiko/DVCEs> and will be available at [https://github.com/berenslab/retinal\\_image\\_counterfactuals](https://github.com/berenslab/retinal_image_counterfactuals).

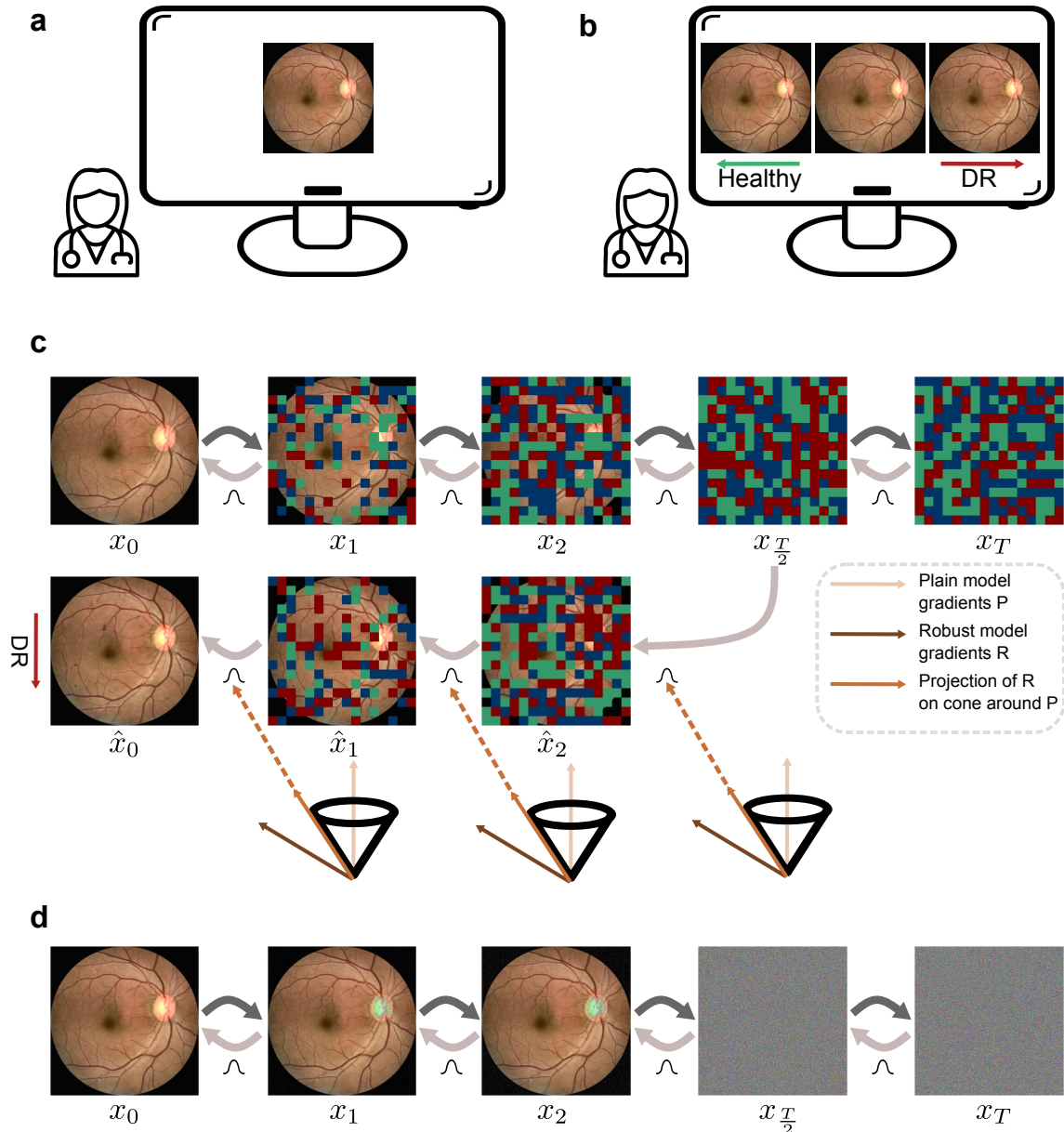


Figure 6.1: **a.** Original retinal fundus image, **b.** Visualization of counterfactuals with the healthy counterfactual on the left, DR counterfactual on the right and original image in middle, **c.** Method to generate diffusion counterfactuals. Top shows the forward and reverse diffusion for an original image  $x_0$ . Bottom shows generation of a DR DVC starting from the  $\frac{T}{2}^{th}$  time step. The mean of distributions in reverse diffusion is shifted using projected gradients (shown in dark orange) of an adversarially robust classifier (shown in brown) on a cone around the gradients of a plain classifier (shown in light orange), **d.** Images from the actual forward diffusion corresponding to the time steps shown in c.

### 6.3.6 Prior work: Sparse Visual Counterfactuals

Previous studies on generating retinal counterfactuals either use StyleGANs [129] or adversarially robust classifiers [69]. While the StyleGAN approach is closer to our approach as it uses a generative model, the code or model information is not adequately provided for reproducing the results presented. Hence, for comparison, we used a previously suggested method for generating Sparse Visual Counterfactuals (SVCs) requiring an adversarially robust classifier [56] or at least an ensemble of plain and adversarially robust classifiers [69]. For an overview of methods used to generate sparse visual counterfactuals, see Chapter 3, Section 3.2.2.

Table 6.2: Evaluation of plain and robust classifiers in terms of standard and balanced accuracy

	Binary fundus		5-class fundus		OCT	
	acc.	bal. acc.	acc.	Quad. $\kappa$	acc.	bal. acc.
Plain	92.39	80.67	86.65	0.67	96.35	95.87
Robust	90.03	74.35	83.69	0.51	95.03	93.29

The main drawback of sparse counterfactuals is that visual inspection showed that healthy counterfactuals from the DR class using retinal fundus images covered up the lesions on the fundus image with artificial looking blood vessels in a previous study [69]. Although this achieved the effect of removing the lesions to make the image look healthy, these changes do not appear realistic (e.g. see Fig. 2, second row in [69]). A more realistic change would have been to cover up the lesions with the background colors instead of adding artificial structures.

### 6.3.7 User study

To evaluate the realism of the generated counterfactuals, we performed a user study with AI experts as well as trained ophthalmologists. We built a web-based image evaluator based on the Python web framework Django (v. 4.2.1) with a PostgreSQL (v. 15.3) backend database, available at <https://github.com/berenslab/retimgtools/tree/v.1.0.0>. On the front-end, we used custom JavaScript to modify various presentation parameters (e.g. hiding the images after a certain number of seconds).

Seven ophthalmologists who had a clinical experience of 2, 4, 5, 9, 9, 10 or 14 years participated in the study (including author LaK). In addition, 4 AI experts working on applying deep learning for clinical tasks in ophthalmology took part and provided their input (including authors PB and LiK).

All participants were given a three-way odd-one-out task where they had to identify the generated counterfactual among three images. This task design is recommended for this type of study as it is highly sensitive for detecting the odd-one-out category [140, 141]. Each trial thus consisted of two real images from the data sets and one counterfactual generated by a model. Images were displayed for a maximum of 20 seconds and then hidden. All 3 images in any question belonged to the same class. For example, for a question showing DR images, we show two real DR images and one generated counterfactual with DR as the target class. The latter is generated from an image which is labeled as healthy in the data set and classified as healthy by the classifier.

For retinal fundus images, a total of 80 trials were performed with a randomly chosen set of 40 questions showing sparse counterfactuals and the remaining half showing diffusion counterfactuals as the generated image. Within each group, 50% questions belonged to the healthy class and the rest to DR. For OCT scans, on the other hand, only diffusion counterfactuals were shown as the generated images in all 80 questions as study time was a limiting factor with four disease categories. Questions were equally split across the four disease categories with 20 questions for each class. Similar to the fundus scenario, OCT counterfactuals for questions belonging to the healthy category are generated from any of the three disease classes and vice-versa.

Ethical approval for the study was obtained from the ethics commission at the University Clinic, Tübingen (Ref No. 250/2023BO2). Statistical analysis was performed using R.

## 6.4 Results

Our goal was to show that the counterfactuals based on diffusion models guided by the gradients of robust classifiers can generate minimal, meaningful and high-confidence changes to an input image such that the DNN classifier alters its decision to a desired target class and that domain experts view the resulting images as realistic. To this end, we first report the result of a user study with domain experts in order to evaluate the realism of our counterfactuals generated with the chosen parameters (Section 6.3.5). We then go into the technical factors necessary for achieving this result, establish that robust classifiers are indeed necessary and illustrate the effect of regularization strength on



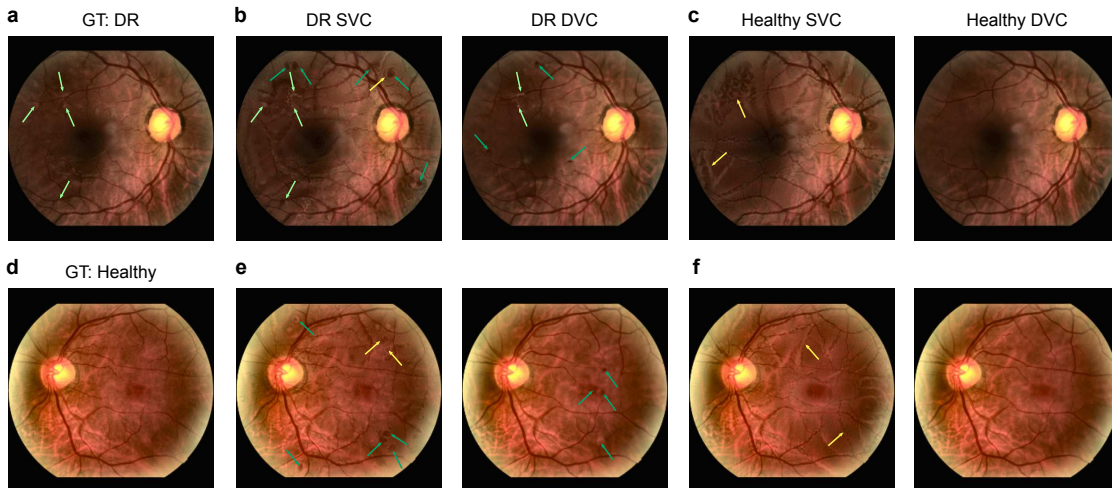


Figure 6.2: DVCs show clinically meaningful changes and appear more realistic than SVCs. **a.** Image with DR and classifier confidence  $p_\phi(\text{DR}) = 0.99$ . **b.** DR SVC (left) and DVC (right) with  $p_\phi(\text{DR}) = 1.00$  for both images. **c.** Healthy SVC (left) with  $p_\phi(\text{healthy}) = 1.00$  and healthy DVC (right) with  $p_\phi(\text{healthy}) = 0.99$ . DVCs show realistically emphasized lesions (light green arrow) and new lesions (dark green arrow). DVC shows more realistic removal of disease related lesions whereas SVCs introduce artifacts (yellow arrow). **d.-f.**, as **a.-c.**, but for a healthy fundus image  $p_\phi(\text{healthy}) = 0.90$ . DR SVC:  $p_\phi(\text{DR}) = 0.98$ ; DR DVC:  $p_\phi(\text{DR}) = 1.00$ ; Healthy SVC:  $p_\phi(\text{healthy}) = 1.00$ ; healthy DVC:  $p_\phi(\text{healthy}) = 0.99$ . All SVCs were generated with  $\ell_4$  norm and  $\epsilon = 0.3$ . DVCs were generated with  $\ell_2$  norm and regularization strength  $\lambda = 0.5$ .

the generated images. Finally, we demonstrate multi-class counterfactuals and counterfactuals for retinal OCT scans, for which we also evaluate the realism in a user study.

#### 6.4.1 Fundus diffusion counterfactuals are realistic

We trained binary plain and robust DNN classifiers for DR based on fundus images with high accuracy on a large and diverse fundus image dataset (Table 6.2). For details of the dataset, see Table 6.1; for details of training procedure, see Section 6.3.4). As expected, the robust classifier had lower accuracy than the plain one. In addition, we also trained a diffusion model on the same dataset augmented by additional datasets to add more diseased examples (for details of the training procedure, see Section 6.3.3). We used the diffusion model with cone projected gradients in order to generate realistic diffusion counterfactuals such that if an image showed signs of DR, the counterfactual could either remove these signs (“healthy diffusion counterfactual”) or reinforce them (“DR diffusion counterfactual”). Likewise, the diffusion counterfactual could either add signs of DR to a healthy original image or strengthen its healthy appearance. Thus, the model was able to generate images that illustrate what the fundus image of a patient might have looked like, had he or she been more or less progressed in their disease (the definition of a counterfactual). We compared the diffusion counterfactual method to the previously published sparse counterfactuals method [69].

