Explainable AI for retinal OCT diagnosis

by

Amitojdeep Singh Brar

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Vision Science & Systems Design Engineering

Waterloo, Ontario, Canada, 2021

Author Declaration

I hereby declare that this thesis consists of materials all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following publications have resulted from the work presented in this thesis:

- A Singh, S Sengupta, J J Balaji, V Jayakumar, M A Rasheed, J S Zelek, and V Lakshminarayanan, "What is the optimal attribution method for explainable ophthalmic disease classification?" In International Workshop on Ophthalmic Medical Image Analysis, Springer, 2020, Presented virtually at 7th OMIA workshop, MICCAI 2020 in Lima, Peru. **(Best Paper Winner)**
- A Singh, A R Mohammed, J S Zelek, and V Lakshminarayanan. "Interpretation of deep learning using attributions: application to ophthalmic diagnosis." Proc. SPIE 11511, Applications of Machine Learning, 2020, Presented virtually at SPIE Optics and Photonics in San Diego, USA.
- A Singh, S Sengupta, and V Lakshminarayanan. "Explainable deep learning models in medical image analysis." Journal of Imaging, vol. 6, no. 6, p. 52, 2020.
- H Leopold, A Singh, S Sengupta, J S Zelek, and V Lakshminarayanan, "Deep Learning on Optical Coherence Tomography for Ophthalmology", State-of-the-Art in Neural Networks, Vol.1, A. El-Baz, and J. Suri, Eds, Elsevier, NY (in press, 2021).

Abstract

Artificial intelligence methods such as deep learning are leading to great progress in complex tasks that are usually associated with human intelligence and experience. Deep learning models have matched if not bettered human performance for medical diagnosis tasks including retinal diagnosis. Given a sufficient amount of data and computational resources, these models can perform classification and segmentation as well as related tasks such as image quality improvement. The adoption of these systems in actual healthcare centers has been limited due to the lack of reasoning behind their decisions. This black box nature along with upcoming regulations for transparency and privacy exacerbates the ethico-legal challenges faced by deep learning systems.

The attribution methods are a way to explain the decisions of a deep learning model by generating a heatmap of the features which have the most contribution to the model's decision. These are generally compared in quantitative terms for standard machine learning datasets. However, the ability of these methods to generalize to specific data distributions such as retinal OCT has not been thoroughly evaluated. In this thesis, multiple attribution methods to explain the decisions of deep learning models for retinal diagnosis are compared. It is evaluated if the methods considered the best for explainability outperform the methods with a relatively simpler theoretical background.

A review of current deep learning models for retinal diagnosis and the state-of-the-art explainability methods for medical diagnosis is provided. A commonly used deep learning model is trained on a large public dataset of OCT images and the attributions are generated using various methods. A quantitative and qualitative comparison of these approaches is done using several performance metrics and a large panel of experienced retina specialists.

The initial quantitative metrics include the runtime of the method, RMSE, and Spearman's rank correlation for a single instance of the model. Later, two stronger metrics - robustness and sensitivity are presented. These evaluate the consistency amongst different instances of the same model and the ability to highlight the features with the most effect on the model output respectively. Similarly, the initial qualitative analysis involves the comparison between the heatmaps and a clinician's markings in terms of cosine similarity. Next, a panel of 14 clinicians rated the heatmaps of each method. Their subjective feedback, reasons for preference, and general feedback about using such a system are also documented.

It is concluded that the explainability methods can make the decision process of deep learning models more transparent and the choice of the method should account for the preference of the domain experts. There is a high degree of acceptance from the clinicians

surveyed for using such systems. The future directions regarding system improvements and enhancements are also discussed.

Acknowledgements

I am most grateful and appreciative of my supervisors Dr. Vasudevan Lakshminarayanan and Dr. John Zelek for this invaluable opportunity to learn and explore a field of study that continuously intrigues me, and for their unwavering support, encouragement and mentorship. I am thankful to both for giving me freedom in research while helping me channelize my efforts to achieve the goals. Dr. Lakshminarayanan has been a continuous source of inspiration and guidance to me throughout this journey and has helped in my overall professional development. He has supported me throughout my program despite all the difficulties imposed by the challenging situation. Dr. Zelek provided valuable ideas and helped to overcome the challenges posed by this new domain of study. The active and complete support of my supervisors was indispensable for undertaking and completing this research. I am especially thankful to my committee members Dr. Ben Thompson and Dr. Kaamran Raahemifar for their valued knowledge, advice, and time invested in guiding me throughout my degree. I am also grateful to my collaborators from the TEEL Lab - Sourya, Abdul, Ibrahim, Henry, and other former and current lab members for their efforts in my research. I would like to express my heartfelt gratitude to my external collaborators - Varadharajan, Jothi, Dr. Rajiv Raman, and his team of ophthalmologists from Sankara Nethralaya, Chennai for sharing their expertise and time. I would also like to thank Compute Canada and Sharcnet for providing computing support. This research was supported by an NSERC discovery grant and a TITAN V GPU grant from NVIDIA to Dr. Lakshminarayanan. I am grateful to my beloved parents and grandparents for their constant encouragement and leaving no stone unturned to help me accomplish my life's goals.

Dedication

To my loving grandmother.

Table of Contents

List of Tables

List of Figures

Acronyms

AI artificial intelligence

AMD age related macular degeneration

BM bruch membrane

CAD computer-aided diagnostic

CNN convolutional neural network

CNV choroidal neovascularization

CT computerized tomography

DeepLIFT Deep Learning Important FeaTures

DME diabetic macular edema

DNN deep neural networks

DR diabetic retinopathy

EG expressive gradients

EHR electronic healthcare record

ELM external limiting membrane

FCN fully convolutional neural network

FD-OCT frequency domain OCT

GAN generative adversarial network

GBP guided backpropagation

GCL ganglion cell layer

GPU graphics processing units

GradCAM gradient weighted class activation mapping

GRU gated recurrent unit

HITL human-in-the-loop

IG integrated gradients

INL inner plexiform layer

IS inner segment

IZ interdigitation zone

LRP layer wise relevance propagation

MA microaneurysms

MLP multilayer perceptron

MRI magnetic resonance imaging

OCT optical coherence tomography

ONL outer nuclear layer

OPL outer plexiform layer

OS outer segment

PCC Pearson's correlation coefficient

PR precision

PS-OCT polarization sensitive OCT

PSNR peak signal to noise ratio

ReLU rectified linear unit

RMSE root mean squared error

RNFL retinal nerve fibre layer

RNN recurrent neural network

RPE retinal pigment epithelium

SD-OCT spectral domain OCT

SHAP SHapley Additive exPlanations

SLO scanning laser ophthalmoscopy

SNR signal to noise ratio

SS-OCT swept source OCT

SSIM structural similarity index measurement

SVM support vector machines

TD-OCT time domain OCT

TL transfer learning

XAI explainable AI

Chapter 1

Deep learning for retinal OCT and the need for explainablity

Based on:

- **A Singh, S Sengupta, and V Lakshminarayanan. "Explainable deep learning models in medical image analysis." Journal of Imaging, vol. 6, no. 6, p. 52, 2020.**
- **H Leopold, A Singh, S Sengupta, J S Zelek, and V Lakshminarayanan, "Deep Learning on Optical Coherence Tomography for Ophthalmology", State-of-the-Art in Neural Networks, Vol.1, A. El-Baz, and J.Suri, Eds, Elsevier, NY (in press, 2021).**

1.1 Introduction

The major contributions of this thesis are summarized in the next subsection followed by the organization of the thesis. The rest of this chapter describes the basic concepts of retinal optical coherence tomography (OCT) imaging, deep learning applications, and the process of explaining the decisions made by a deep learning model in the context of medical imaging.

1.1.1 Major contributions

Retinal image diagnosis has undergone a revolution by the advances in artificial intelligence (AI) methods such as deep learning. These methods have not percolated to the patient care systems due to a lack of reasoning behind their decisions. Attribution based methods are available in the literature to explain these decisions in multiple ways. There is a need to inspect the performance of the explainability methods initially validated on standard computer vision data sets in the context of the OCT data. The major contributions of this thesis are:

- Provide summaries of the deep learning methods for retinal OCT diagnosis and the emerging applications of explainability in medical imaging.
- Train and evaluate a deep learning model for identifying retinal diseases from a large public dataset.
- Implement multiple explainability methods to generate the explanations in the form of heatmaps of the input images.
- Perform a quantitative comparison of various explainability methods to measure their ability to identify the primary features that influence the model decision.
- Perform a qualitative evaluation of these methods in terms of clinical relevance using ratings from clinicians.
- Identify explainability methods fit for retinal diagnosis and indicate the directions of future research.

1.1.2 Organization of the thesis

The organization of this thesis follows the sequence of listed contributions. This chapter serves as the background for discussing the applications of explainability approaches to retinal OCT diagnosis. Section 1.2 provides a summary of deep learning methods for OCT diagnosis. Detailed reviews for OCT analysis [1] and fundus diagnosis [2] are available elsewhere. Section 1.3 introduces the need for explaining the deep learning model used for medical diagnosis. A general categorization of explainability methods and the fundamentals needed for subsequent chapters are discussed. An overview of the applications in various medical imaging domains showing the diverse approaches is also presented with a summary table. A more expansive version of the section is available in the literature [3].

The chapter 2 demonstrates the process of training and evaluating a deep learning model for diagnosing multiple diseases along with the generation of explanations. Chapter 3 uses quantitative metrics such as root mean squared error (RMSE) and sensitivity to compare the explanations from different methods. A study involving the clinicians' ratings of the explanations for their ability to highlight pathologies is presented in chapter 4. Chapter 5 concludes with discussions of future research directions including an ongoing work on uncertainty. The relevant codes are available at https://github.com/amitojdeep.

1.2 Deep learning for retinal OCT diagnosis

Ophthalmology is a branch of medicine and surgery that deals with the anatomy, physiology and diseases of the eye as well as the visual process [4]. The landmark treatise on vision can be attributed to the Arab scholar Ibn Al-Haytham who in his magnum opus, Kitab al Manzir (The Book of Optics) laid down the foundations of vision science [5]. The invention of direct Ophthalmoscope by Hermann von Helmholtz in 1861 revolutionized our understanding of retina [6]. The ophthalmoscope and its modern successors, including the retinal fundus camera and the OCT are indispensable tools for ophthalmic examination. The development, commercialization and the impact of the OCT is well documented in a recent article by one of the inventors of the OCT [7]. This chapter mainly deals with analysis of OCT images of retinal diseases (e.g., glaucoma, diabetic retinopathy (DR), age related macular degeneration (AMD). An analysis of retinal fundus photographs can be found in a recent review article [2] .
In general, patient care system is immensely non-uniform in various parts of the world. Sometimes it is over-burdened due to high demand and paucity of adequate number of trained clinicians. These factors increase the risk of diagnostic errors and degrade the

health-care quality and efficacy. It is estimated that by 2020 the number of glaucoma patients may reach almost 80 million worldwide [8], the DR patient percentage will increase upto 4.4% by 2030 [9], and AMD prevalence is about 12% in people over the age of 80 in the United States [2].
Diagnosis with OCT images is sometimes difficult due to the various factors e.g. presence of noise, variations in camera calibration, aperture, contrast setting, training of the clinicians and ethnicity of patients. Enhancing the efficacy of computer-aided diagnostic (CAD) technique can be a way to mitigate the need of trained clinicians which is unavailable in many places in a cost and time optimal manner.
Prior to the advent of deep learning, traditional image processing and pattern recognition based methods have been used to make CAD systems [10]. But, these traditional methods are too benign to extract useful and distinguishable insights from high-dimensional, complex, unstructured medical data. Also these methods require feature extraction to obtain important information. It is very hard to construct generalized robust automated systems except for some highly specific problems. These traditional methods typically involve steps like image pre-processing, feature extraction and application of traditional classifiers to predict an outcome.
Deep learning based models are powerful architectures to automatically find important patterns or feature maps from different high dimensional data. Unlike traditional methods, without any manual feature extraction it derives necessary representations and provides an efficient paradigm to build an automated end-to-end model to predict and distinguish different tasks. With the advent of graphics processing units (GPU), deep learning methods have become much easier to implement offering considerable savings in computational time when compared to ordinary processors. In various fields like computer vision, natural language processing deep learning models have started to outperform traditional machine learning based models and retinal diagnosis is not an exception. In this section the main focus is on neural networks based CAD systems for retinal disease diagnosis using OCT images.

1.2.1 Optical Coherence Tomography

This section briefly describes OCT, its various types and key advantages. The OCT is a clinical imaging technique to visualize the cross-sectional structure of retina. The basic physics of the OCT is given in detail in the articles.

1.2.1.1 Overview

The OCT is a clinical imaging technique to visualize the cross-sectional structure of retina. OCT uses low-coherence light to capture 2D and 3D images of scattering media. It is based on the interferometric technique invented by Albert A. Michelson and was developed by James Fujimoto. OCT is used to non-invasively image the retinal layers for diagnosis of pathologies like glaucoma, AMD and DR. A review of OCT applications in ophthalmology is given in [11], and general information on OCT can be found in [12]. Figure 1.1 is an example slice from a retinal OCT with the various retinal layers, including retinal nerve fibre layer (RNFL), ganglion cell layer (GCL), inner plexiform layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), external limiting membrane (ELM), inner segment (IS), outer segment (OS), retinal pigment epithelium (RPE), interdigitation zone (IZ), bruch membrane (BM), which will be discussed further in Section 1.2.2 [12]. Figure 1.3 shows OCT cross-sections for common variants discussed in Section 1.2.1.2. Other modalities for retinal imaging include fundus photography [13] and scanning laser ophthalmoscopy (SLO) [14].

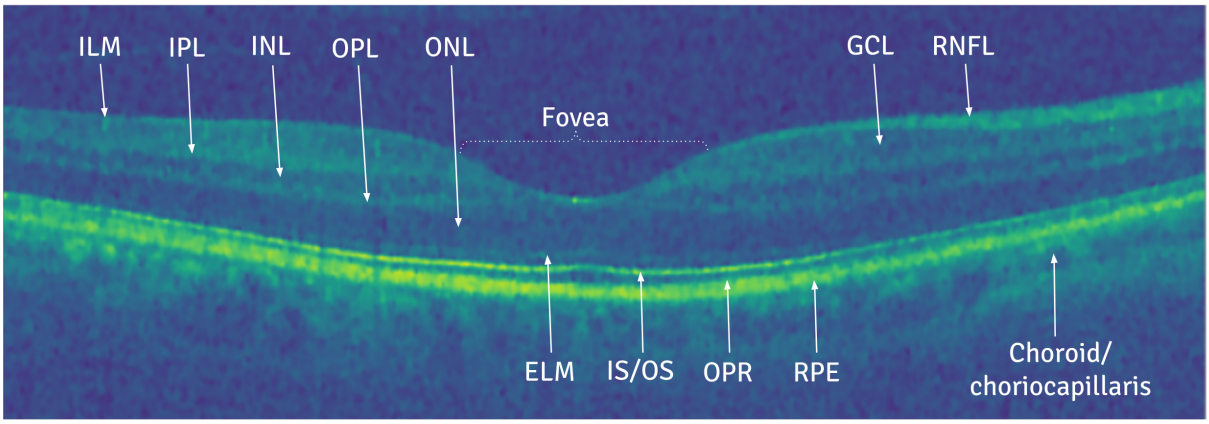


Figure 1.1: OCT image including retinal anatomy [10].

1.2.1.2 Variants of OCT systems

The 4 common variants of OCT used for retinal diagnosis are discussed here.