We found that the diffusion model generated visually realistic counterfactual fundus images from either DR (Fig. 6.2 **a**) or healthy starting images (Fig. 6.2 **d**). For example, a DR diffusion counterfactual generated from a DR fundus images enhanced the existing lesions and added new lesions (Fig. 6.2 **b** right panel). On the other hand, a DR diffusion counterfactual generated from a healthy image produced diverse lesions including images regions that resembled microaneurysms, haemorrhages and exudates (Fig. 6.2 **e** right panel). Further, the structural details in the retina including the blood vessels, macula and optic disc were largely preserved on the diffusion counterfactuals of any given subject’s fundus image. In comparison, baseline sparse counterfactuals appeared more artificial (left panels in Fig. 6.2 **b,c,e,f**). Sparse counterfactuals introduced artifacts such as waves around lesions in DR counterfactuals (Fig. 6.2 **b,e**) and lines in healthy counterfactuals (Fig. 6.2 **c,f**). For more examples, see Appendix B.2.

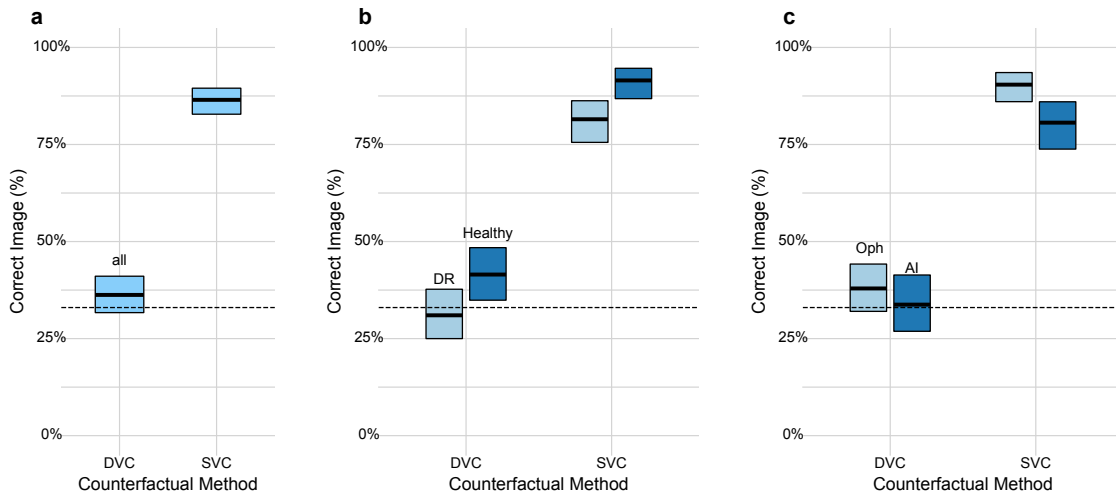


Figure 6.3: User study of realism of generated DVCs. We asked  $n = 4$  AI experts and  $n = 6$  ophthalmologists to identify a counterfactual in a odd-one-out task with three images (two real and one counterfactual). **a.** Overall fraction of correctly identified counterfactuals with binomial 95%-CI. Baseline at 33% chance level (dashed line). **b.** As in **a.** for the healthy and DR classes. **c.** As in **a.** for ophthalmologists and AI experts.

Table 6.3: Generalized Linear Model to assess the influence of factors in Fig. 6.3.  $n = 800$

Predictor	Odds Ratio	CI	p-value
SVC vs. DVC	12.03	8.45 – 17.40	$\lll 0.0001$
healthy vs. DR	1.82	1.30 – 2.57	0.0005
Ophthalmologist vs. AI researcher	0.66	0.47 – 0.94	0.0197

To assess whether the generated diffusion counterfactuals are realistic, we performed a user study with four AI specialists, who worked with ophthalmological data on a regular basis, and six ophthalmologists with different levels of experience (see Sec. 6.3.7). In a three-way odd-one-out task, we asked to identify the image likely to have been generated by an AI model. The shown images included both healthy and DR diffusion and sparse counterfactuals. Interestingly, all participants found it challenging to distinguish diffusion counterfactuals from real fundus images whereas they easily spotted the sparse counterfactuals (Fig. 6.3). In fact, across all images, participants showed a close to chance level (33%) performance for diffusion counterfactuals as opposed to a significantly better performance than chance level for sparse counterfactuals (Fig. 6.3 **a**, DVC vs. SVC: 36.3% [31.7% – 41.1%] correct, 95% CI), confirmed by statistical analysis ( $p \lll 0.0001$ , see Table 6.3).

We further analyzed if DR or healthy could be more easily identified as artificial. We found that participants could identify healthy diffusion counterfactuals more easily compared to DR diffusion counterfactuals (Fig. 6.3 **b**), potentially because diffusion models appear to smooth the image during removal of lesions and sometimes fail to remove all traces of lesions ( $p = 0.0005$ , see Table 6.3). Finally, we studied whether trained ophthalmologists were more likely to identify diffusion counterfactuals than AI specialists. Interestingly, we found that this difference was not major, with all ophthalmologists independent of experience levels being close to chance level, at a similar level as AI specialists (Fig. 6.3 **c**). Ophthalmologists detected sparse counterfactuals at an average rate of 90.4%, significantly better than AI specialists who detected the same at an average rate of 80.6% ( $p = 0.0197$ , see Table 6.3).

In summary, we found that our diffusion counterfactual model can generate realistic looking fundus images from both healthy and DR images, emphasizing or removing signs of the disease. We showed that the images generated by our new model are almost impossible to detect even for highly trained experts, in contrast to images created by previous techniques.

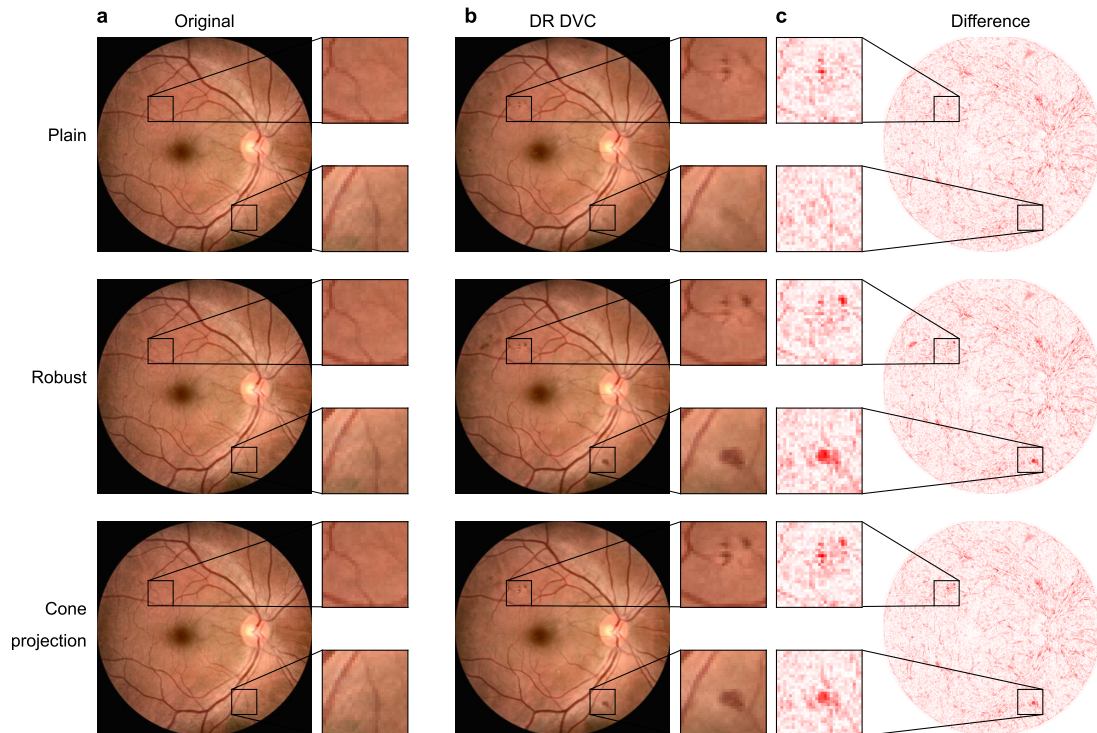


Figure 6.4: Comparison of DVCs generated using the plain model (top row), robust model (middle row) and cone projection of an adversarially robust model onto a plain model (bottom row). **a.** A DR fundus image with  $p_\phi(\text{DR}) = 1.00$  with a zoom in on patches with lesions. **b.** DR DVCs for the image from **a.** for the three different models. **c.** Difference maps between original DR image and the DR DVC show robust and cone projection models produce more realistic changes than the plain model

### 6.4.2 Realistic counterfactual examples require robust classifiers

Now that we established that we are able to generate realistic looking counterfactuals for fundus images, we explore the technical ingredients necessary to achieve this. First, as discussed in Sec. 6.3.4, the gradients of the output a standard classifier with respect to the image often do not represent meaningful changes, but rather lead to the generation of adversarial examples that fool the classifier but are imperceptible for humans [142]. In fact, for our diffusion model, the gradients of a "plain" classifier often were not strong enough to provide guidance towards the target class and hence, the resulting counterfactuals looked quite similar to the original image (Fig. 6.4 **a-c**, top row). This effect was more prominent in DR diffusion counterfactuals generated from healthy images, where the plain classifier's gradients induced hardly noticeable lesions, compared to healthy counterfactuals generated from DR images, where the diffusion model removed lesions even when guided by the plain classifier (compare Fig. 6.4 top row to Appendix. B.3).

In contrast, the gradients of the output of a robust classifier with respect to the image (see Sec. 6.3.4) supported the generation of high quality DR diffusion counterfactuals, with clearly visible and highly realistic lesions (Fig. 6.4 **a-c**, middle row). As discussed, the robust classifier, however, traded robustness against accuracy, leading to a drop in performance (see Table 6.2). To obtain high quality diffusion counterfactuals while maintaining high classification accuracy, we combined the plain and the robust classifier gradients using cone projection (see Sec. 6.3.4). Here, the gradients of the robust classifier are projected onto a cone around the gradients of the plain classifier. In this case, the generated DR diffusion counterfactuals were almost as good for the robust classifier alone ((Fig. 6.4 **a-c**, bottom row and Appendix B.4), while maintaining a high balanced accuracy (Table 6.2). Therefore, our final model evaluated in the user study above used cone projection for generating realistic diffusion counterfactuals. In the sections that follow, all diffusion counterfactuals are generated using the cone projection method.

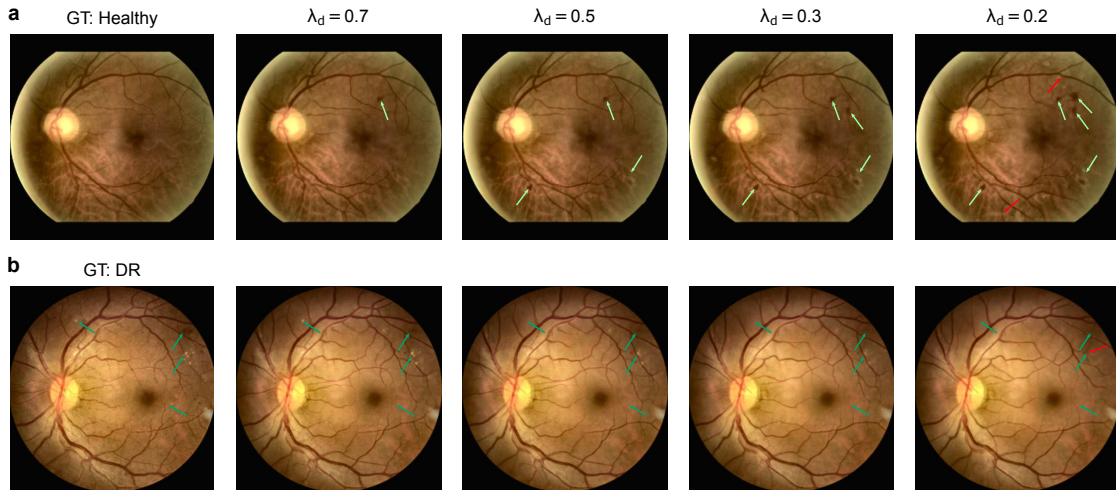


Figure 6.5: Effect of tuning the regularization strength  $\lambda_d$  on generated DVCs. Decreasing  $\lambda_d$  allows for more changes on the original image. **a.** We start with a healthy image and generate DR DVCs with decreasing  $\lambda_d$ . More lesions are generated as  $\lambda_d$  decreases (light green arrows). **b.** We start with a DR image and generate healthy DVCs with different  $\lambda_d$  values. Some traces of the lesions were still visible for  $\lambda_d = \{0.7, 0.5\}$  while they were completely removed for  $\lambda_d = 0.2$  (dark green arrows) at the cost of some changes to the vessel structure (red arrows). While a higher  $\lambda_d = 0.7$  is sufficient to generate the minimum number of lesions required to convert a healthy fundus to DR, it is not sufficient to remove all lesions on a DR image to convert it to a healthy fundus.