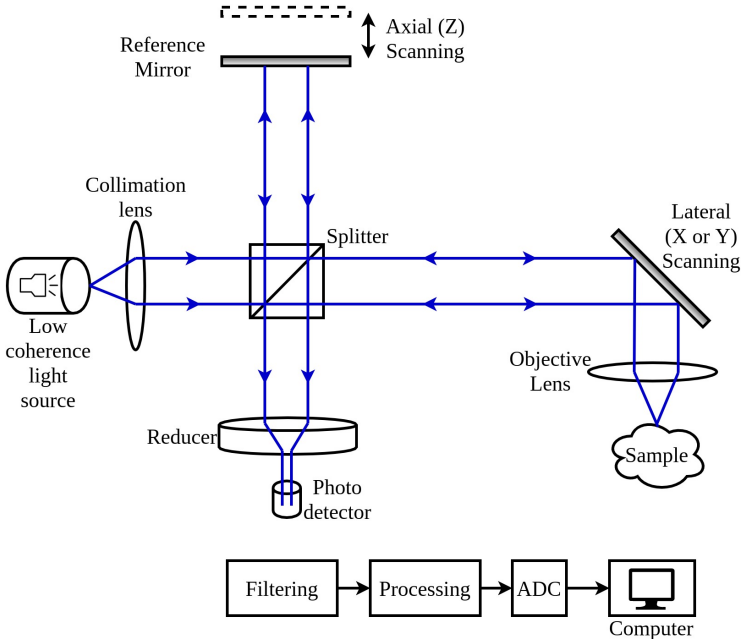


Figure 1.2: Schematic of a single point OCT setup. Adapted from [15]

Time domain OCT (TD-OCT) is the traditional and earliest OCT imaging technique [16]. The reference arm pathlength is varied in time. Interference happens if the path difference is within the range of the light coherence. Typically an OCT setup consists of a low coherence broad bandwidth interferometer. In figure 1.2 a typical schematic of the OCT setup is shown, where light is coming from a monochromatic light source and it is split into two arms, a sample arm and a reference arm by a beam splitter. By translating the reference arm longitudinally, the path length of the reference arm is varied and it results into a series of bright and dark fringes due to interference of light waves reflected from various layers.

Spectral domain OCT (SD-OCT) has similar basic setup to the TD-OCT. The main difference between TD-OCT and SD-OCT is the reference arm length is fixed. It does not obtain the depth information of the sample by scanning reference arm, rather a fourier spectrometer is used to analyze the output light. SD-OCT is also known as frequency domain OCT (FD-OCT). The main advantages of SD-OCT are the higher resolution and the faster speed of image acquisition. SD-OCT offers a more detailed 3D map, facilitating better visualization of inter-retinal layers and a higher possibility of multiple retinal layers' segmentation. It has been shown that the sensitivity of SD-OCT technique is 20dB more than traditional TD-OCT [17], though sometimes in practice the presence of artifacts deter

the efficacy of segmentation.

Polarization sensitive OCT (PS-OCT) enables contrast specific examination of retinal layers. The major drawback of the conventional OCT system is it does not provide tissue-specific contrast. It uses the fact that several materials and tissues can change the polarization state of light providing additional contrast and information. PS-OCT finds applications in a wide range of applications including imaging eyes, muscles, teeth, cancerous tissues, nerves, blood vessels etc due. Detailed information about the PS-OCT method and its applications can be found in [18], [19].

Swept source OCT (SS-OCT) was first used clinically in 2012 and it provides deeper penetration and faster acquisition time than other approaches. This is used to visualize vitreous, choroid and other retinal structures that are covered by dense preretinal hemorrhages. It has been pivotal in the study of the posterior precortical vitreous pocket. SS-OCT devices are not widely available due to the higher costs relative to more popular SD-OCT. The impact of SS-OCT on in-vivo ophthalmic studies is reviewed in [20]. In figure 1.3 different OCT scans from these variants are shown.

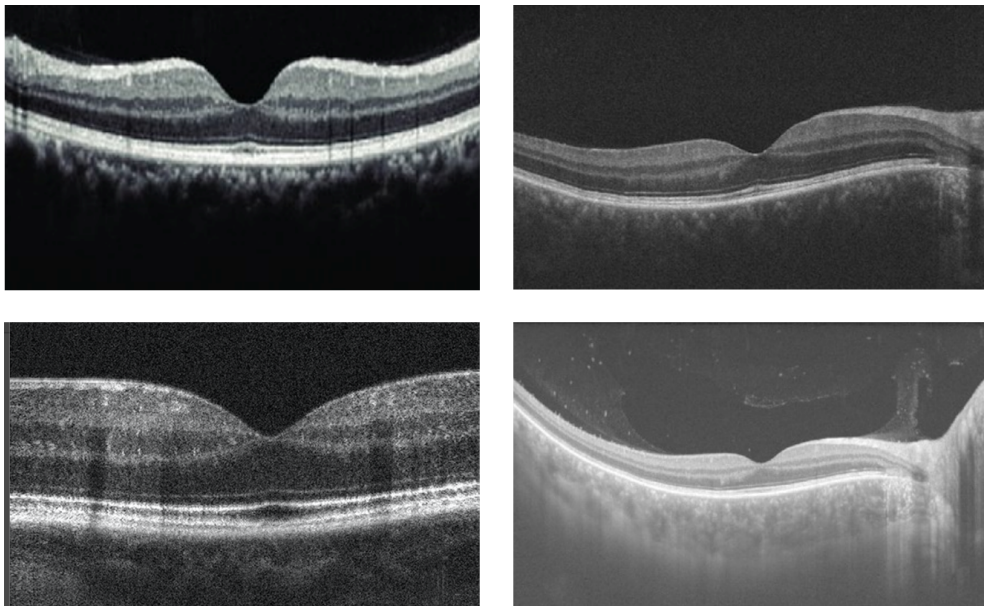


Figure 1.3: Different types of OCT images for normal eye (clockwise from top left): TD-OCT [21], SD-OCT [20], PS-OCT [22] (more contrast) and SS-OCT [20] (more depth)

1.2.2 Retinal Diseases

This subsection briefly describes some important retinal diseases and the associated anatomical and physiological changes.

1.2.2.1 Diabetic Retinopathy and Diabetic Macular Edema

DR is one of the leading causes of blindness. This is a vascular disease of retina, patients with diabetes melitus get affected with this disease. People over the age of 30 years are more prone to this disease [9]. It is suspected that the diabetic patients worldwide will increase from 2.8% in 2000 to 4.4% in 2030. DR is characterized by various abnormalities in retina such as microaneurysms (MA) and other small lesions. DR, a diabetes complication on eye, is majorly caused by the rupture of thin light-sensitive retinal capillaries [10].

Diabetic macular edema (DME) occurs in persons having DR and it involves fluid accumulation in the macula. Since the macula is the region of fine vision, DME greatly impacts vision.

1.2.2.2 Glaucoma

Glaucoma is another major cause of blindness; it is estimated that by 2020, 80 million people will be affected by glaucoma [8]. *Open-angle* glaucoma and *angle closure* glaucoma are the two main types of glaucoma. About 90% of the affected people suffer from primary open-angle glaucoma [23]. Glaucoma is caused due to rise in intra-ocular pressure leading to damage of retinal nerve fibres. This is the pressure of fluid inside the eye and can be measured using a tonometer. Blockage of drainage canals is one of the causes of rise in of this pressure.

1.2.2.3 Age-related Macular Degeneration

AMD is another common retinal disease. It causes the loss of vision in the middle of the visual field. With time there is a high chance of complete loss of central vision [24]. It is reported that in the United States, about 0.4% people from age range 50 to 60 suffer from this disease and almost 12% people above 80 years are affected by this disease. [25].

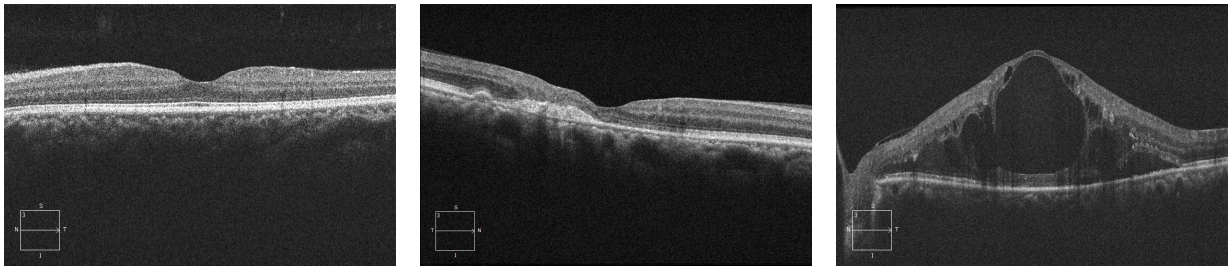


Figure 1.4: Sample OCT images from OCTID dataset [26]. Normal, DR, AMD (L to R)

1.2.3 Deep learning approaches to OCT analysis

In this section applications of deep learning approaches for biomarker detection and ophthalmic disease classification from OCT images are discussed. Further details of neural networks can be found in [27]. It should be emphasized that the segmentation task is essentially a biomarker detection task from pixel level annotations of images.It is notable that much of the work in this area is on using convolutional neural network (CNN)s for classification, whereas generative adversarial network (GAN) and similar architectures are used for more complex tasks such as super-resolution and noise reduction.

1.2.3.1 Convolutional neural network (CNN) applications

CNNs are very common deep learning paradigms. Its applications span computer vision, natural language processing [28], financial forecasting, signal processing, and many more domains. These use the convolution operation instead of simple matrix multiplication in at least one of the layers.

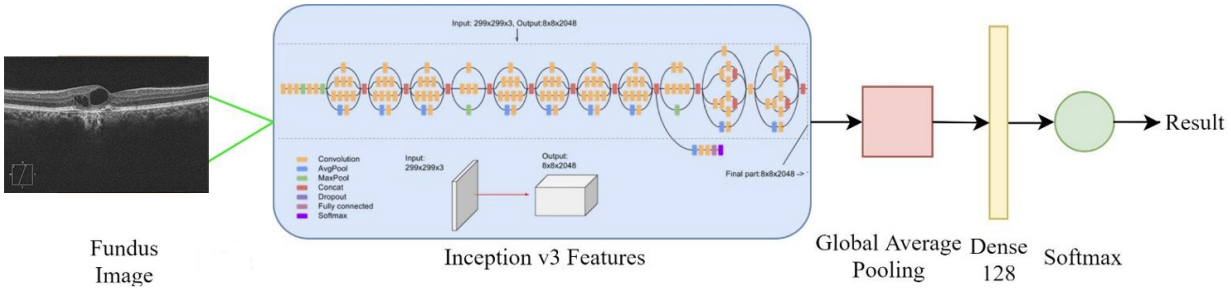


Figure 1.5: An example of Inception-v3 CNN [29] used for classification. Adapted from [30]

The structure of a typical CNN is shown in figure 1.5. The data (an OCT image in this case) is provided to the input layer of the network. A certain number of kernels (feature filters) are convoluted with the input and an activation function is applied to produce the activation matrices corresponding to each kernel. These kernels are usually square matrices in the case of a 2D CNN with size in the range of 3x3 to 7x7 in most applications. A smaller kernel is believed to extract minute features better while larger kernel provides an overall view. The activation functions are designed to replicate the activation of a neuron and some of the common variations are the rectified linear unit (ReLU) [31], softmax, tanh, etc. The process of convolution is repeated at each convolutional layer on the outputs of the preceding layer. The final step is typically a dense layer consisting of a multi-layer perceptron where each neuron is connected to every neuron of the preceding convolutional layer. The output is generated in terms of probabilities of each class for classification problems like the detection of pathology. The entire network is trained using backpropagation [32]. CNN has been extensively applied to retinal diagnostics and the major variations are described in table 1.1.

Architecture	**Application**	**Dataset**	**Performance**	**Others**
AlexNet TL + SVM	DME detection	SERI: 32 3D OCT	Acc: 98.6%, SN: 99.3%, SP: 98.4%	
AlexNet + RF	Glaucoma detection [33]	Private: 102 OCT	Acc:93.1%	
VGG16 TL on 2 scales + patient data with SVM	Glaucoma detection [34]	Private: 8270 OCT	Acc:89.3%, SN: 88.9%, SP: 89.6%, AUC: 0.9456	
VGG based	AMD, DR, DME [35]	A2A, SERI, CUHK	AUC: AMD: 0.99, DR & DME:0.86	Added feature fusion [36]
ResNet + RF	Layer segmentation [37]	DUKE	F1:0.885	
Inception-v3: transfer learning (TL)	AMD detection [38]	200k+ OCT	Acc:96.6%, SN:97.8%, SP:97.4%, AUC:0.999%	[39], [40]
Inception-v3	DME [41]	Private	Acc:85%, SN:80%, SP:89%	

RelayNet TL for seg + layer guided CNN	DME, CNV, DRUSEN [42]	2 datasets: 93k images	Acc:89.9+-0.6%
3 CNN subnets for different scales	Glaucoma [43]	2 private datasets: 9k images	F1: 0.677 & 0.814
Patch based CNN	DME [44]	Private:328	F1:0.926
Custom CNN (OCT-NET)	DME [45]	SERI: 32 SD-OCT volumes	Acc:93.8%
MGRF + deep fusion CNN and autoencoders	DR detection [46]	Private:74	Acc:93%, SN:91%, SP:97%
CNN + Graph search	Layer segmentation [47]	Private: 117 SD OCT	ME:1.26, SD:1.24
TL + PCA with majority voting	DME [48]	SERI	Acc:93.8%
Ensemble of CNN	DME, AMD [49]	Private: 577	AUC:0.998, precision (PR):98.9% [50]
Weakly supervised (DenseNet)	Fluid segmentation [51]	Private: 1217 OCT	Acc:94.8%
Weakly supervised localization	AMD [52]	Private: 10.1k	Acc: 94.9%
3D CNN	Glaucoma [53]	Private: 1110 3D SD-OCT	AUC:0.940
3D CNN	GA segmentation [54]	Private: 200k+ SD OCT scans	Mean OR: upto 87.24%+-7.95%
3D CNN	AMD (NSR) [55]	A2A: 384 SD OCT	MAPE:15.6
CNN + LSTM	Biomarker detection [56]	Private: 416 vols	F1: 0.694 +-0.009
CNN with soft attention map	Lesion detection [57]	UCSD (84k images) + NEH (148 vols)	Acc: 90.1%, SN:86.8%, PR: 86.2%

Table 1.1: CNN applications

1.2.3.2 Fully convolutional network (FCN) including Encoder - Decoder network

A fully convolutional neural network (FCN) consist of only convolutional layers and no fully connected or dense layers. Encoder decoder architecture is a typical example of an FCN and it has two distinct paths - a downsampling path made of convolutional layers and an upsampling path made of deconvolutional layers. The downsampling path represents the data in a high dimensional vector space and an identical but in reverse orientation upsampling path recreates the data from this representation. The FCN is typically used for applications like segmentation and denoising where output has similar dimensionality to the input, unlike one-dimensional classification tasks.

The table 1.2 presents some of the commonly used FCN and encoder-decoder architectures for retinal diagnostics. The measures used in addition to those previously discussed are dice score and mean structural similarity index measurement (SSIM). The most frequently performed task is retinal layer segmentation and it has been used to segment structures such as choroidal vessels, hyperreflective foci, retinal fluid, and drusen.

1.2.3.3 Generative adversarial network (GAN)

A GAN [68] is made of two major components - a generator and a discriminator as shown in figure 1.6. The generator learns to produces the data whereas the discriminator learns to distinguish between the spurious data from the generator and the real data. Both the networks are trained together in an adversarial way such that the generator gets better at producing data whereas the discriminator improves the ability to detect fake data. Most of the GANs used in current practice have both generator and discriminator comprised of convolutional networks.

In the area of retinal diagnostics, GANs have been used to perform tasks requiring the generation of new images/image masks such as segmentation, denoising, image generation and super-resolution of OCT images as described in table 1.3. The various performance measures used include dice score, SSIM, signal to noise ratio (SNR) and peak signal to noise ratio (PSNR).

Architecture	Application	Dataset	Performance	Others
FCN: Patch based U-Net	Layer segmentation [58]	TASMC & OCTexplorer	F1: upto 0.95	[59]
	Cornea segmentation [60]	Private:20k	Acc:99.5%	
FCN: DenseNet + Gaussian process	Layer segmentation [61]	Pvt: UMiami-50 images	Mean error: 1.1	
FCN: ResNet based	Choroidal vessel segmentation [62]	Private: 40 SS-OCT	Avg SA:0.840 +-0.035	
FCN: 4 conv layers	hyper-reflective foci segmentation [63]	Private: 1111 SD-OCT slices	Dice:over 95%	
FCN + RF	Fluid segmentation [64]	Multiple datasets	Acc: IRF:0.9815, SRF:0.9653, PED:0.9931	
Encoder decoder: single encoder multi decoder	Drusen segmentation [65]	Mix of public/private: 366 SD-OCT vols	Absolute surface diff: ILM:0.65+-0.06, IBRPE:1.06+-0.12, BM:0.9+-0.08	
Encoder decoder: DeconvNet and U-Net	Layer segmentation (ReLayNet) [66]	DUKE	F1:0.94	
Encoder decoder: residual blocks and skip conn.	Denoising [67]	Private: 3,880 scans	Mean SSIM:0.65 +- 0.03	

Table 1.2: FCN including encoder-decoder applications

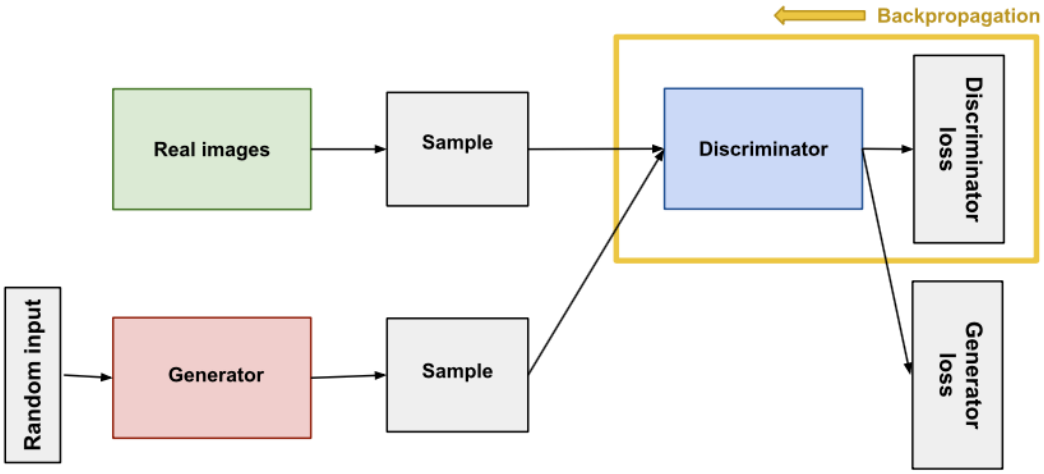


Figure 1.6: GAN schematic diagram [69]

Architecture	**Application**	**Dataset**	**Performance**
cGAN: Unet as generator, PatchGAN as discriminator	Denoising [70]	384 B scan pairs	SNR:60.09 +- 8.00
	Image generation [71]	Private 600 images	SSIM: upto 63.30%
GAN: encoder-decoder with skip conn. as gen.	Fluid segmentation [72]	RETOUCH: 42 volumes	Dice: IRF: 0.69, SRF: 0.67, PED: 0.85
GAN: other	Denoising & super–resolution [73]	Multiple sources	PSNR: upto 28.13

Table 1.3: GAN algorithms

1.2.4 Summary of deep learning for retinal OCT diagnosis

In the previous sections, a plethora of research directions and development of computer-aided diagnosis with OCT images have been discussed. The major difference between traditional machine learning methods and deep learning methods is that the feature maps are generated automatically in deep learning methods which is in contrast to the need for manual feature engineering for traditional algorithms. With the growing availability of high power computing resources, deep learning has become popular and surpassed the results of the other methods for disease detection and image segmentation tasks. Since, medical diagnosis procedure is very crucial and sensitive, the computer-aided models should be made extremely robust to perform in real-world scenarios and sensitivity is a key measure of this goal.

To sum it up, deep learning methods have considerably advanced the automated diagnosis of retinal diseases through OCT images. The major tasks performed include the classification of different diseases like AMD, glaucoma, DME, and DR as well as segmentation of various biomarkers like retinal layers, drusens, retinal fluid, hyper-reflective foci etc. There is a rising interest in denoising, image generation and super-resolution tasks and newer deep learning methods are being exploited for these. Various types of neural networks designs such as CNN, FCN, encoder decoder and GAN have been demonstrated to achieve a high performance on this wide variety of tasks.

A key challenge with the acceptance of deep learning for clinical applications in ophthalmology and other domains is the lack of reasoning for the decisions. It leads to a lack of trust in the model results amongst clinicians, regulators, and patients. The next section describes the concept of explainability of deep learning models with a focus on medical imaging.

1.3 Explainable deep learning for medical images

Deep learning is the leading AI method for a wide range of tasks including medical imaging problems. It is the state of the art for several computer vision tasks and has been used for medical imaging tasks like the classification of Alzheimer's [74], lung cancer detection [75], retinal disease detection [1], [2], etc. Despite achieving remarkable results in the medical domain, AI-based methods have not achieved a significant deployment in the clinics. This is due to the underlying black-box nature of the deep learning algorithms along with other reasons like computational costs. It arises from the fact that despite having the underlying statistical principles, there is a lack of ability to explicitly represent the knowledge for a

given task performed by a deep neural network. Simpler AI methods like linear regression and decision trees are self-explanatory as the decision boundary used for classification can be visualized in a few dimensions using the model parameters. But these lack the complexity required for tasks such as classification of 3D and most 2D medical images. The lack of tools to inspect the behavior of black-box models affects the use of deep learning in all domains including finance and autonomous driving where explainability and reliability are the key elements for trust by the end-user.

A medical diagnosis system needs to be transparent, understandable, and explainable to gain the trust of physicians, regulators as well as the patients. Ideally, it should be able to explain the complete logic of making a certain decision to all the parties involved. Newer regulations like the European General Data Protection Regulation (GDPR) are making it harder for the use of black-box models in all businesses including healthcare because retraceability of the decisions is now a requirement [76]. An AI system to complement medical professionals should have a certain amount of explainability and allow the human expert to retrace the decisions and use their judgment. Some researchers also emphasize that even humans are not always able to or even willing to explain their decisions [76]. Explainability is the key to safe, ethical, fair, and trust-able use of AI and a key enabler for its deployment in the real world. Breaking myths about AI by showing what a model looked at while making the decision can inculcate trust among the end-users. It is even more important to show the domain-specific features used in the decision for non-deep learning users like most medical professionals.

The terms explainability and interpretability are often used interchangeably in the literature. A distinction between these was provided in [77] where interpretation was defined as mapping an abstract concept like the output class into a domain example, while explanation was defined as a set of domain features such as pixels of an image the contribute to the output decision of the model. A related term to this concept is the uncertainty associated with the decision of a model. Deep learning classifiers are usually not able to say "I don't know" in situations with ambiguity and instead return the class with the highest probability, even if by a narrow margin. Lately, uncertainty has been analyzed along with the problem of explainability in many studies to highlight the cases where a model is unsure and in turn make the models more acceptable to non-deep learning users. Deep learning models are considered as non-transparent as the weights of the neurons can't be understood as knowledge directly. [78] showed that neither the magnitude or the selectivity of the activations, nor the impact on network decisions is sufficient for deciding the importance of a neuron for a given task. A detailed analysis of the terminologies, concepts and, use cases of explainable AI is provided in [79].

This section describes the studies related to the explainability of deep learning models

in the context of medical imaging. A general taxonomy of explainability approaches is described briefly in the next section and a comparison of various attribution based methods is performed in section 1.3.2. Section 1.3.3 reviews various explainability methods applied to different medical imaging modalities. The analysis is broken down into subsections 1.3.3.1 and 1.3.3.2 depending upon the use of attributions or other methods of explainability. The evolution, current trends, and some future possibilities of the explainable deep learning models in medical image analysis are summarized in 1.3.4.

1.3.1 Taxonomy of explainability approaches

Several taxonomies have been proposed in the literature to classify different explainability methods[80], [81]. Generally, the classification techniques are not absolute, it can vary widely depending upon the characteristics of the methods and can be classified into many overlapping or non-overlapping classes simultaneously. Different kinds of taxonomies and classification methods are discussed briefly here and a detailed analysis of the taxonomies can be found in [79], [80] and a flow chart for them is shown in 1.7.

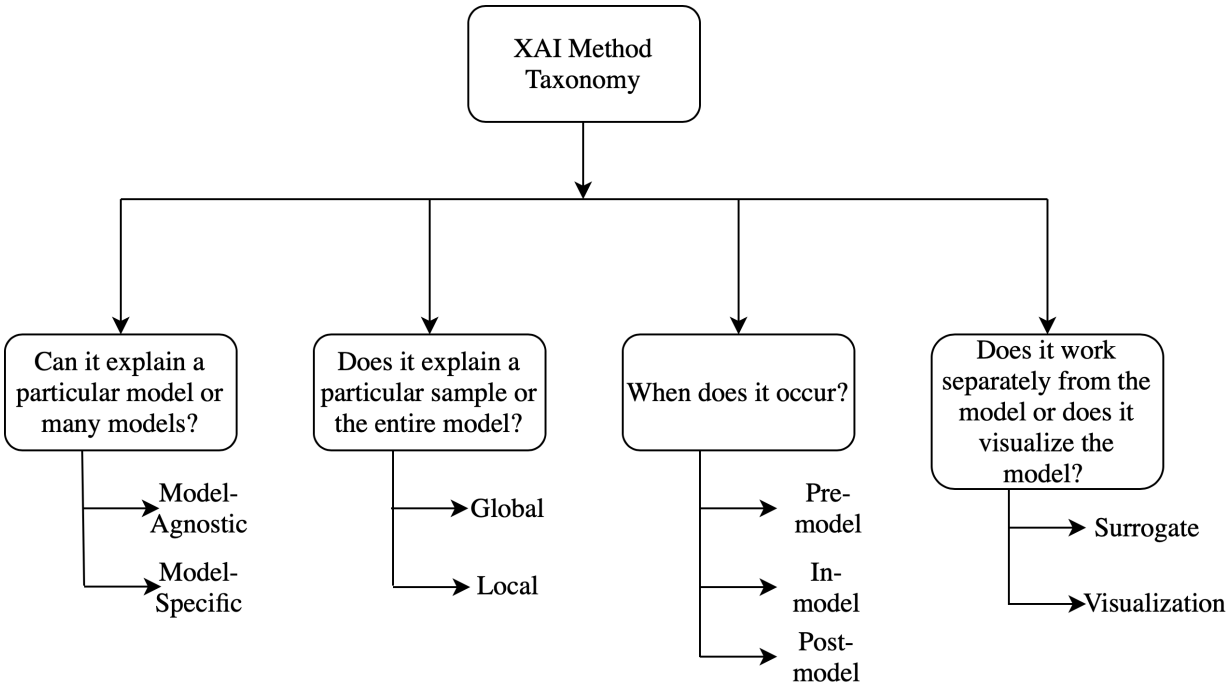


Figure 1.7: Taxonomy of XAI methods

It is to be noted that these classification methods are non-exclusive, these are built upon different logical intuitions and hence have significant overlaps. For example, most of the post-hoc models like attributions can also be seen as model agnostic as these methods are typically not dependent upon the structure of a model. However, some requirements regarding the limitations on model layers or the activation functions do exist for some of the attribution methods. The next section describes the basic concept and subtle difference between various attribution methods to facilitate a comparative discussion of the applications in section 1.3.3.

1.3.2 Explainability methods - attribution based

There are broadly two types of approaches to explain the results of deep neural networks (DNN) in medical imaging - those using standard attribution based methods and those using novel, often architecture or domain-specific techniques. The methods used for the former are discussed in this section with applications provided in 1.3.3.1 while the latter are discussed along with their applications in section 1.3.3.2. The problem of assigning an attribution value or contribution or relevance to each input feature of a network led to the development of several attribution methods. The goal of an attribution method is to determine the contribution of an input feature to the target neuron which is usually the output neuron of the correct class for a classification problem. The arrangement of the attributions of all the input features in the shape of the input sample forms heatmaps known as the *attribution maps*. Some examples of attribution maps for different images are shown in Figure 1.8. The features with a positive contribution to the activation of the target neuron are typically marked in red while those negatively affecting the activation are marked in blue. These are the features or pixels in case of images providing positive and negative evidence of different magnitudes respectively.

The commonly used attribution methods are discussed in this section and the applications in the next section. It must be noted that some of the approaches like DeepTaylor [83] provide only positive evidence and can be useful for a certain set of tasks.The attribution methods can be applied on a black box CNN without any modification to the underlying architecture making them a convenient yet powerful explainable AI (XAI) tool. An empirical comparison of some of the methods discussed in this section and a unified framework called *DeepExplain* is available in [84]. Most of the methods discussed here apart from the newer Deep Learning Important FeaTures (DeepLIFT) and Deep SHapley Additive exPlanations (SHAP) are implemented in the iNNvestigate toolbox [82].