### 6.4.3 Influence of regularisation strength on diffusion counterfactuals

Next, we explored the effect of the regularization strength  $\lambda_d$  (Eqn. (3.14)), which constrained the distance of the generated diffusion counterfactual from the original image. This parameter controls the extent of changes appearing on the diffusion counterfactuals compared to the original image, with high values indicating that the generated image remains closer to the original. Without regularization, the diffusion model is not constrained to keep the generated image close to the original image and guided by the gradients from the classifier/ cone projection it can generate any image belonging to the target class without necessarily preserving the background of the original image including color and vessel structure.

We systematically observed the pattern of changes in both DR and healthy diffusion counterfactuals as we lowered the parameter  $\lambda_d$ . In general, for both DR and healthy diffusion counterfactuals, more changes were visible on images as  $\lambda_d$  was decreased. For DR counterfactuals, the size, number and sharpness of lesions increased with decreasing regularization strength. Thus, with a strong regularization of  $\lambda_d = \{0.7, 0.5\}$ , fewer lesions appeared on DR diffusion counterfactuals which were relatively smaller and in some cases not too sharp and distinct (Fig. 6.5 a). As the strength was decreased to 0.3, more lesions were generated and their sharpness increased compared to the ones generated with higher regularization values. For all these values of  $\lambda_d$ , almost all counterfactuals were labeled as DR (see Table 6.4). At 0.2, the diffusion model had the freedom to make several modifications to the original image and it added various bright and large lesions while also modifying the blood vessel structure to a larger extent compared to the other regularization values.

We repeated the above for healthy counterfactuals. Notably, we found that lesions were generally removed well in mild and moderate DR already at larger regularization strengths of  $\lambda_d = \{0.7, 0.5\}$ , but some traces of lesions were still visible with for severe, proliferative and a few moderate cases (Fig. 6.5 b). As regularization was decreased further to 0.3, the sharpness and clarity of those lesions decreased. At  $\lambda_d = 0.2$ , the lesions were completely removed, however, the blood vessel structure was also heavily altered. For extreme cases of DR, where the entire retina was affected such as in a few proliferate cases, even a regularization of 0.2 was not sufficient to remove all the lesions (Appendix B.5 a-b). Furthermore, with  $\lambda_d = 0.7$ , a third of healthy diffusion counterfactuals

$\lambda_d$	Healthy $\rightarrow$ DR	DR $\rightarrow$ healthy
0.7	4.1%	27.0%
0.5	1.4%	14.0%
0.3	0%	6.0%

Table 6.4: Fraction of images that do not change the class label depending on the choice of regularization parameter  $\lambda_d$ . For this analysis, 40 images were chosen from each of the five classes, such that there were 80 images for the “healthy to DR” direction and 120 for “DR to healthy”. For 73/80 and 100/120, class labels were correctly predicted for the original image. Then we evaluated the class label of the corresponding counterfactual.

generated from DR fundus images did not change the prediction of the DNN classifier to the healthy class. In contrast, for  $\lambda_d = 0.5$  and  $\lambda_d = 0.3$ , 14% and 6% did not convert to healthy class, respectively (Table 6.4). As we evaluated the classifier at a “referable DR” scenario, a healthy diffusion counterfactual may contain small traces of lesions as seen in mild DR fundus images even when the prediction of the classifier changes to healthy. For more examples of diffusion counterfactuals with varying  $\lambda_d$ , see Appendix B.6.

Taken together, we found that using values of  $\lambda_d$  such as 0.2 or lower resulted in larger changes to the original image than necessary for conversion to the target class, producing large changes to the vessel pattern. While high  $\lambda_d$  such as 0.7 and higher was sufficient for the DR diffusion counterfactuals to show minimal features required to convert healthy fundus to DR, the same did not hold for healthy diffusion counterfactuals generated from DR fundus images (Table 6.4). Therefore, we chose a regularization value of 0.5 where we could qualitatively observe the minimal changes on the image necessary to alter the decision and confidences of the classifier in both directions while maintaining image structure close to the original (although within a range of  $\lambda_d = 0.3 - 0.5$ , this is frankly a qualitative judgment).

#### 6.4.4 Diffusion counterfactuals for the multiclass DR grading task

We followed up the counterfactuals for the binary case of healthy versus DR with counterfactuals for a more fine-grained classification scenario with 5 classes, healthy, mild, moderate, severe and proliferative. The latter four categories are the various stages of DR in the order of increasing severity. The mild class often shows only very tiny changes in the form of microaneurysms and is the hardest to detect. Moderate and severe are characterized by the presence of a relatively greater number of microaneurysms and larger lesions such as hemorrhages and exudates. The proliferative class is the most advanced stage with venous bleeding, large haemorrhages and neovascularization. Some of the images in the proliferative and severe stages also show scars resulting from laser treatment. Typically severe and proliferative classes are easier to detect due to larger lesions however due to rare occurrences they are underrepresented in the data set.

We generated diffusion counterfactuals to the 5 different classes from originally healthy, mild and moderate fundus images (Fig. 6.6). First, we looked at diffusion counterfactuals to the various DR stages from a healthy image and found that the diffusion counterfactuals contained meaningful features for both mild and moderate classes. The features included tiny dot-like microaneurysms/exudates for mild class and slightly larger and more exudates, microaneurysms and a few haemorrhages for moderate class (Fig. 6.6 a). However, for the severe and proliferative classes most often only a couple of scattered haemorrhages were generated and most other features such as bleeding or the laser scars were not observed (Fig. 6.6 a). This was likely due to the scarcity of these classes in the data set. Another technical factor could be the choice of parameters such as the regularization value  $\lambda_d$  and the radius  $\varepsilon$  which were more suitable for smaller changes.

For healthy diffusion counterfactuals from both mild and moderate classes, all lesions were removed completely in most cases with a regularization strength of 0.5 (Fig. 6.6 b-c). On a moderate diffusion counterfactual generated from a fundus image originally belonging to the mild class, the number of exudates increased and an existing exudate was slightly enlarged (Fig. 6.6 b). The diffusion counterfactual from moderate to mild interestingly removed the exudates all over the fundus and added a single microaneurysm (Fig. 6.6 c). Here again, diffusion counterfactuals to the more advanced stages of severe and proliferative did not exhibit the relevant features for those

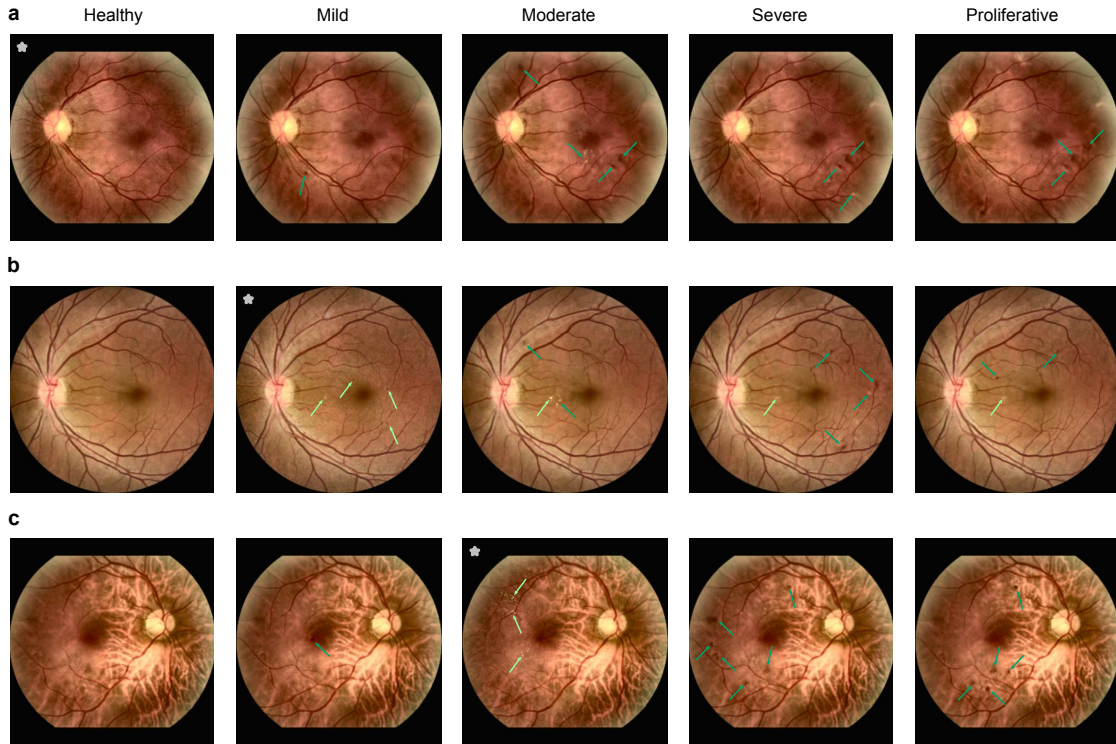


Figure 6.6: DVCs for DR grading task with 5-classes: healthy, mild, moderate, severe and proliferative. Images marked with \* are original images with GT as indicated in the headline. DVCs to the different classes from **a.** healthy fundus, **b.** fundus with mild DR and **c.** fundus with moderate DR. Lesions which are originally present in initial image are indicated with light green arrows while lesions added by DVC are indicated with dark green arrows. In all cases, healthy DVCs removed all lesions. While the number and types of lesions introduced in mild and moderate DVCs are consistent with those observed in real-world data, severe and proliferative DVCs did not reflect the size and intensity of lesions seen in real examples.

classes (Fig. 6.6 **b-c**). Furthermore, in all cases, the plain classifier achieved high target probabilities in the range  $[0.85, 1.00]$  for diffusion counterfactuals to the healthy, mild and moderate classes while having only low target probabilities which dropped to below 0.25 for the severe and proliferative diffusion counterfactuals.

To summarize, the 5-class model could generate meaningful diffusion counterfactuals to the healthy, mild and moderate classes while it was not as efficient at generating severe and proliferative cases. Nonetheless, mild and moderate classes are the clinically more interesting stages as they are challenging to detect and diagnostic decisions are uncertain for not only DNNs but also ophthalmologists around the boundaries of these early stages [143]. Studying the progression of biomarkers closely in these stages with counterfactuals can help prevent conversion to the more advanced stages.

#### 6.4.5 Diffusion counterfactuals of OCT scans are also realistic

Finally, we trained another set of diffusion model and classifiers on a database of 96,441 OCT scans, a different image modality which is also predominantly used in ophthalmology (see Table 6.1). While the diffusion model was trained to generate realistic OCT scans, the task of the classifiers was to detect whether a given scan was healthy or had one among the three conditions: choroidal neovascularization (CNV), drusen or Diabetic Macular Edema (DME), which both plain and robust classifiers were able to do with high accuracy (see Table 6.2).

OCT scans can visualize a cross-section of the retina and typically show the different layers of the retina, from the vitreo-retinal interface, inner retina, outer retina, retinal pigment epithelium (RPE)/Bruch's membrane to the choroid from top to bottom. The biomarkers of CNV on OCT

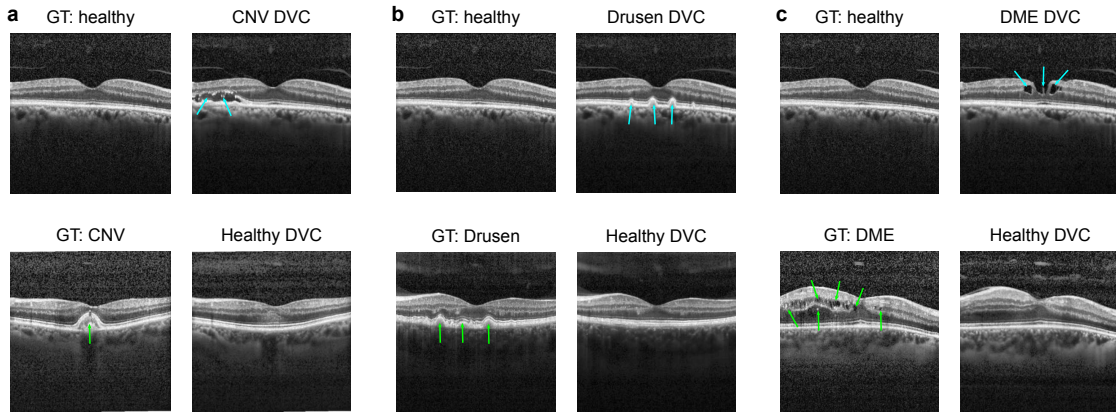


Figure 6.7: DVCs of OCT images from healthy to various disease classes and vice-versa. **a.** DVC from healthy to CNV (top) and from CNV to healthy (bottom). **b-c.** Same as **a** for classes drusen (**b**) and DME (**c**). Similar to fundus DVCs, OCT DVCs show meaningful changes which are consistent with the important features of each class. DVCs from healthy images add features relevant to the disease (blue arrows). DVCs from diseased images to the healthy class remove the disease specific features seen on original image (green arrows).