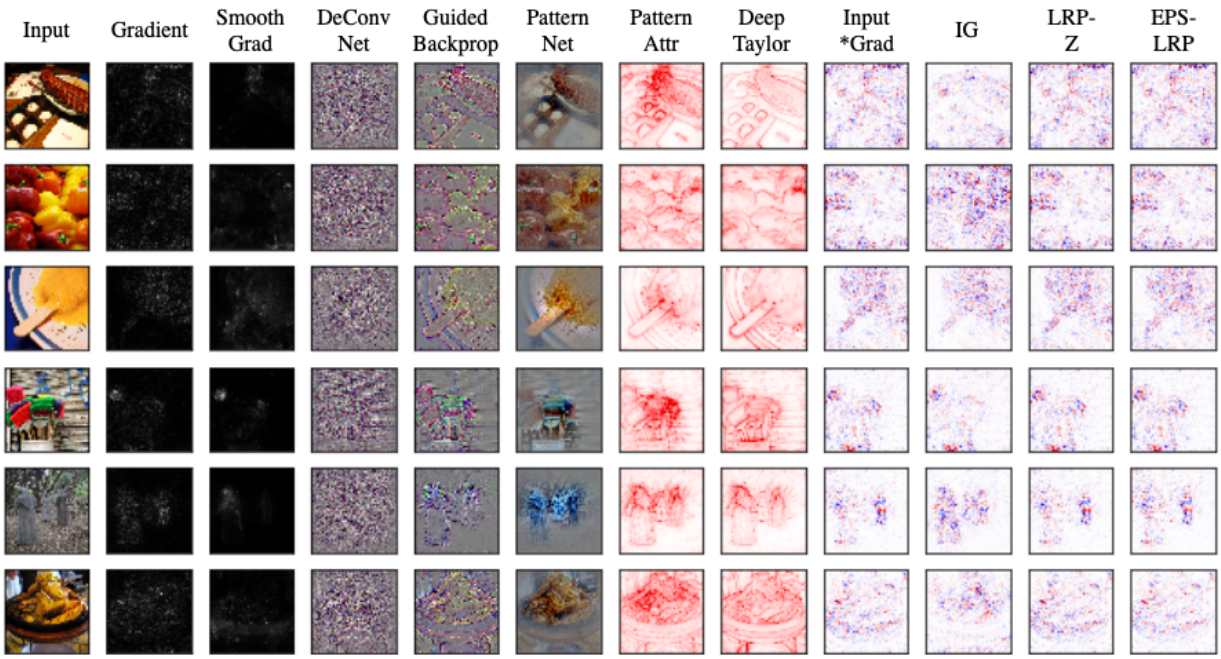


Figure 1.8: Attributions of VGG-16 with images from Imagenet using the methods implemented in [82]

1.3.2.1 Perturbation based methods - Occlusion

Perturbation is the simplest way to analyze the effect of changing the input features on the output of an AI model. This can be implemented by removing, masking, or modifying certain input features, and running the forward pass (output computation), and measuring the difference from the original output. This is similar to the sensitivity analysis performed in parametric control system models. The input features affecting the output the most are ranked as the most important. It is computationally expensive as a forward pass needs to be run after perturbing each group of features of the input. In the case of image data the perturbation is performed by covering parts of an image with a grey patch and hence *occluding* them from the system's view. It can provide both positive and negative evidence by highlighting the responsible features.

This technique was applied by Zeiler and Fergus [85] to the CNN for the image classification task. **Occlusion** is the benchmark for any attribution study as it is a simple to perform model agnostic approach which reveals the feature importance of a model. It can reveal if a model is overfitting and learning irrelevant features as in the case of adversarial

examples [86]. The adversarial examples are the inputs designed to cause the model to make a false decision and are like optical illusions for the models. In that case, the model misclassifies the image (say a cat as a dog) despite the presence of discriminating feature

Occluding all features (pixels) one-by-one and running the forward pass each time can be computationally expensive and can take several hours per image [84]. It is common to use patches of sizes such as 5x5, 10x10, or even larger depending on the size of the target features and computational resources available.

Another perturbation based approach is **Shapley value sampling** which computes approximate Shapely Values by taking each input feature for a sample number of times. It a method from the coalitional game theory which describes the fair distribution of the gains and losses among the input features. It was originally proposed for the analysis of regression [87]. It is slower than all other approaches as the network has to be run samples $\times$ number of features times. As a result it is not a practical method in its original form but has led to the development of game theory-based methods like Deep SHAP as discussed in the next subsection.

1.3.2.2 Backpropagation based methods

These methods compute the attribution for all the input features with a single forward and backward pass through the network. In some of the methods these steps need to be repeated multiple times but it is independent of the number of input features and much lower than for perturbation-based methods. The faster run-time comes at the expense of a weaker relationship between the outcome and the variation of the output. Various backpropagation based attribution methods are described in Table 1.4. It must be noted that some of these methods provide only positive evidence while others provide both positive and negative evidence. The methods providing both positive and negative evidence tend to have high-frequency noise which can make the results seem spurious. [84].

Table 1.4: Backpropagation based attribution methods

Method	**Description**	**Notes**
Gradient	Computes the gradient of the **output neuron** with respect to the input.	The **simplest** approach but is usually not the most effective.
DeConvNet [85]	Applies the **rectified linear unit (ReLU) to the gradient computation instead** of the gradient of a neuron with ReLU activation.	Used to **visualize the features** learned by the layers. **Limited** to CNN models with **ReLU activation**.

Saliency Maps [88]	Takes the **absolute value of the partial derivative** of the target output neuron with respect to the input features to find the features which affect the output the most with least perturbation.	**Can't distinguish between positive and negative** evidence due to absolute values.
Guided backpropagation (GBP) [89]	Applies the **ReLU to the gradient computation in addition** to the gradient of a neuron with ReLU activation.	Like DeConvNet, it is **limited** to CNN models with **ReLU activation**.
Layer wise relevance propagation (LRP) [90]	**Redistributes the prediction score** layer by layer with a backward pass on the network using a particular rule like the **ϵ-rule** while ensuring numerical stability	There are alternative stability rules and **limited** to CNN models with **ReLU activation** when all activations are **ReLU**.
Gradient $\times$ input [91]	Initially proposed as a method to **improve sharpness of attribution maps** and is computed by multiplying the signed partial derivative of the output with the input.	It **can approximate occlusion** better than other methods in certain cases like multilayer perceptron (MLP) with Tanh on MNIST data [84] while being instant to compute.
Gradient weighted class activation mapping (GradCAM) [92]	Produces **gradient-weighted class activation maps** using the gradients of the target concept as it flows to the final convolutional layer	Applicable to **only CNN** including those with fully connected layers, structured output (like captions) and reinforcement learning.
Integrated gradients (IG) [93]	Computes the **average gradient** as the input is varied from the **baseline** (often zero) to the actual input value unlike the Gradient $\times$ input which uses a single derivative at the input.	It is **highly correlated with the rescale rule of DeepLIFT** discussed below which can act as a good and faster approximation.

DeepTaylor [83]	Finds a rootpoint near each neuron with a value close to the input but with output as 0 and uses it to recursively estimate the attribution of each neuron using **Taylor decomposition**.	Provides **sparser explanations** i.e. focuses on key features but provides **no negative evidence** due to its assumptions of only positive effect.
PatternNet [94]	Estimates the input signal of the output neuron using an **objective function**.	Proposed to counter the incorrect attributions of other methods on **linear systems** and generalized to deep networks.
Pattern Attribution [94]	Applies Deep Taylor decomposition by searching the **rootpoints in the signal direction** for each neuron	Proposed along with **PatternNet** and uses decomposition instead of signal visualization
DeepLIFT [95]	Uses a reference input and computes the reference values of all hidden units using a forward pass and then proceeds backward **like LRP**. It has two variants - **Rescale rule** and the one introduced later called **RevealCancel** which treats positive and negative contributions to a neuron separately.	Rescale is strongly related to and **equivalent in some cases to ϵ-LRP** but is **not applicable to models involving multiplicative rules**. **RevealCancel handles such cases** and using RevealCancel for convolutional and Rescale for fully connected layers reduces noise.
SmoothGrad [96]	An improvement on the gradient method which averages the gradient over multiple inputs with additional noise	Designed to visually sharpen the attributions produced by gradient method using class score function.
Deep SHAP [97]	It is a fast **approximation** algorithm to compute the game theory based **SHAP values**. It is connected to DeepLIFT and uses **multiple background samples** instead of one baseline.	Finds attributions for **non neural net models** like trees, support vector machines (SVM) and **ensemble** of those with a neural net using various tools in the the SHAP library.

An important property of attribution methods known as *completeness* was introduced in the DeepLIFT [95] paper. It states that the attributions for a given input add up

to the target output minus the target output at the baseline input. It is satisfied by integrated gradients, DeepTaylor and Deep SHAP but not by DeepLIFT in its rescale rule. A measure generalizing this property is proposed in [84] for a quantitative comparison of various attribution methods. It is called *sensitivity-n* and involves comparing the sum of the attributions and the variation in the target output in terms of Pearson's correlation coefficient (PCC). Occlusion is found to have a higher PCC than other methods as it finds a direct relationship between the variation in the input and that in the output.

The evaluation of attribution methods is complex as it is challenging to discern between the errors of the model and the attribution method explaining it. Measures like sensitivity-n reward the methods designed to reflect the network behavior closely. However, a more practically relevant measure of an attribution method is the similarity of attributions to a human observer's expectation. It needs to be performed with a human expert for a given task and carries an observer bias as the methods closer to the observer expectation can be favored at the cost of those explaining the model behavior. We underscore the argument that the ratings of different attribution methods by experts of a specific domain are potentially useful to develop explainable models which are more likely to be trusted by the end users and hence should be a critical part of the development of an XAI system.

1.3.3 Applications

The applications of explainability in medical imaging are reviewed here by categorizing them into two types - those using pre-existing attribution based methods and those using other, often specific methods. The methods are discussed according to the explainability method and the medical imaging application. Table 1.5 provides a brief overview of the methods.

1.3.3.1 Attribution based

A majority of the medical imaging literature that studied interpretability of deep learning methods used attribution based methods due to their ease of use. Researchers can train a suitable neural network architecture without the added complexity of making it inherently explainable and use a readily available attribution model. This allows the use of either a pre-existing deep learning model or one with a custom architecture for the best performance on the given task. The former makes the implementation easier and allows one to leverage techniques like transfer learning [30], [98] while latter can be used to focus on specific data and avoid overfitting by using fewer parameters. Both approaches are beneficial for medical

imaging datasets which tend to be relatively smaller than computer vision benchmarks like ImageNet [99].

Post-model analysis using attributions can reveal if the model is learning relevant features or if it is overfitting to the input by learning spurious features. This allows researchers to adjust the model architecture and hyperparameters to achieve better results on the test data and in turn a potential real-world setting. Some recent studies using attribution methods across medical imaging modalities such as brain magnetic resonance imaging (MRI) [100], [101], retinal imaging [52], [102]–[104], breast imaging [105], [106], skin imaging [107], [108], computerized tomography (CT) scans [109], and chest X-ray [110], [111],

The attribution based methods were one of the initial ways of visualizing neural networks and have since then evolved from simple class activation map and gradient-based methods to advanced techniques like Deep SHAP. The better visualizations of these methods show that the models were learning relevant features in most of the cases. Any presence of spurious features was scrutinized, flagged to the readers, and brought adjustments to the model training methods. Smaller and task-specific models like [52] along with custom variants of the attribution methods can improve the identification of relevant features.

1.3.3.2 Non-attribution based

The studies discussed in this subsection approached the problem of explainability by developing a methodology and validating it on a given problem rather than performing a separate analysis using pre-existing attributions based methods like those previously discussed. These used approaches like attention maps, concept vectors, returning a similar image, text justifications, expert knowledge, generative modeling, combination with other machine learning methods, etc. It must be noted that the majority of these are still post-model but their implementation usually needs specific changes to the model structure such as in the attention maps or the addition of expert knowledge in case of rule-based methods. In this section, the studies are grouped by the explainability approach they took.

The design of the methods in this case is more involved than the application of attribution based methods on the inputs of a trained model. Specific elements like concept vectors, expert-based rules, image retrieval methods need to be integrated often at a model training level. This added complexity can potentially provide more domain-specific explanations at the expense of higher design effort. Notably, a majority of these techniques are still a post-hoc step but for a specific architecture or domain. Also, here the scope is limited to medical imaging as that is the dominant approach for automated diagnosis because of the detailed information presented by the images. However, patient records also

provide rich information for diagnosis and there were studies discussing their explainability. For example, in [112] a gated recurrent unit (GRU)-based recurrent neural network (RNN) for mortality prediction from diagnostic codes from electronic healthcare record (EHR) was presented. It used hierarchical attention in the network for interpretability and visualization of the results.

Table 1.5: Applications of explainability in medical imaging

Method	**Algorithm**	**Model**	**Application**	**Modality**
Attribution	Gradient*I/P, GBP, LRP, occlusion [100]	3D CNN	Alzheimer's detection	Brain MRI
	GradCAM, GBP [101]	Custom CNN	Grading brain tumor	Brain MRI
	IG [102]	Inception-v4	DR grading	Fundus images
	EG [52]	Custom CNN	Lesion segmentation for AMD	Retinal OCT
	IG, SmoothGrad [105]	AlexNet	Estrogen receptor status	Breast MRI
	Saliency maps [106]	AlexNet	Breast mass classification	Breast MRI
	GradCAM, SHAP [107]	Inception	Melanoma detection	Skin images
	Activation maps [108]	Custom CNN	Lesion classification	Skin images
	DeepDreams [109]	Custom CNN	Segmentation of tumor from liver	CT imaging
	GSInquire, GBP, activation maps [110]	COVIDNet CNN	COVID-19 detection	X-ray images
Attention	Mapping between image to reports [113]	CNN & LSTM	Bladder cancer	Tissue images
	U-Net with shape attention stream [114]	U-net based	Cardiac volume estimation	Cardiac MRI

Concept vectors	TCAV [115]	Inception	DR detection	Fundus images
	TCAV with RCV [116]	ResNet101	Breast tumor detection	Breast lymph node images
	UBS [117]	SqueezeNet	Breast mass classification	Mammogram images
Expert knowledge	Domain constraints [118]	U-net	Brain MLS estimation	Brain MRI
	Rule-based segmentation, perturbation [119]	VGG16	Lung nodule segmentation	Lung CT
Similar images	GMM and atlas [77]	3D CNN	MRI classification	3D MNIST, Brain MRI
	Triplet loss, kNN [120]	AlexNet based with shared weights	Melanoma	Dermoscopy images
	Monotonic constraints [121]	DNN with two streams	Melanoma detection	Dermoscopy images
Textual justification	LSTM, visual word constraint [122]	Breast mass classification	CNN	Mammogra-phy images
Intrinsic explainability	Deep Hierarchical Generative Models [123]	Auto-encoders	Classification and segmentation for Alzheimer's	Brain MRI
	SVM margin [124]	Hybrid of CNN & SVM	ASD detection	Brain fMRI

1.3.4 Discussion

There has been significant progress in explaining the decisions of deep learning models, especially those used for medical diagnosis. Understanding the features responsible for a

certain decision is useful for the model designers to iron out reliability concerns for the end-users to gain trust and make better judgments. Almost all of these methods target local explainability, i.e. explaining the decisions for a single example. This then is extrapolated to a global level by averaging the highlighted features, especially in cases where the images have the same spatial orientation. However, emerging methods like concept vectors [115] provide a more global view of the decisions for each class in terms of domain concepts.

It is important to analyze the features of a black-box which can make the right decision due to the wrong reason. It is a major issue that can affect performance when the system is deployed in the real world. Most of the methods, especially the attribution based are available as open source implementations. However, some methods like GSInquire [111] which show higher performance on some metrics are proprietary. There is an increasing commercial interest in explainability, and specifically the attribution methods which can be leveraged for a variety of business use cases.

Despite all these advances, there is still a need to make the explainability methods more holistic and interwoven with uncertainty methods. More studies like [102] need to be conducted to observe the effect of the explainability models on the decision time and accuracy of the clinical experts. Expert feedback must be incorporated into the design of such explainability methods to tailor the feedback for their needs. Initially, any clinical application of such explainable deep learning methods is likely to be a human-in-the-loop (HITL) hybrid keeping the clinical expert in the control of the process. It can be considered analogous to driving aids like adaptive cruise control or lane keep assistance in cars where the driver is still in control and responsible for the final decisions but with a reduced workload and an added safety net.

Another direction of work can be to use multiple modalities like medical images and patients' records together in the decision-making process and attribute the model decisions to each of them. This can simulate the diagnostic workflow of a clinician where both images and physical parameters of a patient are used to make a decision. It can potentially improve the accuracy as well as explain in a more comprehensive way. To sum it up, explainable diagnosis is making convincing strides but there is still some way to go to meet the expectations of end-users, regulators, and the general public.