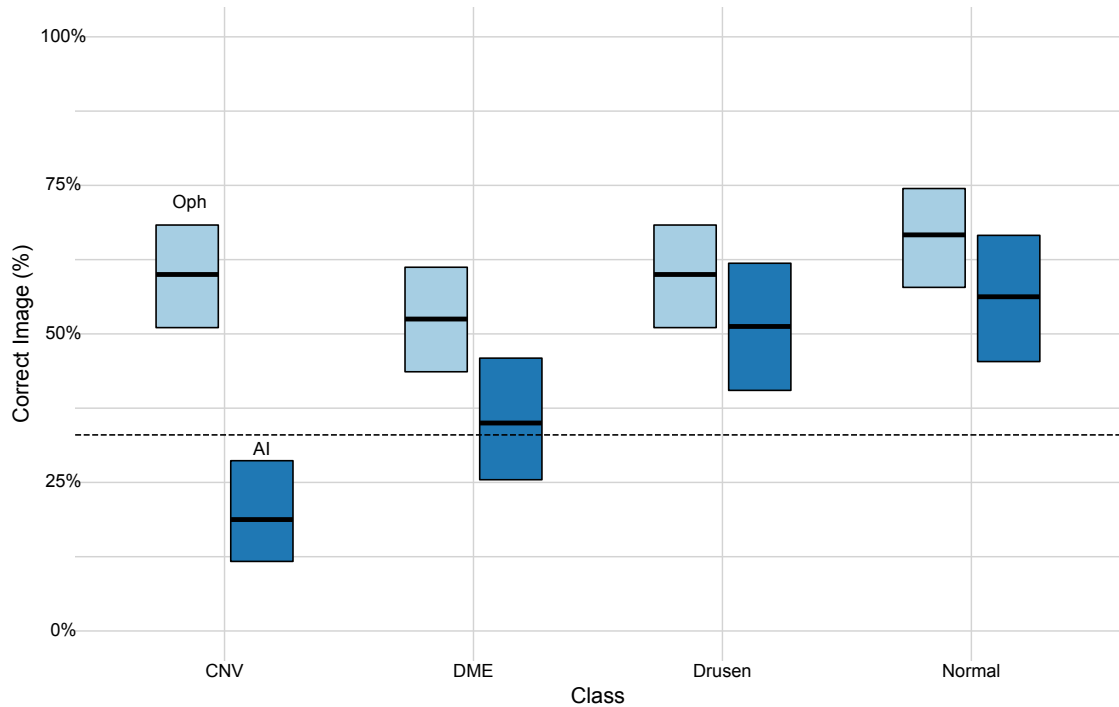


Figure 6.8: Clinical evaluation of realism of generated OCT DVCs. We asked  $n = 4$  AI experts and  $n = 6$  ophthalmologists to identify a DVC in a odd-one-out task with three images (two real and once DVC). **a.** Overall fraction of correctly identified DVCs with binomial 95%-CI. Baseline at 33% (dashed line). **b.** As in **a**.

scans include subretinal neovascular membrane, subretinal fluid and intra-retinal fluid. These occur due to abnormal growth of new vessels in the choroid creating a rupture in the retinal layers above the Bruch's membrane. Drusen are characterized by a bumpy or irregular RPE layer due to lumps of deposits under the RPE. OCT scans of subjects with DME contain several cavity-like structures in the inner and sub retinal layers which represent intraretinal and subretinal fluid that accumulates due to vascular leakage [26]. Upon visual inspection, we found that the diffusion counterfactuals seemed to effectively capture salient features of the various classes.

Table 6.5: Generalized Linear Model to assess the influence of factors in Fig. 6.8.  $n = 800$ 

Predictor	Odds Ratio	CI	p-value
CNV vs. DME	1.09	0.73 - 1.63	0.6817
CNV vs. drusen	1.72	1.15-2.58	0.0082
CNV vs. normal	2.23	1.49 - 3.37	0.0001
Ophthalmologist vs. AI researcher	0.44	0.33 - 0.60	< 0.0001

We generated diffusion counterfactuals using the cone projection method with the OCT scans to the three disease categories from healthy and vice-versa. Diffusion counterfactuals contained meaningful and superficially realistic changes similar to fundus images. Diffusion counterfactual to CNV from healthy added subretinal fluid below the RPE (Fig. 6.7 a, top row). On the other hand, the diffusion counterfactual from CNV to healthy removed the subretinal neovascular membrane and flattened out the portion where it was present (Fig. 6.7 a, bottom row). While diffusion counterfactual from healthy to drusen class added bumps to the RPE layer (Fig. 6.7 b top row), in the reverse case the irregularities were removed to make the RPE layer smooth and flat (Fig. 6.7 b bottom row). DME diffusion counterfactuals generated from the healthy class contained cavities in the inner retinal layers (Fig. 6.7 c top row). On the other hand, healthy diffusion counterfactual from a DME OCT scan covered up the cavities with the original tissue reflectivity (color) in those layers (Fig. 6.7 c bottom row). Hence, in all cases, diffusion counterfactuals generated meaningful structures associated with the target class.

To assess the degree of realism of the generated images, we performed a user study again with six ophthalmologists and four AI researchers. Similar to the fundus user study, the participants were assigned a three-way odd one out task here too although they were shown only diffusion counterfactuals across the four different categories. Ophthalmologists consistently performed better than chance (33.3%) in all classes (Fig. 6.8, indicated by non-overlapping 95%-CIs). Interestingly, they detected diffusion counterfactuals to the normal class from the various disease classes with the highest rate of 66.7%. This could have been due to the normal diffusion counterfactuals generated from OCT scans with signs of extreme CNV or DME. In such scenarios, the normal DVCEs generally tended to fill up the cavities or tears with original tissue reflectivities but did not restore the thickness of the layers (Fig. 6.7 c bottom row) thereby resulting in easier detection. They found diffusion counterfactuals to the CNV and drusen classes also easier to detect (Fig. 6.8). This could be due to certain features that looked artificially generated for e.g. the perfect waves on drusen diffusion counterfactuals. On the other hand, diffusion counterfactuals to the DME class was the hardest to detect for the ophthalmologists (Fig. 6.8). AI experts performed significantly worse overall (see Table 6.5) compared to ophthalmologists, this was especially true in the CNV and DME classes (Fig. 6.8). We attribute this to the relatively low experience of AI experts with these disease categories such that realistic looking images not showing realistic disease features led to performance at chance level for AI experts.

Taken together, OCT diffusion counterfactuals were able to generate the primary features associated with each class convincingly although with a few imperfections which led to their easier detection in the user study compared to fundus diffusion counterfactuals.

## 6.5 Discussion

In conclusion, we showed that diffusion models guided by the gradients of robust classifiers can be used to generate realistic counterfactuals for retinal fundus and OCT images. We found that domain experts in ophthalmology including clinicians and AI researchers could hardly distinguish these images from real images, opening up new opportunities to include counterfactual images in medical reasoning [127].

While the counterfactuals for retinal fundus images were nearly indistinguishable from real ones for clinicians, OCT counterfactuals were relatively easier to detect. It is interesting to speculate what could have caused these differences. Likely, one factor is that lesions for early DR stages visible in fundus images are mostly localized and not too large, therefore not requiring major structural changes to large parts of the image, in contrast to what is needed for generating OCT



counterfactuals. Also, OCT counterfactuals typically looked too regular and symmetric, such as in the case of drusen counterfactuals. Qualitative feedback after the user study indicated that raters were quickly able to pick up on these regularities. In addition, healthy OCT counterfactuals from extremely diseased cases often covered the abnormalities with appropriate texture but did not alter the thickness of the retina at that stage of the image which is an important factor for clinicians to classify the image as healthy. Since most cases in the chosen OCT data set also belong to extreme disease stages, this could have further impacted the overall performance of clinicians in the detection of OCT counterfactuals. In contrast, fundus images covered the whole disease spectrum, allowing the model to learn about the gradual changes along the disease trajectory.

It is possible that the different degrees of realism of fundus vs. OCT counterfactuals come either from the generative capabilities of the diffusion model or what is learned by the classifiers. In fact, we noticed that supplementing the fundus image dataset with more examples from diseased classes helped to generate better disease counterfactuals. However, it is also possible that the disease concepts learned by the classifiers which guide the diffusion model are insufficient. For example, with the OCT dataset consisting mostly of extreme examples at advanced disease stages and missing much of the borderline cases in between [138], a classifier might have easily taken a shortcut towards distinguishing healthy from diseased images, picking up on the most informative feature (e.g. texture) while ignoring more subtle ones (retinal thickness) [144]. In turn, improved classifiers based on more realistic and varied datasets may therefore yield even better counterfactuals. In fact, it would be interesting to study how the generative capacities of the model change with even larger datasets, such as those used to train retinal foundation models, or based on robust classifiers derived from such foundation models [145].

Diffusion models have been largely used in combination with plain classifiers for realistic counterfactual generation in the natural image domain using data sets such as ImageNet and CelebA [134, 146]. The quality and realistic nature of counterfactuals for such images has been shown to improve when adversarially robust classifiers are used [57]. In the medical setting, diffusion models have been used primarily for generating healthy counterfactuals [133, 147, 132], which is an easier task for the diffusion model compared to generating disease related features. We demonstrated that both diffusion models and adversarially robust classifiers play a major role in generating realistic medical counterfactuals for high-resolution retinal fundus and OCT images. Moreover, our counterfactuals are bi-directional, i.e. from healthy to diseased and diseased to healthy. In parallel work, BioMedJourney [148] uses two consecutive Chest-*X*Ray images of a subject and a summary of their medical reports to generate longitudinal counterfactuals. Here, a latent diffusion model is trained on embeddings of the textual descriptions and a starting image to obtain an estimate of the progressing image. This method relies on the availability of detailed medical reports in addition to longitudinal imaging data.

Realistic counterfactuals have the potential to be used in clinical decision support where the DNN can provide human-like and human-understandable reasoning for its prediction. For example, decision support is conceivable that illustrates for a given patient with an uncertain diagnosis how the imaging data might look if it provided less ambiguous evidence for the presence of a disease, potentially even with more than one sample for a given input image. The clinician could then use similarity of the present image to judge the presence or absence of disease signs. Realistic counterfactual images could also be used to synthesize data for training clinicians and augmenting DNN models, as medical data sets are often imbalanced and diseased samples may be less readily available. Due to the realistic nature of counterfactuals, diseased data points could then be synthesized based on a preliminary classifier from the more prevalent healthy examples [149, 150]. In fact, adding a diseased counterfactual for each healthy image and vice versa would effectively create a paired dataset, where structurally similar images derived from the same base image are contained in the healthy and diseased class, allowing the classifier to focus more easily on the disease patterns. In the reverse case of generating healthy examples from diseased, the counterfactuals could help in anomaly detection and identification of bio-markers [133, 147, 132]. Further, they could be used as a testing tool to ensure that the classifiers do not use any shortcuts to make the decisions, such as hospital or device logos instead of disease related features.

A natural extension of this work would be to generate counterfactuals from multi-task DNNs which learn several attributes simultaneously [151]. With such a DNN, it would be possible to generate counterfactuals for one attribute keeping another fixed. For instance, a multi-task classifier which

is trained on both age and disease type can be used to generate counterfactuals for increasing age keeping the disease fixed or vice-versa. Such counterfactuals could potentially be used in tracking the progression of a disease with age. Similarly, it would be interesting to study counterfactuals for longitudinal data or data with interventions, such as the administration of a drug. For example for OCT images during age-related macular degeneration, treatment effects for the injection of anti-VGEF drugs might be simulated for the different available drugs, and the most promising drug chosen.

## Acknowledgments

We acknowledge support by the German Ministry of Science and Education (BMBF; 01IS18039A), the Deutsche Forschungsgemeinschaft through a Heisenberg Professorship (BE5601/8-1) and under Germany’s Excellence Strategy – Excellence Cluster ”Machine Learning — New Perspectives for Science” EXC2064/1 — Project number 390727645), the Carl Zeiss Foundation (project “Certification and Foundations of Safe Machine Learning Systems in Healthcare”) and Gemeinnützige Hertie Stiftung. This research utilized compute resources at the Tübingen Machine Learning Cloud, INST 37/1057-1 FUGG. PB is a member of the Else Kröner Medical Scientist Kolleg ”ClinbrAI: Artificial Intelligence for Clinical Brain Research”.

# Chapter 7

## Conclusion

### 7.1 Summary and contributions

In the future, clinical workflows are predicted to increasingly rely on the synergy between clinicians and automated systems that use deep learning algorithms. In such situations, explanations will play an important role in facilitating co-operation between humans and automated diagnostic systems. This work is a preliminary step in this direction that illustrates how CNN models can offer visual explanations for their decisions in diverse clinical diagnostic tasks in ophthalmology through the use of two distinct methods. One method utilized BagNets, which are inherently interpretable models, while the other method generated clinically realistic counterfactual explanations.

Human-friendly and easily understandable explanations enhance the trust of human users on deep learning model decisions [42]. They enable clinicians and other users to visualize errors made by the model, identify the reason behind the errors and systematically debug the models. They aid in identifying biases and shortcuts picked up by the model during training. More importantly, explanations can provide scientific insights, for instance, by highlighting a new biomarker related to a particular disease that was previously unknown to clinicians. Ideally, explanations can also provide a mechanism for clinicians to review their decisions and spot any missed diagnostic features in medical images [43]. Besides, these use cases humans also tend to rely on explanations to learn the task better and aid developers in building better classification models [152].

The landscape of explanations from CNN models, especially in medical classification tasks, has been dominated by methods that produce saliency or heat maps such as Integrated Gradients [55], Guided Backpropagation [54] or GradCAM [153]. Such saliency maps are not found to be human-friendly and are shown to be ineffective in highlighting features that affect the model's decision [154, 104]. For example, some of these methods produce similar maps irrespective of whether a trained model or a model with random weights is used [154]. Particularly in medical classification tasks, saliency map methods fail to produce useful explanations for the healthy cases [104]. In this work, we presented methods that overcome these limitations and provide a step forward towards producing explanations from CNN models for medical tasks that are closer to the principles of human explanations.

### 7.2 Future work

Firstly, both explainability approaches investigated in this work hold considerable promise for various real-world clinical applications. There is a huge scope for using BagNets to aid clinicians in medical diagnostic tasks such as the detection of diabetic retinopathy from fundus images. One example is the Sparse BagNet [155] model which is capable of effectively localizing patches with disease features in the early stages of diabetic retinopathy. In this case, one could make the explanations more informative and user-friendly by annotating the identified patches. This could be achieved by labelling the patches with the help of another CNN model that classifies the different types of lesions. Likewise, counterfactual explanations show immense potential for use in clinical decision-making scenarios. They could be generated to answer questions that could help

in treatment interventions such as: how would a diabetic retinopathy patient’s medical image look if eye injections were administered? They could also be used to visualize progression of disease by generating different possibilities of the current medical image instance with respect to varying ages of the patient. Such applications could be realized with multi-task models with several heads [151] that are trained to perform different tasks simultaneously such as detecting diseases alongside factors such as treatment type or age. Furthermore, there are future possibilities to improve the usability of these methods by combining them together. For example, a BagNet model could help in localizing the patches on a counterfactual explanation where changes are applied.

Secondly, despite the foundational similarities between the presented methods and human explanation mechanisms, there are still gaps to be addressed. These methods mainly focused on the cognitive and generative process of explanations and not on the social process of explanations. Besides, the explanations were derived solely from image classification models. Recently, large language models (LLMs) based on the Generative Pretrained Transformer (GPT) model [156, 157] and multimodal neural networks such as Contrastive Language Image Pretraining (CLIP) [158] have risen to the fore and opened up possibilities to generate natural language explanations similar to the human social processes. A concrete approach could be to develop chatbots by pairing general feature representations from retinal foundational models [159] with large language models trained on medical data such as the MedPaLM model [160]. Another prospective future extension of this work is to generate context-specific explanations and explanations specific to the role of the explainee. For example, the explanations provided to clinicians could include more medical jargon compared to an explanation provided to a patient.

Finally, in order to reap the benefits of explanations from machine learning models, it is essential to evaluate the explanations by including clinicians and various clinical stakeholders in the loop. Although human evaluation of explanations is a time-consuming and laborious process, this will pave the path for identifying explanations which are most suited to the task and favored by the user group. For example, on a bird classification task, humans preferred part based explanations such as prototypes or concepts rather than saliency map based explanations [152]. Similar studies could prove to be successful at identifying user preferences for explanations in the medical domain. While we evaluated the realism of generated counterfactual explanations with a clinical study involving experienced ophthalmologists, evaluation of the presentation formats and usefulness of the various explanation methods including counterfactual explanations in a clinical context still remains an open question.

# Appendix A

## Supplementary materials to Chapter 5








Orig. (GT:DR)	$l_{1.5}$ -VCE, $\epsilon=30$	$l_2$ -VCE, $\epsilon=6$	$l_4$ -VCE, $\epsilon=0.2$
DR:0.99	$\rightarrow$ DR: 1.00	$\rightarrow$ DR: 1.00	$\rightarrow$ DR: 1.00
			
	$\rightarrow$ healthy: 0.99	$\rightarrow$ healthy: 1.00	$\rightarrow$ healthy: 0.86
			

Figure A.1: Failure: when transforming to healthy using ensemble,  $l_2$ - and  $l_{1.5}$ -VCEs have visible artifacts (yellow spots), unlike  $l_4$ -VCE.








Orig.(GT:DR)	$l_{1.5}$ -VCE, $\epsilon=30$	$l_2$ -VCE, $\epsilon=6$	$l_4$ -VCE, $\epsilon=0.2$
DR:1.00	$\rightarrow$ DR: 1.00	$\rightarrow$ DR: 1.00	$\rightarrow$ DR: 1.00
			
	$\rightarrow$ healthy: 0.99	$\rightarrow$ healthy: 1.00	$\rightarrow$ healthy: 0.82
			

Figure A.2: Failure: when transforming to DR using ensemble,  $l_{1.5}$ -VCE has visible artifacts, unlike  $l_2, l_4$ -VCEs.

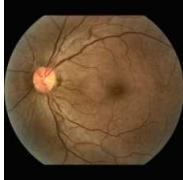
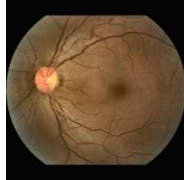
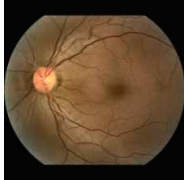
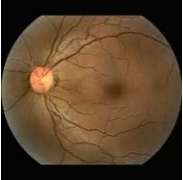
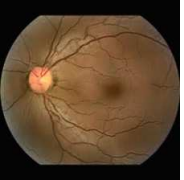
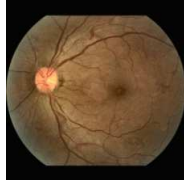
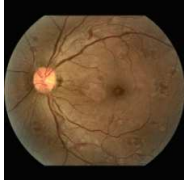
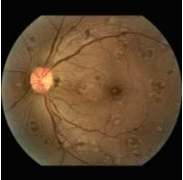
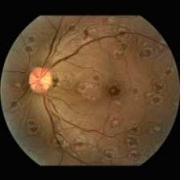
Orig.(GT:healthy)	$l_4$ -VCE, $\epsilon=0.1$	$l_4$ -VCE, $\epsilon=0.2$	$l_4$ -VCE, $\epsilon=0.3$	$l_4$ -VCE, $\epsilon=0.4$
healthy:0.77	$\rightarrow$ healthy: 0.95	$\rightarrow$ healthy: 0.99	$\rightarrow$ healthy: 1.00	$\rightarrow$ healthy: 1.00
				
	$\rightarrow$ DR: 0.95	$\rightarrow$ DR: 1.00	$\rightarrow$ DR: 1.00	$\rightarrow$ DR: 1.00
				

Figure A.3: For one correctly classified healthy and one incorrectly classified healthy image, we show for the ensemble the  $l_4$ -VCEs for  $\epsilon \in \{0.1, 0.2, 0.3, 0.4\}$  when transforming into the healthy and DR class, respectively.


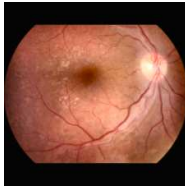
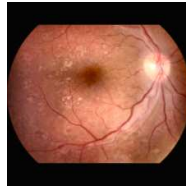
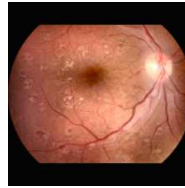
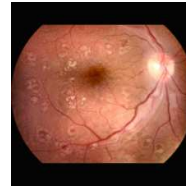
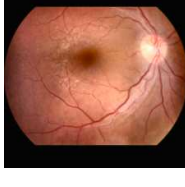
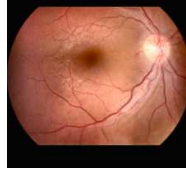
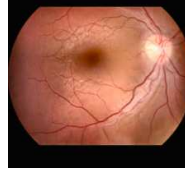
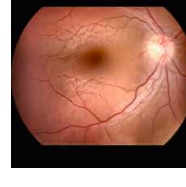
Orig.(GT:DR)	$l_4$ -VCE, $\epsilon=0.1$	$l_4$ -VCE, $\epsilon=0.2$	$l_4$ -VCE, $\epsilon=0.3$	$l_4$ -VCE, $\epsilon=0.4$
healthy:0.69	→DR: 0.77	→DR: 0.98	→DR: 1.00	→DR: 1.00
				
	→healthy: 0.99	→healthy: 1.00	→healthy: 1.00	→healthy: 1.00
				

Figure A.4: For one wrongly classified DR and one incorrectly classified healthy image, we show for the ensemble the  $l_4$ -VCEs for  $\epsilon \in \{0.1, 0.2, 0.3, 0.4\}$  when transforming into the DR and healthy class, respectively.

## Appendix B

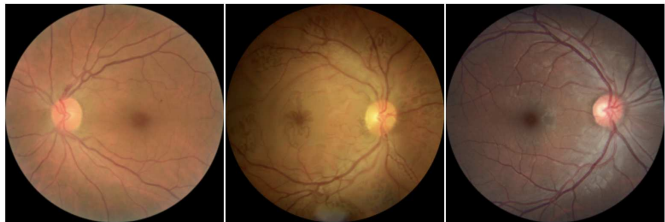
# Supplementary materials to Chapter 6

Retinal Image Tools admin ▾

---

**Question 3/80: Which image was generated by AI?**

a                      b                      c



Select the image generated by AI.

a     b     c

Figure B.1: Web interface for evaluating realism of counterfactuals. Three images are shown on the page where two are real and one is generated. User is asked to select the generated image



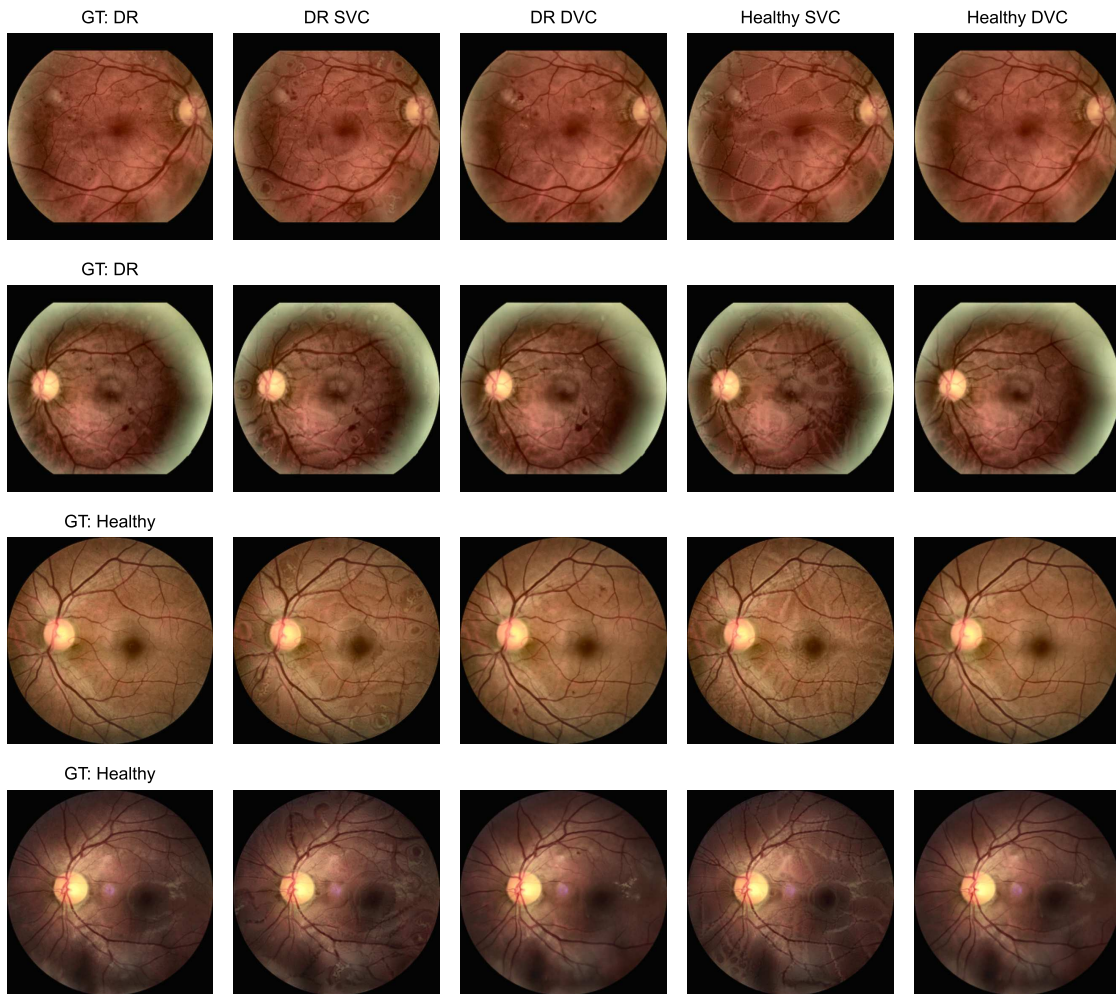


Figure B.2: More examples comparing SVCs and DVCs. Top two rows show counterfactuals from DR fundus images. Bottom two rows show counterfactuals from healthy images. In all cases, changes in DVCs are more realistic compared to SVCs.