Chapter 2

Explaining deep learning models for retinal OCT diagnosis

This chapter and the next two are based on:

- **A Singh, S Sengupta, J J Balaji, M A Rasheed, I Faruq, V Jayakumar, J S Zelek, and V Lakshminarayanan, "What is the optimal attribution method for explainable ophthalmic disease classification?" In International Workshop on Ophthalmic Medical Image Analysis, Springer, 2020.**
- **A Singh, A R Mohammed, J S Zelek, and V Lakshminarayanan. "Interpretation of deep learning using attributions: application to ophthalmic diagnosis." Proc. SPIE 11511, Applications of Machine Learning, 2020.**

2.1 Introduction

Retinal diseases are prevalent among large sections of society, especially amongst the aging population and those with other systemic diseases such as diabetes [125]. It is estimated that the number of Americans over 40 years with a diabetic retinopathy (DR) diagnosis will rise threefold from 5.5 million in 2005 to 16 million in 2050 [126]. For each decade of age after 40, the prevalence of low vision and blindness increases by a factor of three [127]. There is a widespread shortage of trained medical professionals leading to longer wait times and unavailability of medical aid to remote communities. The situation is especially acute in developing countries. It is critical to have an early diagnosis for retinal diseases such as glaucoma where delayed treatment can cause irreversible vision loss. Automated screening and diagnostic assistance using computer-aided methods like deep learning have been suggested as a potential solution to make the diagnosis faster and more accessible. These methods can be used to assist clinicians in making more accurate and faster decisions.

Despite the emergence of many deep learning methods for retinal diagnosis [2], [128] their adoption in clinical settings is very limited [129]. The main hurdle is the lack of trust of the clinical end-users, regulators, and patients due to the black-box nature of the algorithms. These models can detect diseases with high accuracy which is often comparable to human experts [130] but can not explain the logic for their decision. The explainability methods provide reasoning for model decisions. A majority of those evaluate the contribution of each pixel of the image to the model output and hence are called attribution methods. An overview of explainability methods is discussed in section 1.3 and more details are given in [3]. There are a very limited number of studies for explaining the retinal diagnosis performed by deep learning models [52], [102]. Studies are evaluating the impact of explainability on machine learning practitioners [131] and for comparing attribution methods quantitatively [100], [132]. Almost all the studies, especially the ones for ophthalmic diagnosis utilize a single explainability method and do not provide comparisons with alternatives. However, to the best of our knowledge, there is no study evaluating multiple attribution methods both quantitatively and qualitatively for retinal disease diagnosis.

In this study, we performed a quantitative analysis of the attribution methods using multiple measures - robustness, runtime, and sensitivity. The quantitative analysis is important to understand the ability of an attribution method to highlight the features according to their impact on the model output. However, we strongly believe that for any explainability method to be successfully used in the field it must be evaluated by trained experts as an assistive tool. A panel of retinal experts consisting of ophthalmologists and optometrists evaluated the methods for their ability to justify the predicted class in terms of the similarity to the clinical concepts. The use of attributions as a tool to improve the

models and inculcate the trust of clinician end-users through visualizations is discussed in this study.

Section 2.2 describes the studies on explainability and more specifically the applications for retinal diagnosis. Commonly used Inception-v3 [133] architecture was trained on the large UCSD OCT dataset [134] to classify images among 4 classes - choroidal neovascularization (CNV), DME, drusen and normal. CNV refers to the formation of new leaky blood vessels in the choroid beneath the retina, DME is the accumulation of fluid in the most visually active region called the macula, and drusen are the yellowish deposits of lipids and proteins under the retina. The experiment is discussed in section 2.3 highlighting the dataset, model training, and the generation of explanations. The quantitative analysis using multiple metrics is described in chapter 3. These methods were rated by a panel of 14 eye care professionals (10 ophthalmologists and 4 optometrists). Their observations regarding the clinical significance of these methods, preference regarding AI systems, and suggestions for future implementations are analyzed in chapter 4. The findings are concluded in section 5 with directions for future research. Further details can be found in [103], [104], [135].

2.2 Related studies

It is imperative for both the machine learning practitioners and the end-users to observe the relevant features used by an AI system for making decisions. It leads to a better understanding of the interaction between the model and the data [131] enabling the former to design better models. It can inculcate confidence and trust in the domain experts leading to the more responsible use of deep learning methods. Explaining the diagnostic and treatment decisions to all the parties involved is an integral part of the modern healthcare system. The ethical and legal challenges of the domain require decisions to be more transparent, explainable, and understandable for the users. This has lead to advances in the development of XAI systems for medical diagnosis [76]. The key challenges and opportunities for XAI are presented in [79] and a detailed categorization of the methods is provided in [80]. [3] discussed the applications of explainability in the medical imaging and highlighted a need for evaluation of these methods by end-users.

Deep learning methods are used for tasks like classification, segmentation, image enhancement, and image generation from retinal images captured by two common modalities - fundus camera and OCT scans. The classification problem deals with detection of diseases like glaucoma, DR, and AMD from retinal images. Segmentation involves identifying regions of interest such as optic cup and disc, retinal layers, drusen deposits, etc. Image

enhancement refers to denoising OCT scans, increasing the details with super-resolution, and generation of synthetic data for model training. Reviews of deep learning methods for ophthalmic diagnosis are available in [1], [2], [10], [128]. IDx-DR was a method for DR classification from fundus images [129] which received FDA approval.

In a study for weakly-supervised segmentation of lesions for AMD diagnosis [52], an extension of IG called expressive gradients (EG) was proposed. The EG method added the high-level attributions to the input only attributions of IG, outperforming it when applied along with a relatively small custom CNN. The impact of the model predictions and attributions generated by IG on DR grading by ophthalmologists was studied in [102]. The combination of class probabilities and attributions was found to be the most effective in improving the grading accuracy of the users compared to only the probabilities or no assistance. The grading time of the users increased initially but it reduced from the initial levels after prolonged use of the assistance showing the potential to increase the patient throughput and improve the diagnosis simultaneously.

Recent studies have looked into the quantitative analysis of multiple attribution methods [84], [132] in terms of the theoretical principles. A study in the domain of brain imaging [100] performed a robustness analysis to measure the repeatability of the attributions generated by various methods. Motivated by the findings of these as well as the studies for explainable retinal diagnosis [52], [102], we explore the efficacy of different attribution methods to highlight the clinically relevant regions of the images. Instead of evaluating them as a tool for weakly-supervised segmentation and then comparing with markings of a clinician, we suggest their use as a framework for understanding the model - data interactions and assisting clinical end-users.

2.3 Methods

This section describes the dataset, computational hardware, model training, and generation of attributions using various methods.

2.3.1 Dataset

In this study, we used a large publicly available dataset known as the UCSD dataset [38]. It has images in "training" and "test" folders with 4 classes and the details are as shown in table 2.1. It is observed that the dataset is not balanced with the most images in the CNV class. However, no issues due to class imbalance are observed in training due to the

large size of the dataset. It consisted of 83.5k training images from four classes - CNV (37.2k), DME (11.3k), drusen (8.6k) and normal (26.3k). The test set of 1000 images had 250 images from each class.

Table 2.1: Dataset description showing the class level split for training and test sets.

Data	**CNV**	**DME**	**Drusen**	**Normal**	**Total**
Training	37205	11348	8616	26315	83483
% of total	*44.57%*	*13.59%*	*10.32%*	*31.52%*	*100%*
Test	250	250	250	250	1000
% of total	*25.00%*	*25.00%*	*25.00%*	*25.00%*	*100%*

2.3.2 Computational hardware

Two kind of systems were used for model training and attribution generations:

1. *Intel Core i7 9700K 3.60GHz 8 core CPU, 64GB RAM, Nvidia Titan V 12GB GPU*: Used for training the model and hyperparameter tuning of a single model.
2. *Compute Canada Beluga nodes: Intel Gold 6148 Skylake 2.4 GHz processor, up to 128GB RAM, upto 4 Nvidia Tesla V100 16GB GPU*: Used for training multiple model instances and calculating attributions in parallel by using more than one GPUs.

2.3.3 Model

The UCSD OCT dataset [134] was used to train an Inception-v3 [133] network and generate the attributions. The Inception-v3 model was chosen due to its prevalence in medical imaging, especially ophthalmic diagnosis [30], [38], [136], due to ease of implementation and availability of pre-trained weights. TensorFlow 2 library [137] with Keras API [138] was used for building the model. The model was trained from random weights to avoid any irrelevant features from pre-training. 20% of the training images from each class were separated for validation. The test accuracy for the ten training instances ranged between 99.00% and 99.90% with an average of 99.42%.

The confusion matrix for an instance with 99.30% accuracy is shown in figure 2.1. The model learned the labels in a balanced way despite the class imbalance discussed in subsection 2.3.1. Drusen has the most misclassification as CNV. The potential reasons could

be harder detection due to the relatively small size and presence of secondary diagnosis. Also, larger drusen deposits could be confused with smaller CNV structures, especially in the presence of high noise. In some cases, drusen and CNV are also hard to distinguish for human experts. Some normal images with little imperfections were misdiagnosed as drusen.

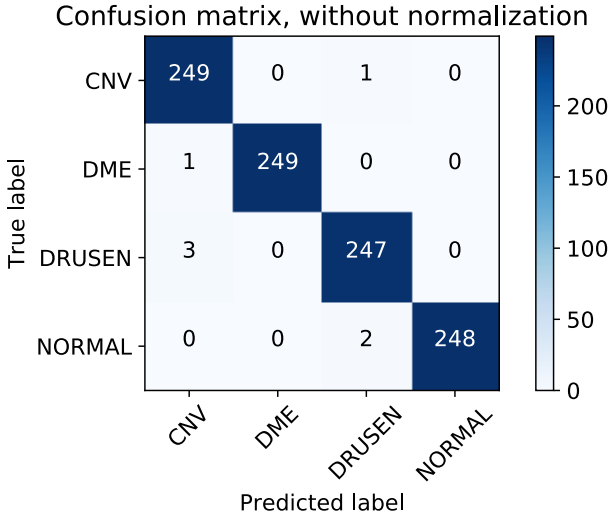


Figure 2.1: Confusion matrix with true labels on Y-axis and predictions on X-axis

The attributions were generated using variants of DeconvNet [85], Saliency maps [88], GBP [89], LRP [90], gradient times input, IG [93], DeepTaylor [83], DeepLIFT [91], SmoothGrad [96], DeepSHAP [97] as well as the baselines from gradient and occlusion. Three libraries were used for implementation of these methods - Innvestigate [82], Deep Explain [84], and SHAP [97]. SHAP and DeepLIFT are considered as state-of-the-art on standard machine learning datasets and have superior theoretical background while IG is commonly used for retinal images Note that some images of the source. A summary of these methods is provide in table 1.4 in chapter 1.

For LRP, the ϵ rule was used while DeepLIFT was used in the original rescale variant implemented by[84]. The reveal cancel rule of DeepLIFT in [95] was incompatible with the bias term of the Inception model. SHAP was the only model that required background distribution and we selected a random set of 20 normal images for the same denoting it as *SHAP random*. It was observed to be sensitive to artifacts and noise in the background

images. Hence, another variant with 20 normal images with low noise and artifacts was also used and denoted as *SHAP selected*. A window size of 64x64 and a step size of 16 were used for occlusion as the runtime is very high when every pixel is perturbed separately as explained in the section 1.3.2.1.

The heatmaps using the attribution methods for one correctly classified example of each disease and an incorrectly classified example of drusen are shown in figures 2.2, 2.3, 2.4 and, 2.5 respectively. A brief description of the output is provided with each figure. The source images in the dataset were cropped and rotated. Notably, certain methods such as DeepTaylor and Saliency provide only positive evidence. Those providing both positive and negative evidence have a high-frequency noise (negative evidence) that can be removed in practice but retained here to compare original outputs. The attributions generated by all the methods were analyzed both quantitatively and qualitatively and are discussed in chapters 3 and 4 respectively.

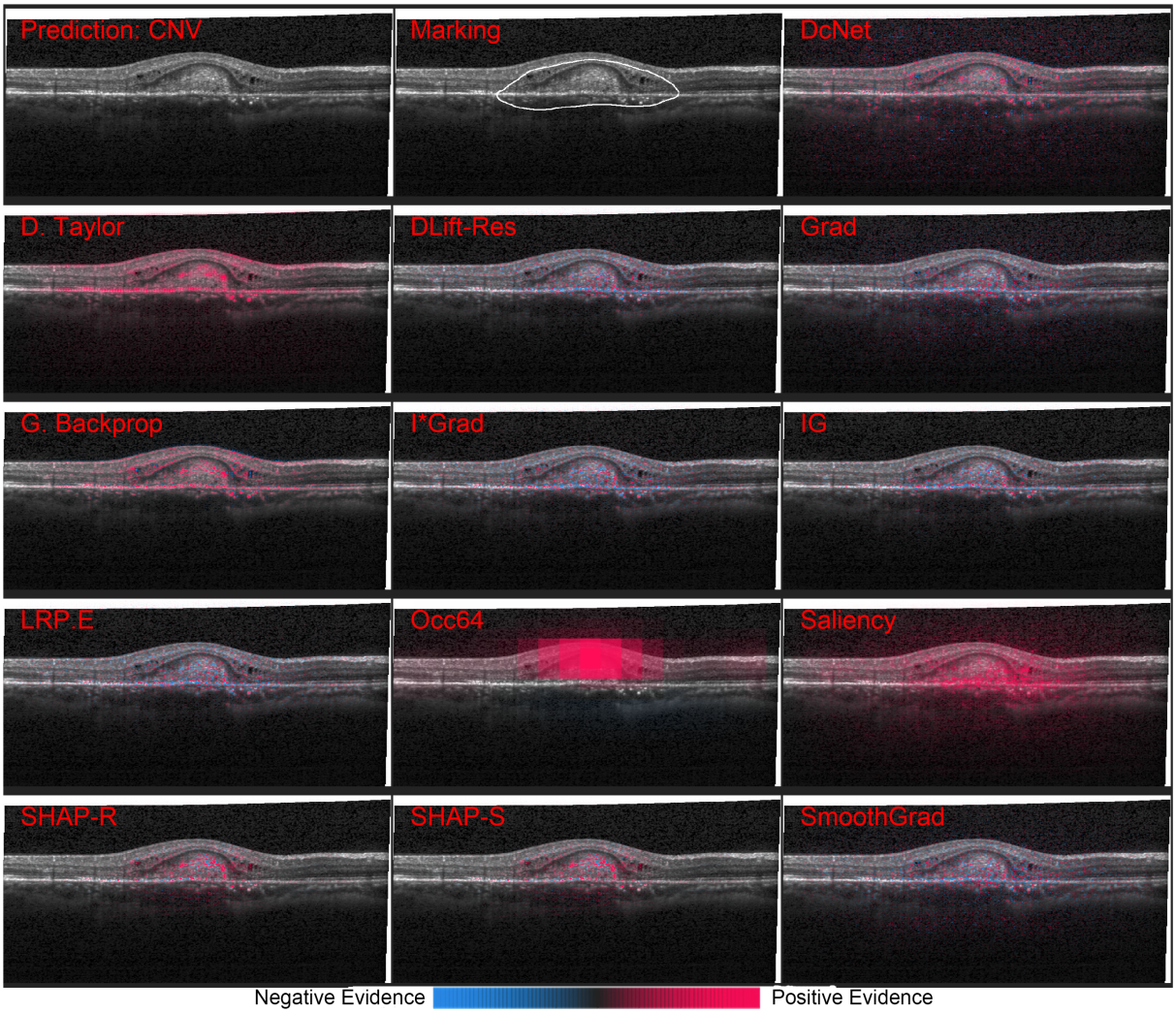


Figure 2.2: Heatmaps for a scan with choroidal neovascularization (CNV).The scale in the bottom shows that the parts highlighted in magenta color provide positive evidence regarding presence of a disease while those in blue color provide a negative evidence indicating that the image doesn't belong to the target class. DeepTaylor, GBP perform the best, SHAP highlights partial but precise regions, and the rest of the methods have varying amounts of noise. The fluid accumulation for CNV was highlighted by better performing methods.