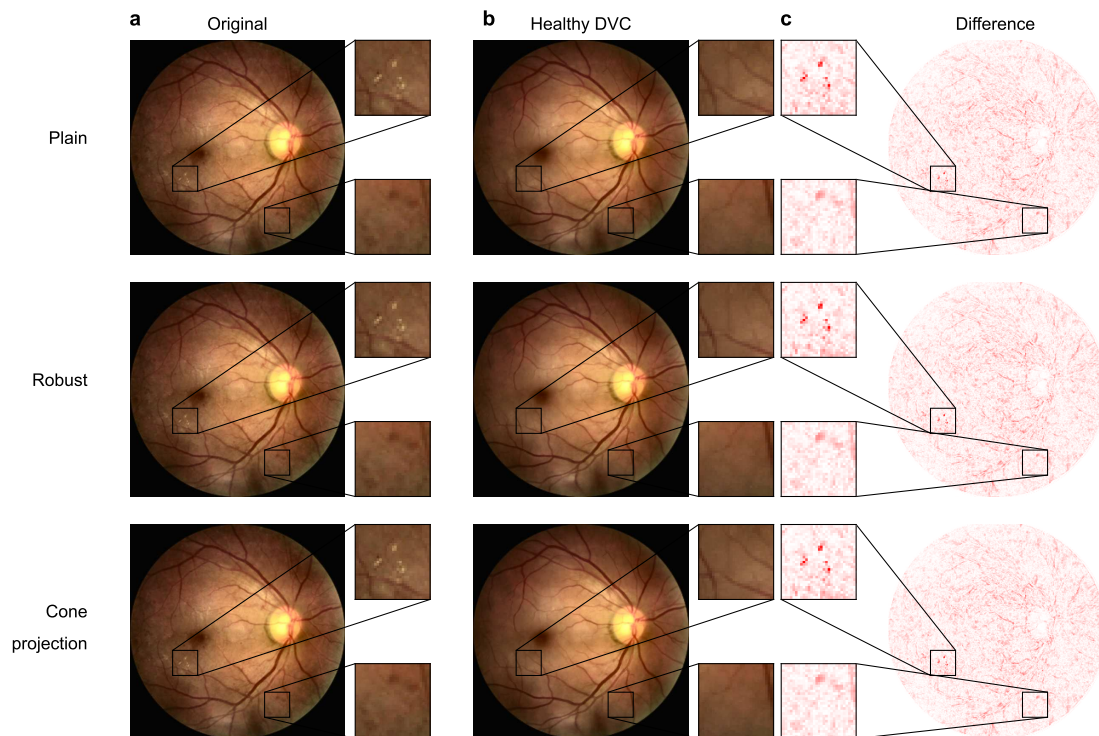


Figure B.3: As in Fig. 6.4 for a DR fundus image. **a**. Original image with ground truth label DR. **b**. DVC to the healthy class. Plain model (top row) removes lesions to a similar extent as robust (middle row) and cone projection (bottom row). DVCs to healthy class are more easily generated than to DR class. **c** Difference maps between the original DR image and generated healthy counterfactual highlighting lesion locations.

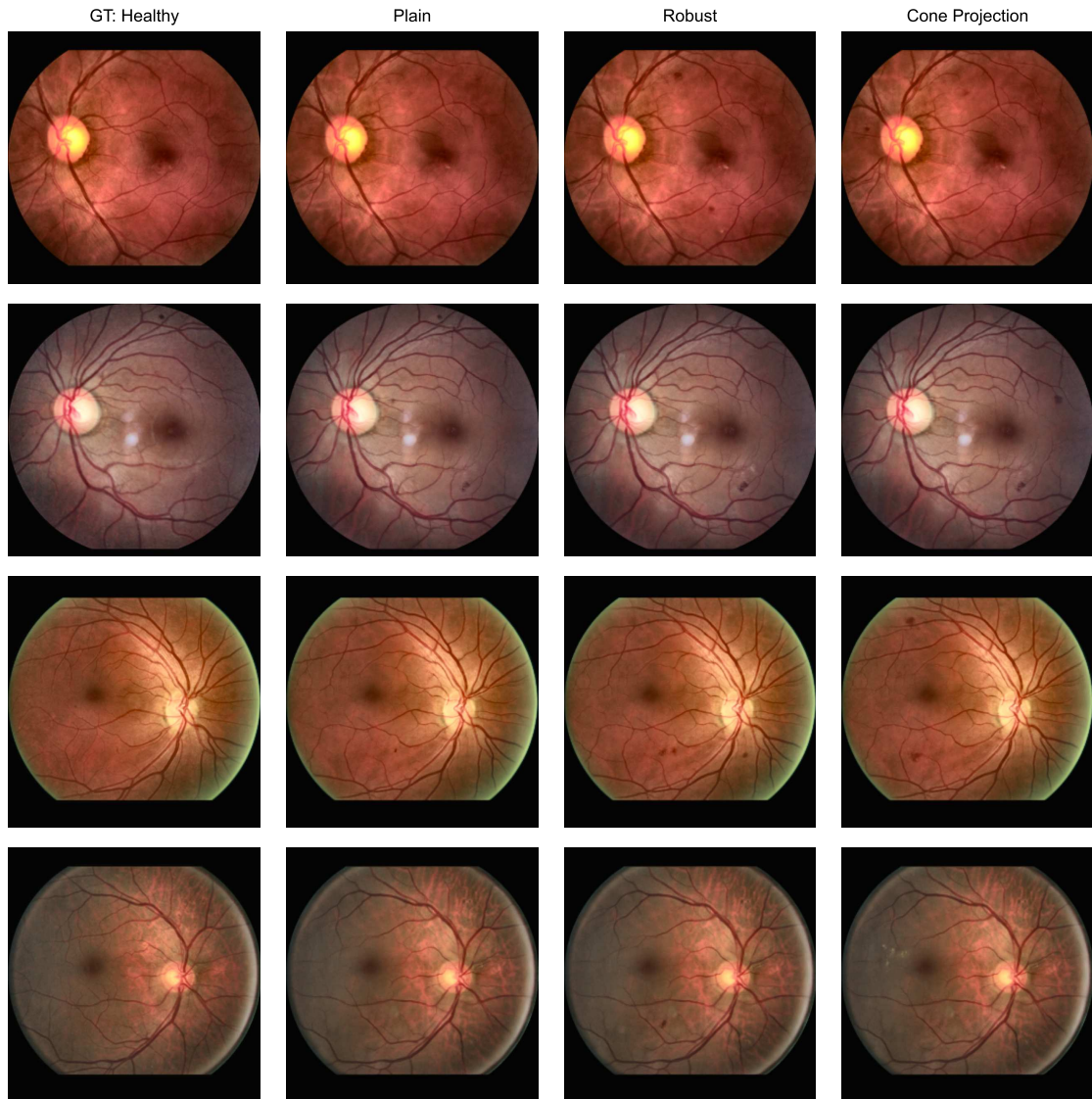


Figure B.4: More examples of DR DVCs generated from healthy fundus images (leftmost column) using the plain model gradients (second column), robust model gradients (third column) and cone projected gradients (rightmost column). In all examples, plain models either show no or fewer and weaker lesions compared to robust and cone projection models.

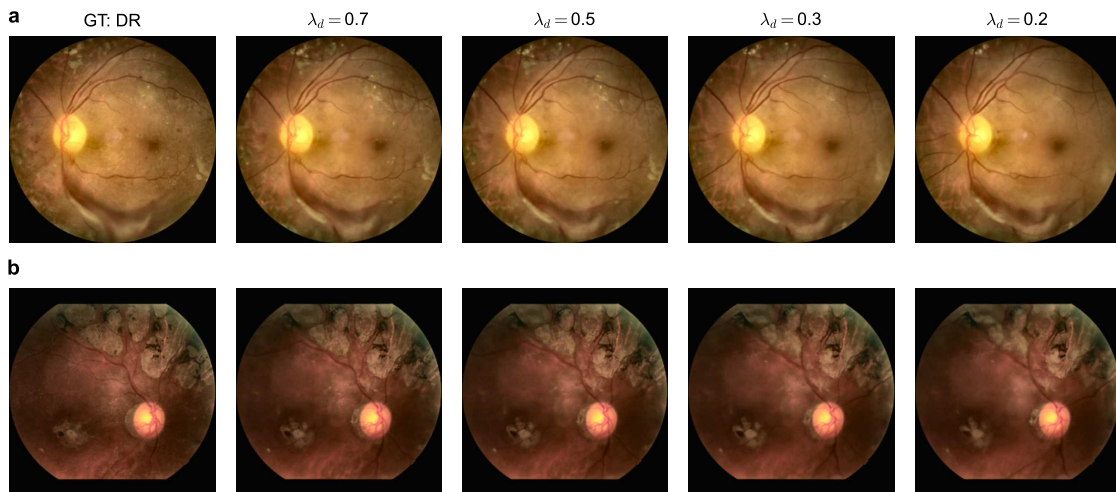


Figure B.5: Effect of regularization strength on retinal fundus images severely affected by DR. In such extreme cases, even a small regularization of 0.2 is not sufficient to convert the image to healthy.

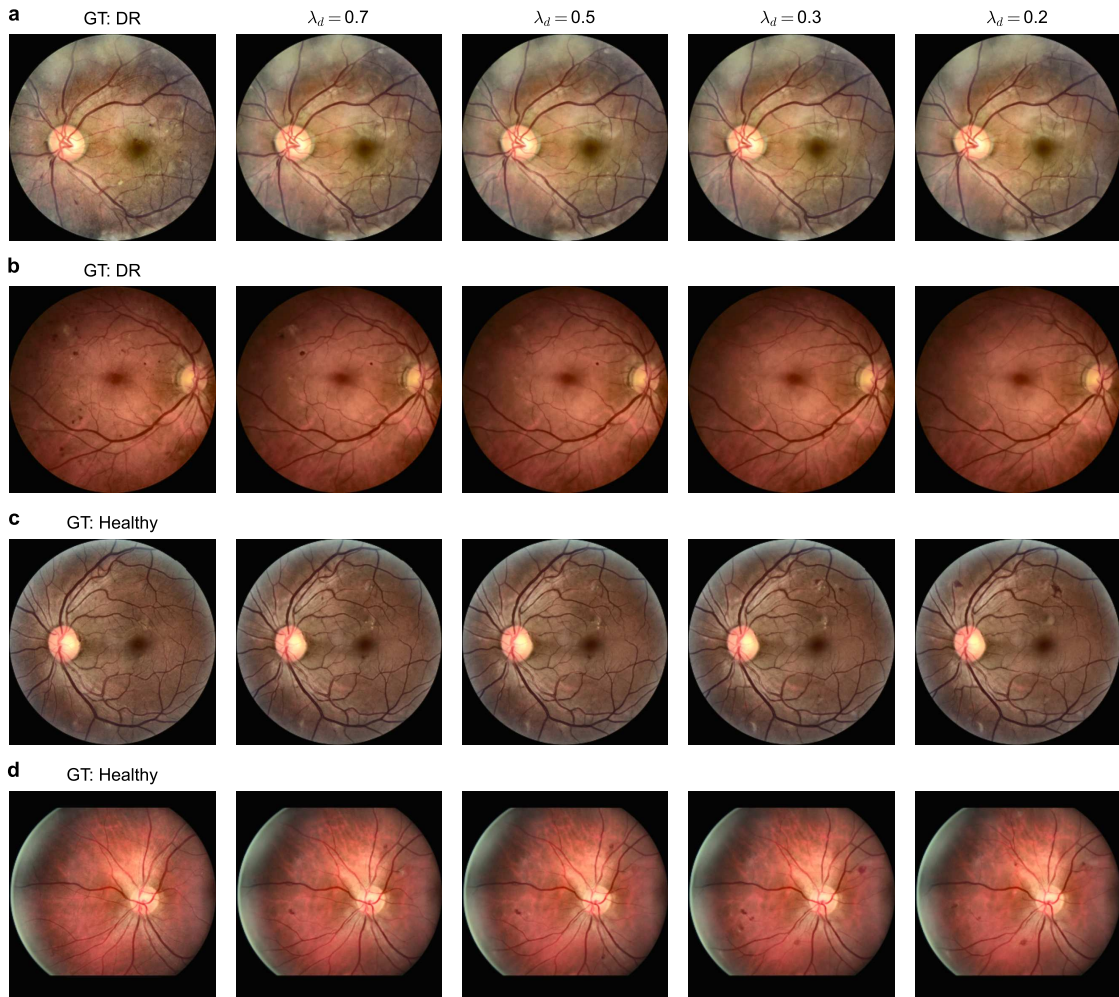


Figure B.6: Effect of regularization strength on retinal fundus images belonging to DR and healthy class. **a-b**. Healthy DVCs of DR fundus images with different values of  $\lambda_d$ . With  $\lambda_d = 0.5$ , examples are converted to the healthy class with either no lesions (**a**) or very few remaining lesions (**b**) such as in the mild DR class. **c-d** DR DVCs of healthy fundus images with varying  $\lambda_d$ . Here too, with  $\lambda_d = 0.5$ , the DVC adds enough lesions to change the decision of the classifier to the DR class with high confidence.

# Bibliography

- [1] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In: *Nature* 529 (2016), pp. 484–503.
- [2] Volodymyr Mnih et al. “Playing Atari with Deep Reinforcement Learning”. In: (2013). URL: <http://arxiv.org/abs/1312.5602>.
- [3] J. Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.
- [4] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [5] Christian Szegedy et al. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.
- [6] Tomáš Mikolov et al. “Recurrent neural network based language model”. In: *Proc. Interspeech 2010*. 2010, pp. 1045–1048. DOI: 10.21437/Interspeech.2010-343.
- [7] Kyunghyun Cho et al. “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179.
- [8] Michael A. Arbib. *The Handbook of Brain Theory and Neural Networks, Second Edition*. The MIT Press, Nov. 2002. ISBN: 9780262267267. DOI: 10.7551/mitpress/3413.001.0001.
- [9] Warren Mcculloch and Walter Pitts. “A Logical Calculus of Ideas Immanent in Nervous Activity”. In: *Bulletin of Mathematical Biophysics* 5 (1943), pp. 127–147.
- [10] F. Rosenblatt. “The perceptron: A probabilistic model for information storage and organization in the brain.” In: *Psychological Review* 65.6 (1958), pp. 386–408. ISSN: 0033-295X. DOI: 10.1037/h0042519.
- [11] M. L. Minsky and S. Papert. *Perceptrons, An Essay on Computational Geometry*. MIT Press, 1969.
- [12] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning Internal Representations by Error Propagation”. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. MIT Press, 1986, pp. 318–362.
- [13] David Beniaguev, Idan Segev, and Michael London. “Single cortical neurons as deep artificial neural networks”. In: *Neuron* 109.17 (2021), 2727–2739.e3. ISSN: 0896-6273.
- [14] Jürgen Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural Networks* 61 (2015), pp. 85–117. ISSN: 0893-6080.
- [15] D. H. Hubel and T. N. Wiesel. “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. In: *The Journal of Physiology* 160.1 (1962), pp. 106–154.
- [16] Kuniyuki Fukushima. “Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position”. In: *Biological Cybernetics* 36 (1980), pp. 193–202.
- [17] Y. LeCun et al. “Backpropagation Applied to Handwritten Zip Code Recognition”. In: *Neural Computation* 1.4 (1989), pp. 541–551. DOI: 10.1162/neco.1989.1.4.541.

- [18] Yann LeCun et al. “Handwritten Digit Recognition with a Back-Propagation Network”. In: *Advances in Neural Information Processing Systems*. Vol. 2. Morgan-Kaufmann, 1989.
- [19] C. Szegedy et al. “Going deeper with convolutions”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.
- [20] Haris Iqbal. *PlotNeuralNet*. Accessed: 2021-02-26. 2018. URL: <https://github.com/HarisIqbal88/PlotNeuralNet>.
- [21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
- [22] Eric J Topol. “High-performance medicine: the convergence of human and artificial intelligence”. In: *Nature medicine* 25.1 (2019), p. 44.
- [23] Michael David Abràmoff et al. “Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning.” In: *Investigative ophthalmology & visual science* 57 13 (2016), pp. 5200–5206.
- [24] Michael Abràmoff and Christine N. Kay. “Chapter 6 - Image Processing”. In: *Retina (Fifth Edition)*. Fifth Edition. London: W.B. Saunders, 2013, pp. 151–176. ISBN: 978-1-4557-0737-9.
- [25] Carlos Alexandre de Amorim Garcia Filho et al. “Chapter 3 - Optical Coherence Tomography”. In: *Retina (Fifth Edition)*. Fifth Edition. London: W.B. Saunders, 2013, pp. 82–110. ISBN: 978-1-4557-0737-9.
- [26] “4.1 - Normal Retinal Anatomy and Basic Pathologic Appearances”. In: *Handbook of Retinal OCT: Optical Coherence Tomography (Second Edition)*. Ed. by Jay S. Duker, Nadia K. Waheed, and Darin R. Goldman. Elsevier, 2022, pp. 24–35. ISBN: 978-0-323-75772-0.
- [27] Huisi Wu et al. “SCS-Net: A Scale and Context Sensitive Network for Retinal Vessel Segmentation”. In: *Medical Image Analysis* 70 (2021), p. 102025. ISSN: 1361-8415.
- [28] Hao Xiong et al. “Weak label based Bayesian U-Net for optic disc segmentation in fundus images”. In: *Artificial Intelligence in Medicine* 126 (2022), p. 102261. ISSN: 0933-3657.
- [29] Varun Gulshan et al. “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs”. In: *Jama* 316.22 (2016), pp. 2402–2410.
- [30] Daniel Shu Wei Ting et al. “Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes”. In: *JAMA* 318.22 (Dec. 2017), pp. 2211–2223. ISSN: 0098-7484. DOI: 10.1001/jama.2017.18152.
- [31] Felix Grassmann et al. “A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography”. In: *Ophthalmology* 125.9 (2018), pp. 1410–1420. ISSN: 0161-6420.
- [32] Ryan Poplin et al. “Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning”. In: *Nature Biomedical Engineering* 2 (2019), pp. 158–164.
- [33] P. Costa et al. “EyeQual: Accurate, Explainable, Retinal Image Quality Assessment”. In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2017, pp. 323–330.
- [34] Jeffrey De Fauw et al. “Clinically applicable deep learning for diagnosis and referral in retinal disease”. In: *Nature medicine* 24.9 (2018), p. 1342.
- [35] Donghuan Lu et al. “Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network”. In: *Medical Image Analysis* 54 (2019), pp. 100–110. ISSN: 1361-8415.
- [36] Cecilia S. Lee et al. “Deep-Learning Based, Automated Segmentation of Macular Edema in Optical Coherence Tomography”. In: *bioRxiv* (2017).
- [37] Daniel S. Kermany et al. “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning”. In: *Cell* 172.5 (2018), 1122–1131.e9. ISSN: 0092-8674.
- [38] Murat Seçkin Ayhan et al. “Interpretable Detection of Epiretinal Membrane from Optical Coherence Tomography with Deep Neural Networks”. In: *medRxiv* (2023). DOI: 10.1101/2022.11.24.22282667.

- [39] Pearse Keane and Eric Topol. “With an eye to AI and autonomous diagnosis”. In: *npj Digital Medicine* 1 (Dec. 2018). DOI: 10.1038/s41746-018-0048-y.
- [40] Kevin Wu et al. “Characterizing the Clinical Adoption of Medical AI Devices through U.S. Insurance Claims”. In: *NEJM AI* 1.1 (2024), AIoa2300030. DOI: 10.1056/AIoa2300030.
- [41] Wieland Brendel and Matthias Bethge. “Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet”. In: *International Conference on Learning Representations*. 2019.
- [42] Thomas Grote and Philipp Berens. “On the ethics of algorithmic decision-making in health-care”. In: *Journal of medical ethics* 46.3 (2020), pp. 205–211.
- [43] Finale Doshi-Velez and Been Kim. “Towards A Rigorous Science of Interpretable Machine Learning”. In: *arXiv: Machine Learning* (2017).
- [44] EU Commission. “Regulation for laying down harmonised rules on AI”. In: *European Commission* (2021). URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021PC0206&from=EN>.
- [45] B. Goodman and S. Flaxman. “European Union regulations on algorithmic decision-making and a “right to explanation””. In: *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. 2016.
- [46] Joseph Y. Halpern and Judea Pearl. “Causes and Explanations: A Structural-Model Approach. Part II: Explanations”. In: *The British Journal for the Philosophy of Science* 56.4 (2005), pp. 889–911. DOI: 10.1093/bjps/axi148.
- [47] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial Intelligence* 267 (2019), pp. 1–38. ISSN: 0004-3702.
- [48] Sandeep Prasada and Elaine M. Dillingham. “Principled and statistical connections in common sense conception”. In: *Cognition* 99.1 (2006), pp. 73–112. ISSN: 0010-0277.
- [49] Ruth MJ Byrne. “Counterfactual thought”. In: *Annual review of psychology* 67 (2016), pp. 135–157.
- [50] Ann McGill and Jill Klein. “Contrastive and counterfactual thinking in causal judgment”. In: *Journal of Personality and Social Psychology* 64 (June 1993), pp. 897–905. DOI: 10.1037/0022-3514.64.6.897.
- [51] Moritz Böhle, Mario Fritz, and Bernt Schiele. “B-Cos Networks: Alignment Is All We Need for Interpretability”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 10329–10338.
- [52] Chaofan Chen et al. “This Looks Like That: Deep Learning for Interpretable Image Recognition”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.
- [53] Amirata Ghorbani et al. “Towards Automatic Concept-based Explanations”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.
- [54] Jost Tobias Springenberg et al. “Striving for simplicity: The all convolutional net”. In: *ICLR (Workshop Track)*. 2014.
- [55] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic Attribution for Deep Networks”. In: *ICML*. 2017.
- [56] Valentyn Boreiko et al. “Sparse Visual Counterfactual Explanations in Image Space”. In: *Pattern Recognition*. Springer International Publishing, 2022, pp. 133–148.
- [57] Maximilian Augustin et al. “Diffusion Visual Counterfactual Explanations”. In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., 2022, pp. 364–377.
- [58] H. Zhang et al. “Theoretically Principled Trade-off between Robustness and Accuracy”. In: *ICML*. 2019.
- [59] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 6840–6851.
- [60] Kevin Jarrett et al. “What is the best multi-stage architecture for object recognition?” In: *2009 IEEE 12th International Conference on Computer Vision*. 2009, pp. 2146–2153.



- [61] Ilya Sutskever et al. “On the importance of initialization and momentum in deep learning”. In: *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, June 2013, pp. 1139–1147.
- [62] Alaa Tharwat. “Classification assessment methods”. In: *Applied Computing and Informatics* (2020).
- [63] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655.
- [64] Margherita Grandini, Enrico Bagli, and Giorgio Visani. “Metrics for Multi-Class Classification: an Overview”. In: *ArXiv abs/2008.05756* (2020).
- [65] Jacob Cohen. “A Coefficient of Agreement for Nominal Scales”. In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46.
- [66] Mary McHugh. “Interrater reliability: The kappa statistic”. In: *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB* 22 (Oct. 2012), pp. 276–82.
- [67] Olga Russakovsky et al. “ImageNet Large Scale Visual Recognition Challenge”. In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252.
- [68] Indu Ilanchezian et al. “Interpretable gender classification from retinal fundus images using BagNets”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 477–487.
- [69] Valentyn Boreiko et al. “Visual Explanations for the Detection of Diabetic Retinopathy from Retinal Fundus Images”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. 2022, pp. 539–549.
- [70] Indu Ilanchezian et al. “Generating Realistic Counterfactuals for Retinal Fundus and OCT Images using Diffusion Models”. In: *arXiv 2311.11629* (2023).
- [71] Prafulla Dhariwal and Alexander Nichol. “Diffusion Models Beat GANs on Image Synthesis”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794.
- [72] Shibani Santurkar et al. “Image Synthesis with a Single (Robust) Classifier”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.
- [73] David Stutz, Matthias Hein, and Bernt Schiele. “Relating Adversarially Robust Generalization to Flat Minima”. In: (2021).
- [74] Francesco Croce and Matthias Hein. “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks”. In: *ICML*. 2020.
- [75] Vasile Moraru. “An Algorithm for Solving Quadratic Programming Problems”. In: *Computer Science Journal of Moldova* (1997).
- [76] Martin Jaggi. “Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization”. In: *ICML*. 2013.
- [77] Francesco Croce and Matthias Hein. “Mind the box:  $l_1$ -APGD for sparse adversarial attacks on image classifiers”. In: *ICML*. 2021.
- [78] Jascha Sohl-Dickstein et al. “Deep Unsupervised Learning using Nonequilibrium Thermodynamics”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. PMLR, 2015, pp. 2256–2265.
- [79] Alexander Quinn Nichol et al. “GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models”. In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 16784–16804.
- [80] Omri Avrahami, Dani Lischinski, and Ohad Fried. “Blended Diffusion for Text-Driven Editing of Natural Images”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 18208–18218.
- [81] Scott Mayer McKinney et al. “International evaluation of an AI system for breast cancer screening”. In: *Nature* 577.7788 (2020), pp. 89–94.