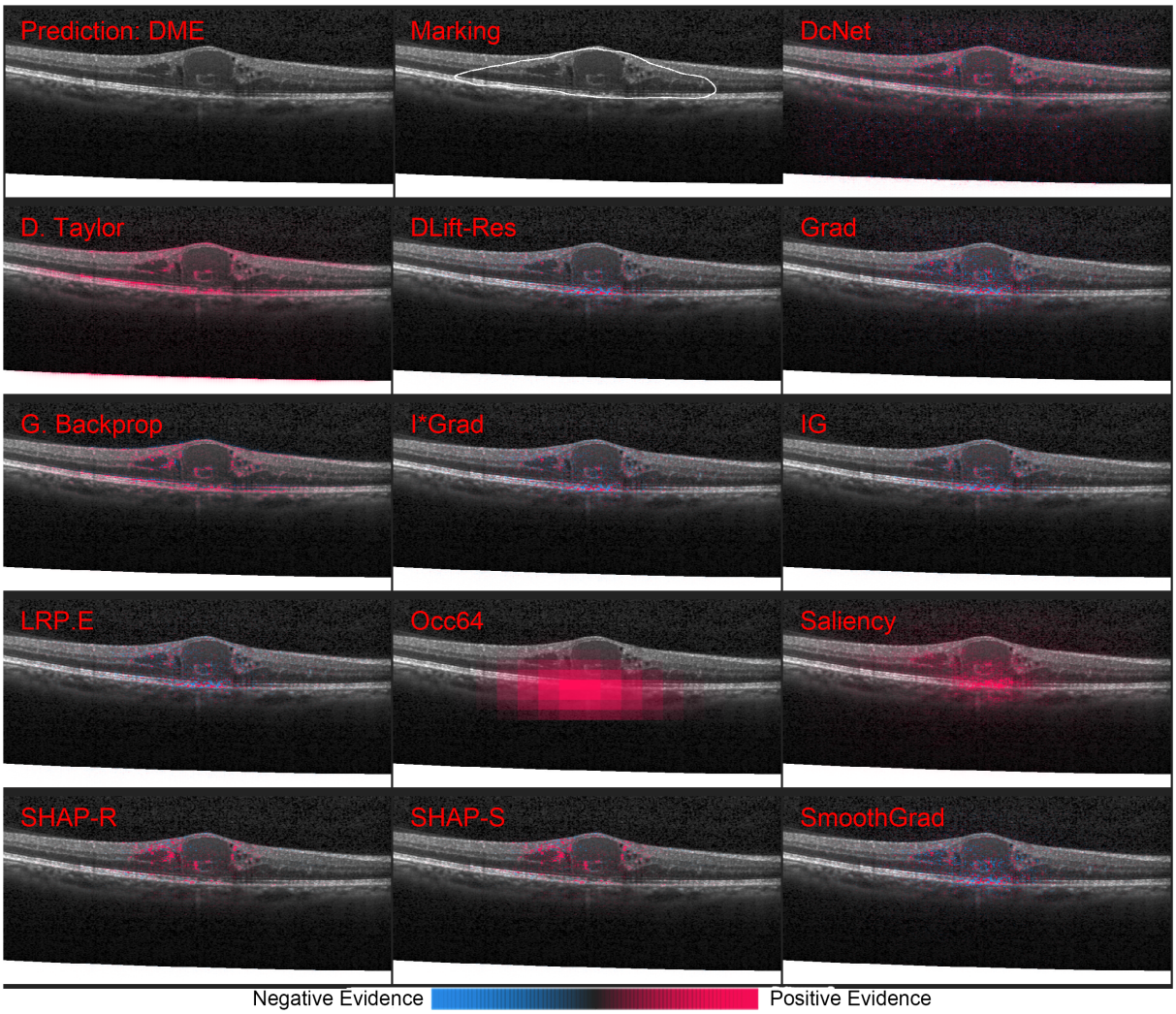


Figure 2.3: Heatmaps for a scan with diabetic macular edema (DME). Overall results are consistent with the CNV case. The edges of the edema for DME were highlighted by better performing methods.

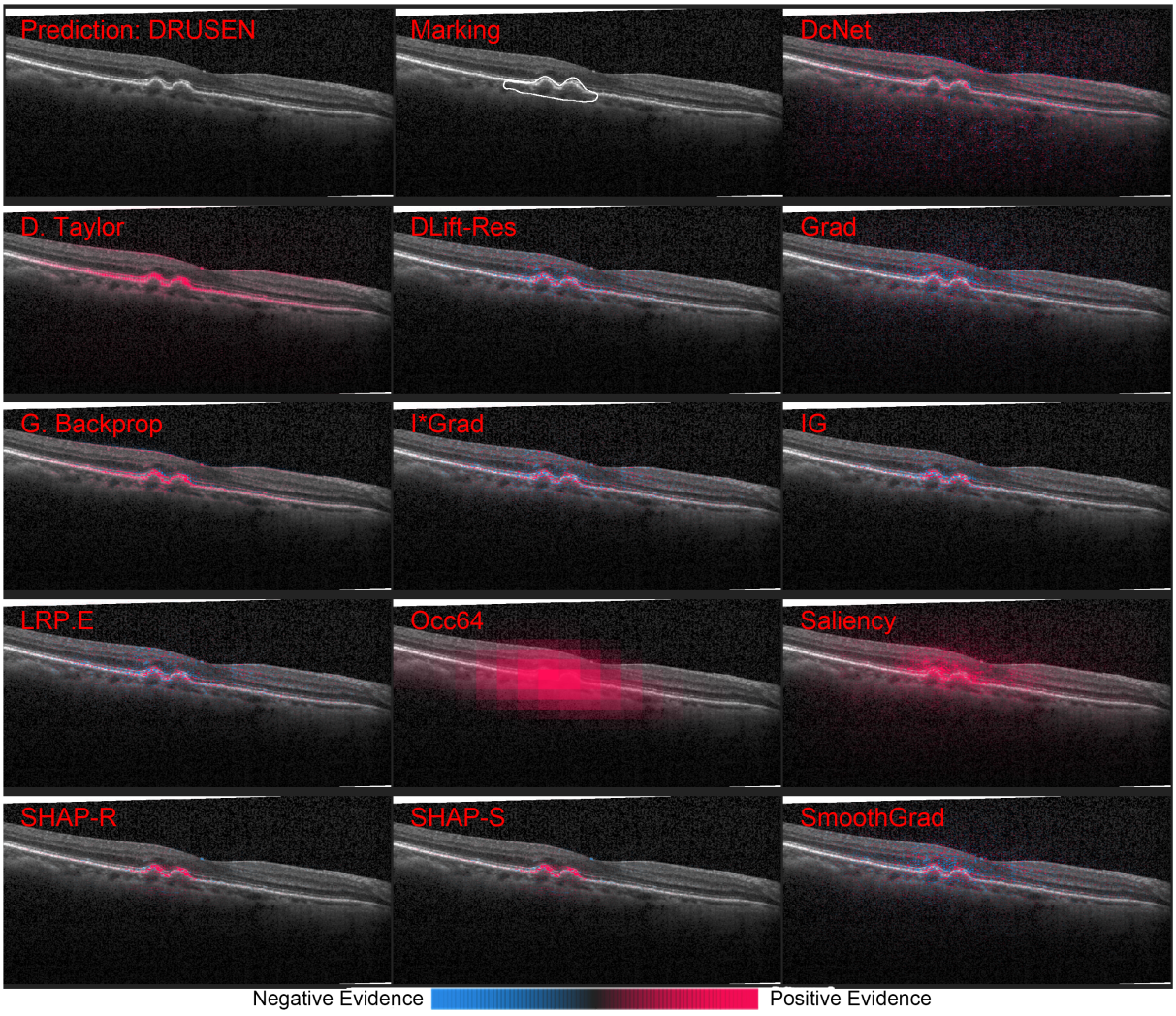


Figure 2.4: Heatmaps maps for a scan with drusen using various attribution methods. The pathological structures are smaller than the previous two and as a result most of the methods highlight regions outside too. SHAP is the most precise here. The performance of the methods can be observed in terms of positive highlights of the bumpy RPE.

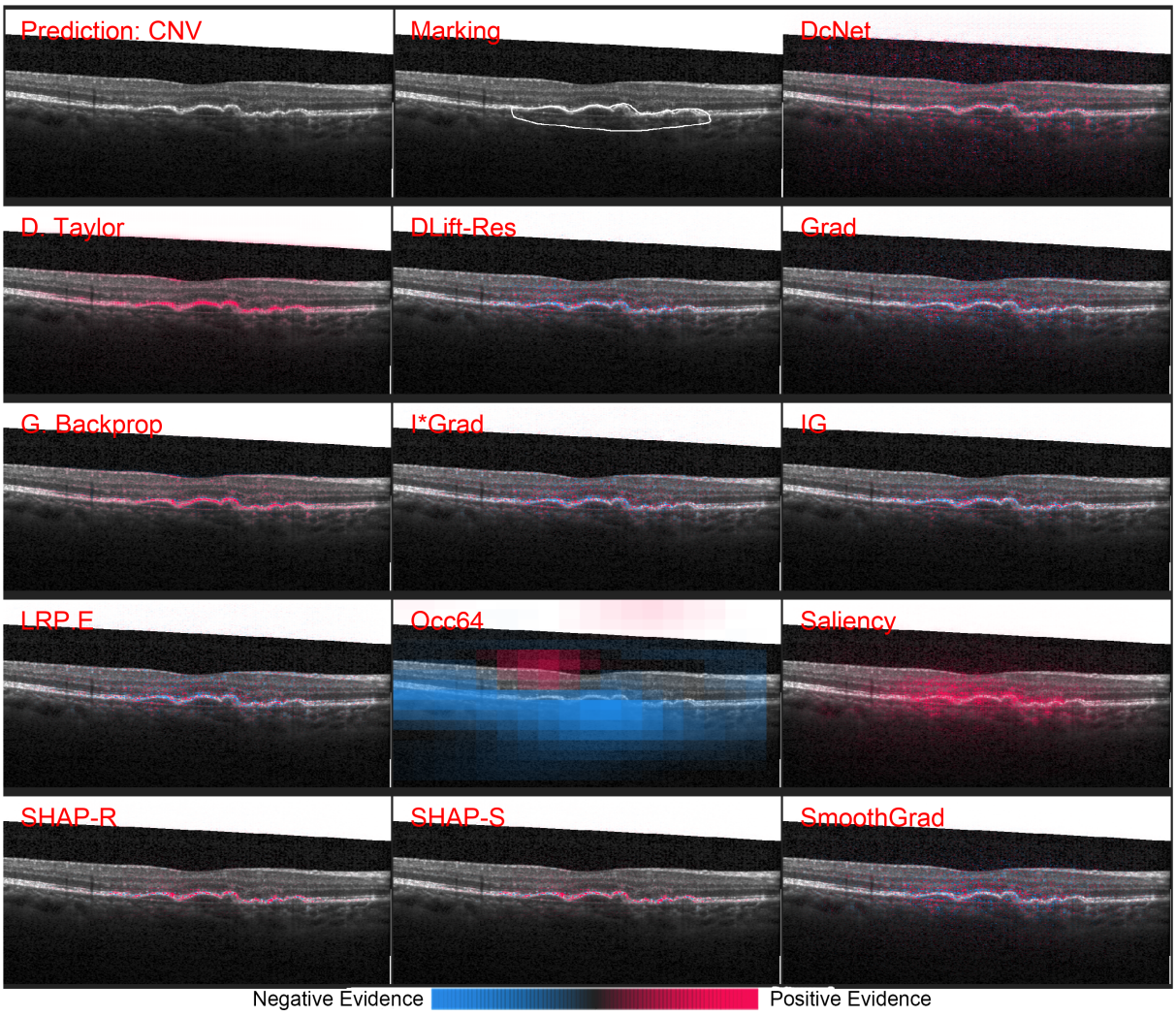


Figure 2.5: Heatmaps for a drusen scan misclassified as CNV. Most of the methods show a higher than usual amount of negative evidence as blue marks and there is a prominent blue glow over the drusen in occlusion.

Chapter 3

Quantitative evaluation of attribution methods

3.1 Introduction

The quantitative evaluation of attribution methods is tedious in the absence of ground truth unlike those for conventional segmentation tasks. Both the analyses were performed in two stages in separate studies. In the first stage deconvnet, DeepLIFT, GBP, input times gradient, IG, LRP - ϵ, occlusion, saliency maps, and SHAP were used. In the later stage, all the methods were used to compute improved metrics.

3.2 Stage 1 analysis

In the initial stage, three metrics - runtime, RMSE, and Spearman's rank correlation were calculated for a subset of the methods. These metrics describe the computational cost, absolute error, and the agreement between rankings of features with their impact on model output [139] respectively. It must be noted that these were calculated for a more recent deep learning model known as Inception-Resnet-v2 [29], an improvement over the popular Inception-v3. More details about the implementation are available in [103].

3.2.1 Runtime

The methods were run on the same system and benchmarked for the average time for heatmap generation of a single image. Figure 3.1 shows the runtimes along with error bars showing maximum and minimum bars for all the methods. SHAP had the highest runtime of over 1.5 seconds per image as it had to calculate the reference value of the neurons for the background distribution of normal images before finding the SHAP values for the sample. Occlusion was the second slowest despite the speed-up caused by the sliding window. The integrated gradient averages the gradient by varying it from a reference value and hence needs more computational effort. All other methods gave results in under 300ms, indicating their potential suitability for systems with lower computational power.

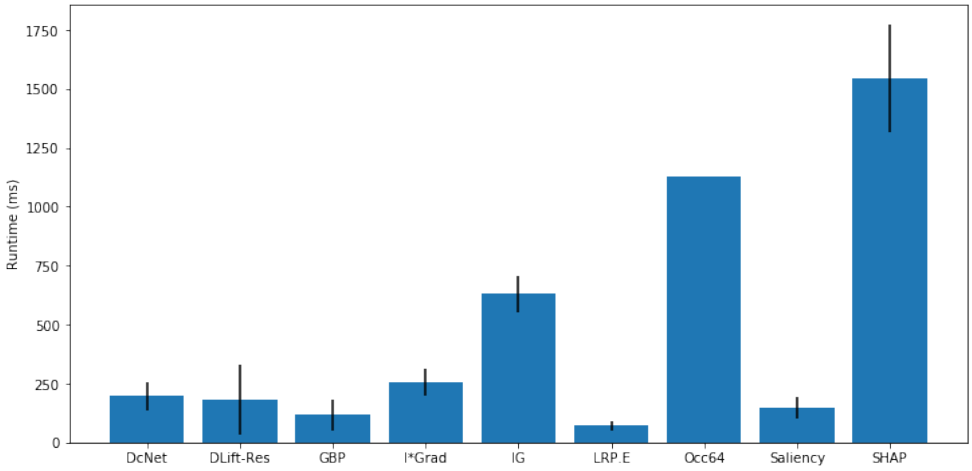


Figure 3.1: Runtime of the attribution methods with error bars showing upper and lower bounds

3.2.2 RMSE

The rootmean squared error (RMSE) was used to quantify the absolute average error of the attribution values. The difference in attribution values between various images of a disease was calculated. A lower score indicated a more robust performance. As indicated in figure 3.4, LRP had the least RMSE score and all methods but occlusion had an RMSE under 0.8 with a relatively small variation. RMSE can be affected by the number of features

highlighted by each method, and hence methods like DeConvNet which highlight more features tend to be less robust. As can be seen in figure 2.3.3, highlighting more pixels indicated either marking of areas outside the retina or better coverage.