- [82] Andre Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature* 542.7639 (2017), p. 115.
- [83] Amirhossein Kiani et al. “Impact of a deep learning assistant on the histopathologic classification of liver cancer”. In: *npj Digital Medicine* 3.1 (2020), pp. 1–8.
- [84] Simon Dieck et al. “Factors in Color Fundus Photographs That Can Be Used by Humans to Determine Sex of Individuals”. In: *Translational Vision Science & Technology* 9.7 (June 2020), pp. 8–8. ISSN: 2164-2591.
- [85] Kuan-Ming Chueh et al. “Prediction of Sex and Age from Macular Optical Coherence Tomography Images and Feature Analysis Using Deep Learning”. In: *medRxiv* (2020).
- [86] Cathie Sudlow et al. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *Plos med* 12.3 (2015), e1001779.
- [87] Stephen O’Hara and Bruce A Draper. “Introduction to the bag of features paradigm for image classification and retrieval”. In: *arXiv preprint arXiv:1101.3354* (2011).
- [88] Takehiro Yamashita et al. “Factors in Color Fundus Photographs That Can Be Used by Humans to Determine Sex of Individuals”. In: *Translational Vision Science & Technology* 9.2 (Jan. 2020), pp. 4–4. ISSN: 2164-2591.
- [89] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. “Methods for interpreting and understanding deep neural networks”. In: *Digital Signal Processing* 73 (2018), pp. 1–15.
- [90] Murat Seçkin Ayhan et al. “Clinical Validation of Saliency Maps for Understanding Deep Neural Networks in Ophthalmology”. In: *medRxiv* (2021).
- [91] Magdalini Paschali et al. “Deep learning under the microscope: improving the interpretability of medical imaging neural networks”. In: *arXiv preprint arXiv:1904.03127* (2019).
- [92] Francois Chollet et al. *Keras*. 2015. URL: <https://github.com/fchollet/keras>.
- [93] L. V. D. Maaten and Geoffrey E. Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.
- [94] George C. Linderman et al. “Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data”. In: *Nature Methods* 16 (2019), pp. 243–245.
- [95] Dmitry Kobak and Philipp Berens. “The art of using t-SNE for single-cell transcriptomics”. In: *Nature communications* 10.1 (2019), pp. 1–14.
- [96] Dmitry Kobak et al. “Heavy-tailed kernels reveal a finer cluster structure in t-SNE visualisations”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2019, pp. 124–139.
- [97] Dian Li et al. “Sex-specific differences in circumpapillary retinal nerve fiber layer thickness”. In: *Ophthalmology* 127.3 (2020), pp. 357–368.
- [98] François C Delori et al. “Bimodal spatial distribution of macular pigment: evidence of a gender relationship”. In: *JOSA A* 23.3 (2006), pp. 521–538.
- [99] Murat Seçkin Ayhan et al. “Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection”. In: *Medical Image Analysis* (2020), p. 101724.
- [100] Xiaoxuan Liu et al. “A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis”. In: *The lancet digital health* 1.6 (2019), e271–e297.
- [101] Cristina González-Gonzalo et al. “Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice”. In: *Progress in retinal and eye research* (2021), p. 101034.
- [102] Toon Van Craenendonck et al. “Systematic Comparison of Heatmapping Techniques in Deep Learning in the Context of Diabetic Retinopathy Lesion Detection”. In: *Translational Vision Science & Technology* 9.2 (Dec. 2020), pp. 64–64. ISSN: 2164-2591. DOI: 10.1167/tvst.9.2.64.
- [103] Murat Seçkin Ayhan et al. “Clinical validation of saliency maps for understanding deep neural networks in ophthalmology”. In: *Medical Image Analysis* (2022), p. 102364.

- [104] Nishanth Arun et al. “Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging”. In: *Radiology: Artificial Intelligence* 3.6 (2021), e200267.
- [105] R. Geirhos et al. “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”. In: *ICLR*. 2019.
- [106] Robert Geirhos et al. “Shortcut learning in deep neural networks”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.
- [107] Adriel Saporta et al. “Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation”. In: *medRxiv* (2021).
- [108] Aleksander Madry et al. “Towards Deep Learning Models Resistant to Adversarial Attacks”. In: *ICLR*. 2018.
- [109] Christian Etmann et al. “On the connection between adversarial robustness and saliency map interpretability”. In: *ICML*. 2019.
- [110] Andrei Margeloiu et al. “Improving interpretability in medical imaging diagnosis using adversarial training”. In: *arXiv preprint arXiv:2012.01166* (2020).
- [111] Maximilian Augustin, Alexander Meinke, and Matthias Hein. “Adversarial Robustness on In- and Out-Distribution Improves Explainability”. In: *ECCV*. 2020.
- [112] Dimitris Tsipras et al. “Robustness May Be at Odds with Accuracy”. In: *ICLR*. 2019.
- [113] H. Zhang et al. “Theoretically Principled Trade-off between Robustness and Accuracy”. In: *ICML*. 2019.
- [114] *Kaggle competition on diabetic retinopathy detection*. Accessed: 2022-02-02. 2015. URL: <https://www.kaggle.com/c/diabetic-retinopathy-detection/data>.
- [115] Etienne Decencière et al. “Feedback on a publicly distributed database: the Messidor database”. en. In: *Image Analysis & Stereology* 33.3 (Aug. 2014), pp. 231–234. ISSN: 1854-5165. DOI: 10.5566/ias.1155.
- [116] Prasanna Porwal et al. “Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research”. In: *Data* 3.3 (2018), p. 25.
- [117] Karel Zuiderveld. “Contrast limited adaptive histogram equalization”. In: *Graphics gems* (1994), pp. 474–485.
- [118] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *ICML*. 2019.
- [119] Naveed Younis et al. “Incidence of sight-threatening retinopathy in patients with type 2 diabetes in the Liverpool Diabetic Eye Study: a cohort study”. In: *The Lancet* 361.9353 (2003), pp. 195–200.
- [120] Murat Seckin Ayhan et al. “Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection”. In: *Medical Image Analysis* 64 (2020).
- [121] Chuan Guo et al. “On Calibration of Modern Neural Networks”. In: *ICML*. 2017.
- [122] K. He et al. “Deep residual learning for image recognition”. In: *CVPR*. 2016, pp. 770–778.
- [123] Utku Ozbulak. *PyTorch CNN Visualizations*. <https://github.com/utkuozbulak/pytorch-cnn-visualizations>. 2019.
- [124] Cristina González-Gonzalo et al. “Iterative augmentation of visual evidence for weakly-supervised lesion localization in deep interpretability frameworks”. In: *IEEE Transactions on Medical Imaging* (2019).
- [125] Oran Lang et al. “Explaining in Style: Training a GAN to explain a classifier in StyleSpace”. In: *arXiv preprint arXiv:2104.13369* (2021).
- [126] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.
- [127] Mattia Prosperi et al. “Causal inference and counterfactual prediction in machine learning for actionable healthcare”. In: *Nature Machine Intelligence* 2.7 (2020), pp. 369–375.
- [128] Susu Sun et al. “Inherently Interpretable Multi-Label Classification Using Class-Specific Counterfactuals”. In: *Medical Imaging with Deep Learning*. 2023.

- [129] Oran Lang et al. “Explaining in Style: Training a GAN to explain a classifier in StyleSpace”. In: *arXiv preprint arXiv:2104.13369* (2021).
- [130] Joseph Paul Cohen et al. “Gifsplanation via Latent Shift: A Simple Autoencoder Approach to Counterfactual Generation for Chest X-rays”. In: *Medical Imaging with Deep Learning*. 2021.
- [131] Martin J. Menten et al. “Exploring Healthy Retinal Aging with Deep Learning”. In: *Ophthalmology Science* 3.3 (2023), p. 100294. ISSN: 2666-9145.
- [132] Pedro Sanchez et al. “What is Healthy? Generative Counterfactual Diffusion for Lesion Localization”. In: *Deep Generative Models*. Cham: Springer Nature Switzerland, 2022, pp. 34–44.
- [133] Julia Wolleb et al. “Diffusion Models for Medical Anomaly Detection”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Cham: Springer Nature Switzerland, 2022, pp. 35–45.
- [134] Pedro Sanchez and Sotirios A. Tsaftaris. “Diffusion Causal Models for Counterfactual Estimation”. In: *First Conference on Causal Learning and Reasoning*. 2022.
- [135] Veronica Elisa Castillo Benítez et al. “Dataset from fundus images for the study of diabetic retinopathy”. In: *Data in Brief* 36 (2021), p. 107068. ISSN: 2352-3409.
- [136] Y. Zhou et al. “A Benchmark for Studying Diabetic Retinopathy: Segmentation, Grading, and Transferability”. In: *IEEE Transactions on Medical Imaging* 40.3 (2021), pp. 818–828. DOI: 10.1109/TMI.2020.3037771.
- [137] Sarah Mueller et al. *fundus circle cropping*. Version 0.1.0. 2023. DOI: 10.5281/zenodo.10137935. URL: [https://github.com/berenslab/fundus\\_circle\\_cropping](https://github.com/berenslab/fundus_circle_cropping).
- [138] Daniel Kermany, Kang Zhang, Michael Goldbaum, et al. “Labeled optical coherence tomography (oct) and chest x-ray images for classification”. In: *Mendeley data* 2.2 (2018), p. 651.
- [139] Logan Engstrom et al. *Robustness (Python Library)*. 2019. URL: <https://github.com/MadryLab/robustness>.
- [140] Felix A. Wichmann et al. “Methods and measurements to compare men against machines”. In: *Electronic Imaging* 29.14 (), pp. 36–36.
- [141] Guillermo Aguilar, Felix A. Wichmann, and Marianne Maertens. “Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment”. In: *Journal of Vision* 17.1 (Jan. 2017), pp. 37–37. ISSN: 1534-7362. DOI: 10.1167/17.1.37.
- [142] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *International Conference on Learning Representations*. 2014.
- [143] Murat Seçkin Ayhan et al. “Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection”. In: *Medical image analysis* 64 (2020), p. 101724.
- [144] Robert Geirhos et al. “Shortcut learning in deep neural networks”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.
- [145] Yukun Zhou et al. “A foundation model for generalizable disease detection from retinal images”. In: *Nature* (2023), pp. 1–8.
- [146] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. “Diffusion Models for Counterfactual Explanations”. In: *Computer Vision – ACCV 2022*. Cham: Springer Nature Switzerland, 2023, pp. 219–237.
- [147] Finn Behrendt et al. “Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI”. In: *Medical Imaging with Deep Learning*. 2023.
- [148] Yu Gu et al. *BiomedJourney: Counterfactual Biomedical Image Generation by Instruction-Learning from Multimodal Patient Journeys*. 2023. arXiv: 2310.10765 [cs.CV].
- [149] Jiarong Ye et al. “Synthetic Augmentation with Large-Scale Unconditional Pre-training”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Cham: Springer Nature Switzerland, 2023, pp. 754–764.

- [150] Luke William Sagers et al. “Improving dermatology classifiers across populations using images generated by large diffusion models”. In: *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*. 2022.
- [151] Murat Seçkin Ayhan et al. “Multitask Learning for Activity Detection in Neovascular Age-Related Macular Degeneration”. In: *Translational Vision Science & Technology* 12.4 (Apr. 2023), pp. 12–12. ISSN: 2164-2591. DOI: 10.1167/tvst.12.4.12.
- [152] Sunnie S. Y. Kim et al. “”Help Me Help the AI”: Understanding How Explainability Can Support Human-AI Interaction”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. New York, NY, USA: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581001.
- [153] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *International Journal of Computer Vision* 128.2 (2019), pp. 336–359.
- [154] Julius Adebayo et al. “Sanity Checks for Saliency Maps”. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018.
- [155] Kerol R. Djoumessi Donteu et al. “Sparse Activations for Interpretable Disease Grading”. In: *Medical Imaging with Deep Learning*. 2023.
- [156] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.
- [157] Tom Brown et al. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.
- [158] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 8748–8763.
- [159] Yukun Zhou et al. “A foundation model for generalizable disease detection from retinal images”. In: *Nature* 622.7981 (2023), pp. 156–163.
- [160] K. Singhal et al. “Towards Expert-Level Medical Question Answering with Large Language Models”. In: *ArXiv* abs/2305.09617 (2023).