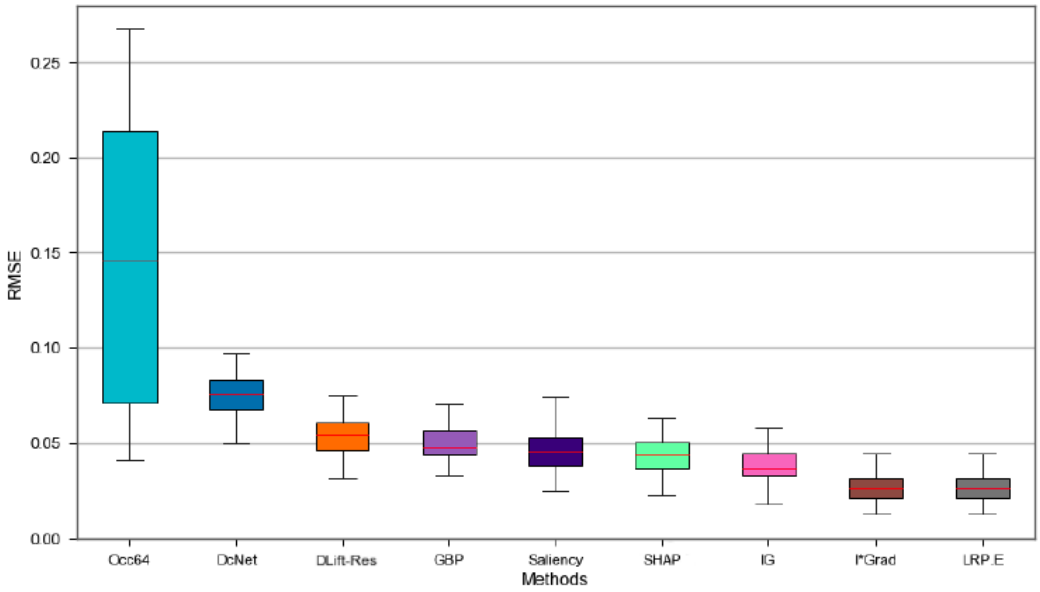


Figure 3.2: RMSE of the output of each method

3.2.3 Spearman's rank correlation

The Spearman's rank correlation was used to measure the agreement between the magnitude of attribution values and their relative effect on the model output. A model with higher values showing more impact on the output would have a higher Spearman's rank correlation score. Similar to the previous measure, LRP had the best score, closely followed by input×gradient. Occlusion, saliency maps, and DeConvNet performed poorly which agreed with their noisy attribution maps in figure 2.3.3. The mediocre performance of SHAP despite highlighting the pixels well could be due to smaller highlighted areas and giving high attribution values to most of the marked pixels. Interestingly, GBP did not perform as well on this metric despite one of the more clinically relevant markings. Hence, we looked at these methods qualitatively to observe the overlap of the heatmaps with clinical markings.

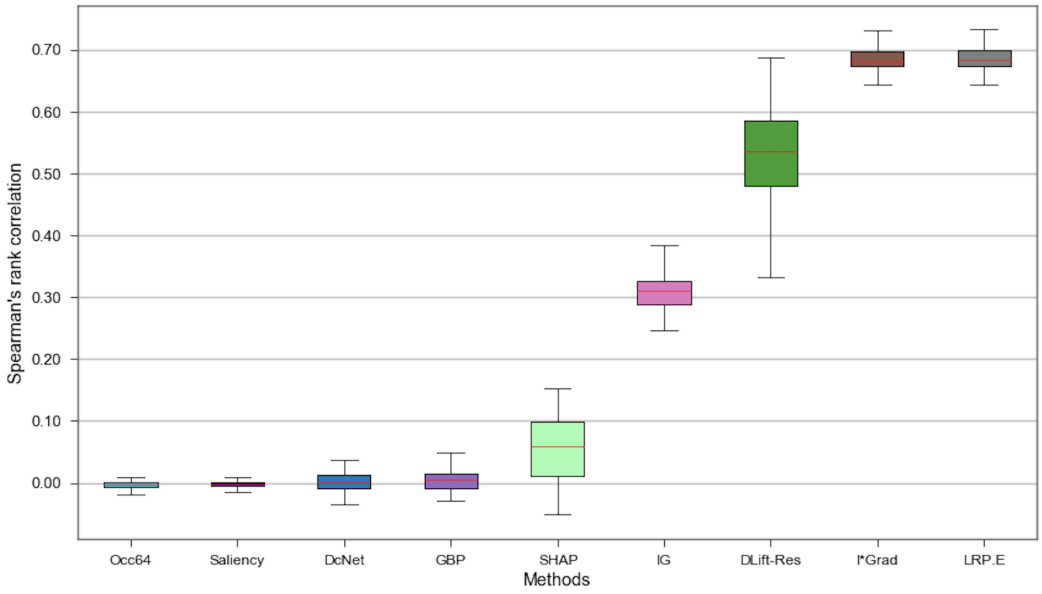


Figure 3.3: Spearman rank correlation showing agreement between methods and output

3.3 Stage 2 analysis

The results from metrics in the first stage were insufficient to discern the methods. The RMSE suffered from the displacement of the images in the scans and favored the methods with smaller highlighted areas. Spearman's rank correlation scores were not attuned with expected results from visual inspection, for example, GBP was strongly penalized despite good heatmaps. Hence, in a later study [104], two stronger metrics from recent literature were evaluated for all the methods. The robustness of a given method between the trained model weights [100] and sensitivity analysis [84], [139] were performed to compare the various attribution methods. All 1000 images of the test set were used for the robustness analysis to achieve better estimates while 80 images (20 per class) were used for sensitivity analysis due to computational constraints.

3.3.1 Robustness between models and runtime

The RMSE between the attributions of a method from all pairs of 10 separately trained instances of the model was used as a measure of robustness. Ideally, the models would

Table 3.1: RMSE between the attributions for different model instances and average runtime

Method	**RMSE**					**Avg. runtime**
	CNV	**DME**	**Drusen**	**Normal**	**Total**	**(LRP - ϵ base)**
DeconvNet	424.74	415.37	400.05	465.75	1705.91	2.70x
Deep Taylor	198.06	211.55	187.24	211.77	808.62	1.58x
DeepLIFT - Rescale	**79.49**	**70.80**	93.72	**64.77**	**308.77**	2.51x
Gradient	493.56	457.87	432.02	438.49	1821.93	1.63x
GBP	267.25	277.97	240.25	285.48	1070.95	1.61x
Input $\times$ Gradient	392.54	378.22	343.07	371.32	1485.15	3.53x
IG	368.8	347.24	311.86	346.18	1374.07	8.70x
LRP - ϵ	392.34	378.09	342.89	371.18	1484.50	**1x (72.44ms)**
Occlusion 64	196.36	306.85	441.94	598.44	1543.59	15.53x
Saliency	107.67	86.60	112.78	84.59	391.64	2.03x
SHAP - Random	117.19	85.89	**93.29**	65.83	362.20	21.30x
SHAP - Selected	122.48	75.08	99.41	63.90	360.87	15.69x
SmoothGrad	465.35	429.34	405.97	409.66	1710.32	5.04x

have learned similar features for all the runs and the attribution methods would, therefore, provide similar results. However, the stochastic nature of model training and the algorithmic differences between the attribution methods leads to non-zero RMSE values as shown in table 3.1. The DeepLIFT rescale rule had the least RMSE followed by SHAP selected, while SHAP random gave similar to that of SHAP selected. SHAP random had slightly better results for drusen as it highlighted smaller areas as shown in figure 2.4. The gradient had the highest RMSE as it is directly influenced by the variation in the model's features. It should be noted that this analysis inherently favored the methods highlighting a smaller area and was affected by the difference in distributions of attributions despite normalizing them from -1 to 1. The code was run on an Intel Gold 6148 Skylake 2.4 GHz processor with 16GB RAM and Nvidia Tesla V100 16GB GPU to benchmark the runtimes. LRP had the least runtime while SHAP random had the most due to high computation cost incurred by having a background of normal images. However, using a selected background with lower artifacts reduced the runtime of SHAP by 26.33%.

3.3.2 Sensitivity analysis

Sensitivity does not suffer the pitfalls of robustness and is a better indicator of the top features identified by an attribution method. The pixels in the original image were ranked by their attribution value and removed sequentially by setting them to 0. The value of a pixel provided its relative importance and was expected to have a positive correlation to its contribution to the output. The faster the drop in target neuron value on eliminating the top pixels, the better a method was able to rank the most important pixels and hence more sensitive to the output of the target neuron [84], [140]. The analysis was performed for the top 20% of the features of the same weights.

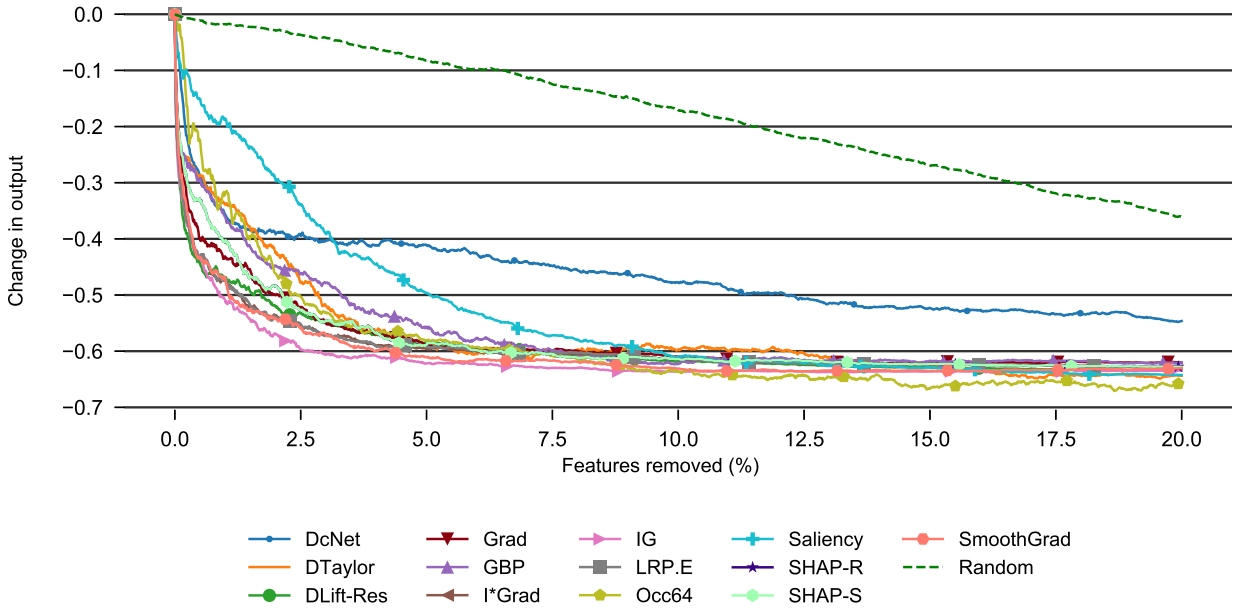


Figure 3.4: Sensitivity analysis by removing the top features of each attribution map and observing the effect on the output neuron. The methods with lower curves identify the relevant features better. The random selection of features shows a linear effect.

The removal of the features identified by all the methods showed a similar exponentially decreasing behavior. Due to the small area of pathology, it resulted in asymptotic curves beyond the 10% mark as shown in figure 3.4. The initial drop was fastest in DeepLIFT and IG but IG continued to be most sensitive till about 10% of top features beyond which occlusion 64 had the most sensitivity. The rapid decrease in model output showed that the features influencing the model ouput the most received the highest attribution values.

Saliency and deconvnet had the worst performance which is also reflected by their noisy heatmaps as shown in fig. 3.4. All the methods performed significantly better than the linear curve obtained by randomly dropping the features.

3.4 Discussion

The quantitative analysis performed here measured the ability of all the attribution methods to find the features that had the most effect on the model output. This study compared 13 different attribution methods for explaining a deep learning model for retinal OCT classification. The quantitative comparison showed high robustness between the models for DeepLIFT and SHAP while IG had marginally more sensitivity for detecting the features that impacted the decision the most.

The quantitative analysis, though important for measuring the ability to detect the features that impact the model decisions, may disagree with the visual inspection of the highlighted pathological regions. In a clinical context, it is essential to identify the features that cover the pathology the best. To this end, the next chapter describes the comparison with markings of clinicians and the ratings by clinicians for different methods.

Chapter 4

Qualitative evaluation of attribution methods

With additional content from:

- **A Singh, J J Balaji, V Jayakumar, M A Rasheed, R Raman, V Lakshminarayanan "Quantitative and Qualitative Evaluation of Explainable Deep Learning Methods for Ophthalmic Diagnosis." arXiv preprint arXiv:2009.12648 (2020).**

4.1 Introduction

As discussed earlier, the qualitative analysis provides an evaluation of the methods by the end-users. A method providing explanations that are both quantitatively sound and closer to the regions analyzed by an expert is likely to have more trust and acceptance. In this section, two qualitative analysis measures - agreement with clinician's markings and ratings from a panel of clinicians are demonstrated. The latter was performed as a pilot study with 3 optometrists initially. It was further expanded to include a group of 14 clinicians (both optometrists and ophthalmologists specializing in the retina) for a more diverse opinion. More details regarding the same are available in [135].

Table 4.1: Average cosine similarity between the heatmap and the clinical grading

Method	CNV	DME	Drusen	Average
DeconvNet	0.3503	0.3253	0.2142	0.2966
DeepLIFT	0.3701	0.2541	0.2763	0.3002
GBP	0.4297	0.3889	**0.3697**	0.3961
Input × Gradient	0.3784	0.2949	0.2779	0.3170
IG	0.2886	0.2450	0.3264	0.2867
LRP - ϵ	0.3684	0.3520	0.2827	0.3344
Occlusion	**0.6052**	0.4170	0.3401	0.4541
Saliency	0.5818	**0.4187**	0.3680	**0.4562**
SHAP	0.3322	0.3208	0.3096	0.3209

4.2 Agreement with clinical markings

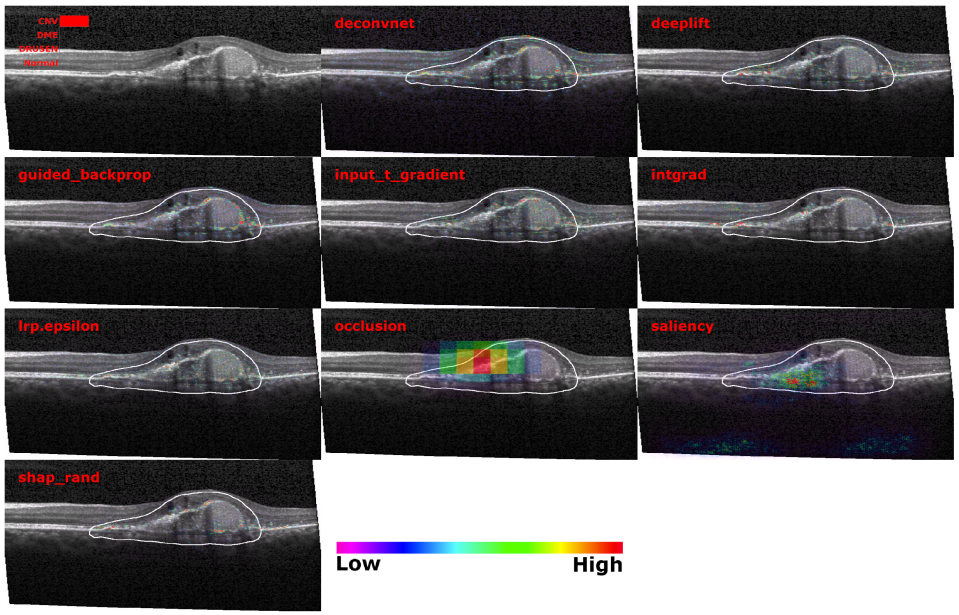
In this subsection, the heatmaps of different methods are compared to the markings done by a clinician. An optometrist with 4 years of experience with retinal OCT images graded 10 images from each disease and marked the pathological areas. This was done using ImageJ on a Microsoft Surface Pro 4 tablet with a stylus. Figure 2.3.3 shows the original images with class probabilities and heatmaps of each attribution method against the clinician's grading. The examples for CNV, DME, drusen, and drusen misclassified as CNV are presented in the figure 4.1. Unlike the figures in chapter 2, here negative values were truncated and colored heatmaps were produced on the basis of attribution scores. The number and size of the drusen present in the scan could have influenced the model to predict is as CNV in figure 2.5. It also showed some degree of confusion in the probability between CNV and drusen in figure 2.4 as well as other images in the data as shown in figure 2.1.

The cosine similarity [1] was used to measure the similarity between the heatmaps and the gradings as shown in table 4.1. It is a metric used to compare two vectors by computing their cosine angle in a multi-dimensional space. It was observed that all the methods except occlusion produced heatmaps of regions which were smaller parts within the graded area. Saliency also highlighted relatively larger regions and had more overlap with the gradings than occlusion. This led to the highest average cosine score of 0.4582 along with the highest for DME and the second-highest for CNV. The occlusion had the highest cosine score for

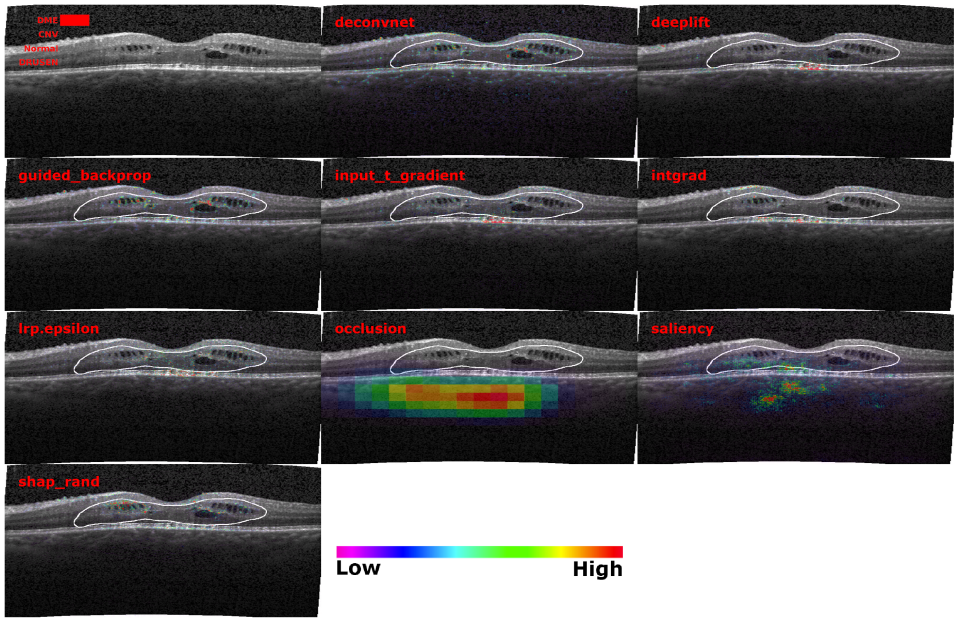
[1]Cosine similarity measures the similarity between two vectors in terms of the cosine of the angle between them

CNV as it is the pathology covering the largest area in the scans. GBP had the highest score for the smallest pathology, i.e. drusen, and had the second-highest score for DME despite covering more targetted regions. The effect is further pronounced for SHAP which highlighted even more focussed parts of the pathology structure leading to low cosine score despite clinically relevant heatmaps.

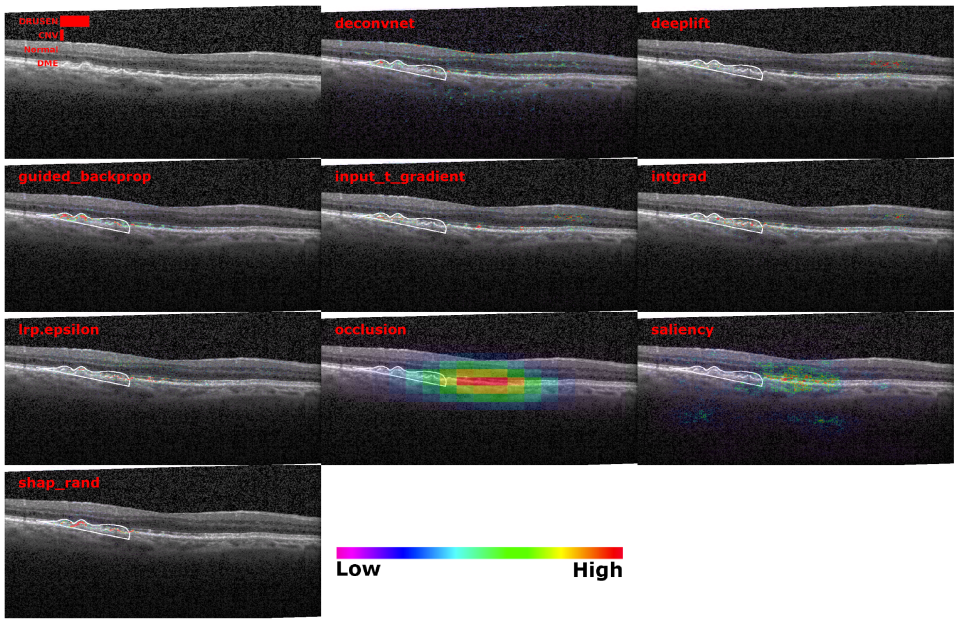
The rest of the methods were somewhat useful as they highlighted both within and outside the gradings. Notably, IG which was previously used in literature for both fundus and OCT images had a mediocre result which means comparing a wider range of methods before selecting one for deployment in real-world is important. Overall, GBP performed the best qualitatively despite having the third-highest cosine score as it highlighted in a more targeted way. SHAP worked notably better for drusen as it effectively highlighted even smaller regions.


CNV
DME
DRUSEN
NORMAL
deconvnet
deeplift
guided_backprop
input_t_gradient
intgrad
lrp.epsilon
occlusion
saliency
shap_rand
Low
High

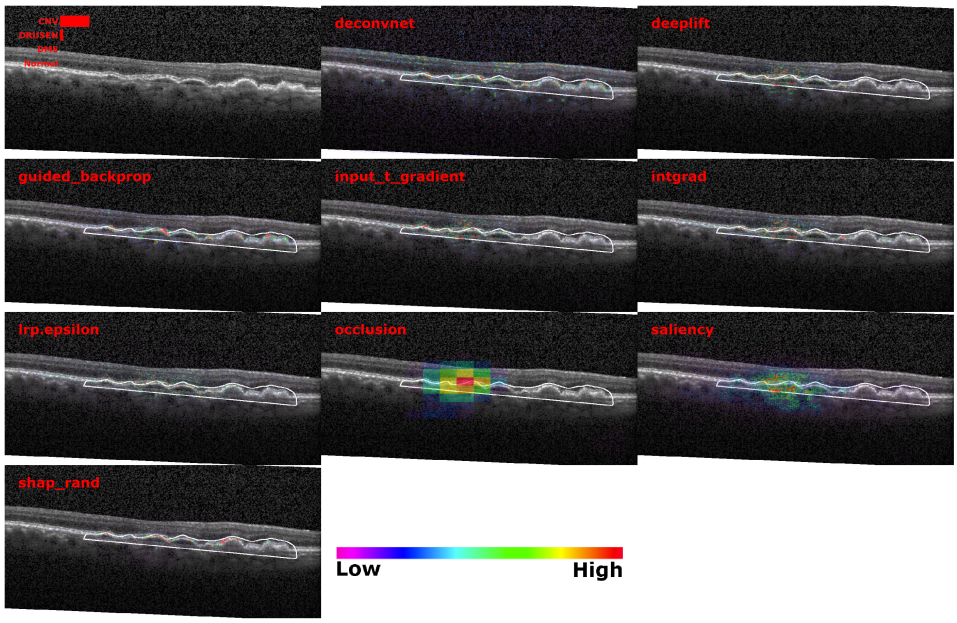
(a) CNV


DME
CNV
Normal
DRUSEN
deconvnet
deeplift
guided_backprop
input_t_gradient
intgrad
lrp.epsilon
occlusion
saliency
shap_rand
Low
High

(b) DME



(c) Drusen



(d) Drusen classified as CNV

Figure 4.1: Input images of the classes with output probabilities and heatmaps of the attribution methods. Clinician's markings are shown in white.

Table 4.2: Statistics of ratings for all data and the best rated method - Deep Taylor

Rater	Mean all	Median all	Mean best	Median best	Spearman ρ
P1	1.30	1	3.15	3	2: 0.11 3: 0.51
P2	2.30	2	4.28	4	1: 0.11 3: 0.22
P3	2.33	2	4.78	5	1: 0.51 2: 0.22

4.3 Ratings by 3 clinicians

This study was part of [104] and served as the foundation of a larger study involving a partnership with leading ophthalmologists who specialize in retinal diseases. Three expert clinicians with different levels of experience in making diagnoses from OCT rated the explanations from all the attribution methods for 20 images from each of the 3 disease classes from a scale of 0 to 5 with 0 indicating no clinical significance. P1 has clinical optometry experience of more than 25 years and has imaged and reviewed around 2500 retinal images in the last 5 years. P2 is an optometrist with 4 years of clinical experience. P3 has over several years of clinical experience as an ophthalmologist and now as an optometrist.

Fig. 4.2 shows box plots of the ratings given to explanations of different methods. To adjust for harshness, each clinician's ratings are normalized by the respective average and then the minimum of all clinicians' ratings is added to them. It was observed that the clinicians preferred Deep Taylor with a mean rating of 4.42 due to clinically coherent explanations, better coverage of pathology, and lack of high-frequency noise. GBP had a mean of 3.79 while SHAP-selected had marginally better mean of 2.85 compared to 2.81 of SHAP-random. LRP, IG, and input×gradient have a consistent but mediocre rating of around 2, and occlusion performed the worst as expected. It was observed that the ratings of methods changed over pathologies, e.g. SHAP performed close to Deep Taylor for detecting relatively small drusen deposits as it highlighted smaller areas.

It must be noted that there were differences in rating preferences between the clinicians as shown in table 4.2. The clinical experience profile might have some influences on grading OCT especially in the absence of written criteria. P1, with the most experience in OCT grading, gave a lower rating to the methods and gave a more accurate diagnosis. Spearman's rank-order correlation indicated a strong correlation between the ratings of P1 and P3 despite the difference in mean and median values.

The clinicians pointed out several differences between these methods and actual areas of relevance. It was observed that some low rated methods highlighted the vitreous regions

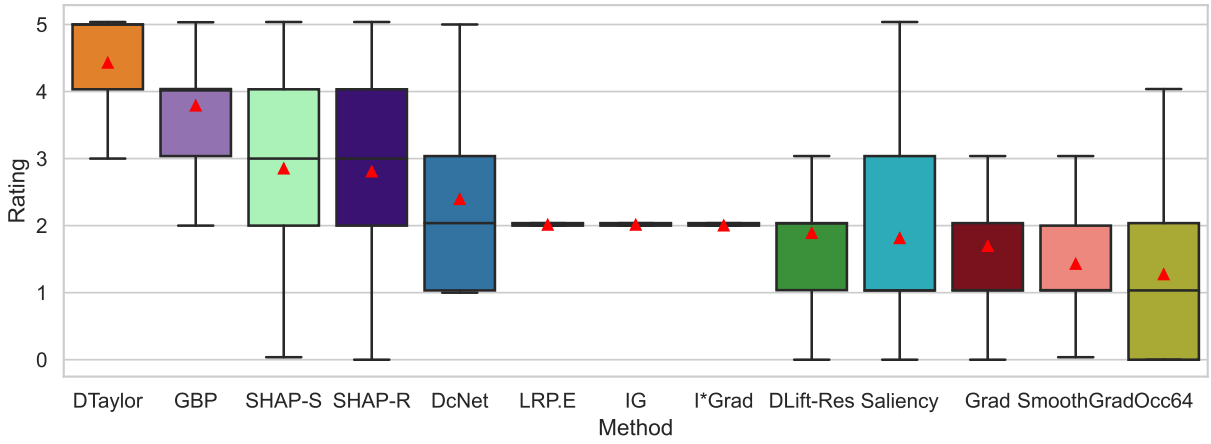


Figure 4.2: The box plots of the normalized ratings of the clinicians for explanations of different methods sorted by mean (red marker). Deep Taylor and GBP had the high mean and short whiskers indicating consistently good ratings.

outside the retina as explanations. This is due to the system's lack of awareness about the bounds of the retinal region. P1 pointed out that CNV was the hardest to explain for the methods while drusen had the most consistent results. P2 and P3 highlighted discrepancies in the source data such as lack of information about the cross-sectional plane of scans and incomplete view of the affected region. The clinicians also found secondary diagnosis for 4 out of the 60 images indicating potential noise or confounds in the source data.

4.4 Rating by a panel of 14 clinicians

The study in the previous section led to further evaluation of the methods for their suitability for clinical use. A panel of 14 clinicians was set up including 10 ophthalmologists and 4 optometrists [135]. The heatmaps generated by the 13 methods for 20 images from each disease category were evaluated by the clinicians. The group had a median experience of 5 years in retinal diagnosis, including 4 years with OCT imaging. The average number of images rated per week was approximately 40 with all the clinicians having prior experience analyzing retinal SD-OCT images. They rated the explanations from 0 (not relevant) to 5 (fully relevant). The scores of each clinician were normalized by subtracting the respective mean and then rescaling between 0 to 5. A comparison between the scores of the raters was also performed in addition to the comparison between methods. The clinicians also

provided qualitative feedback in addition to the ratings.

4.4.1 Comparison between methods

One way to represent the rating scores are through the use of violin plots. In these figures, the estimated probability density of each method is shown by the thickness of the violin plot. The plots of normalized scores of raters for all the methods across 60 scans are shown in figure 4.3. Table 4.3 gives the rating data for all conditions and methods. Deep Taylor with the highest median rating of 3.85 was judged as the best performing method. It had the highest ratings for all conditions as shown in table 4.3. It is relatively simple to compute and involves Taylor series expansion of the signal at neurons. It was considerably ahead of GBP, the next best method which was closely followed by SHAP with selected and then random background.

Table 4.3: Median ratings (with IQR) for each disease for all attribution methods. Deep Taylor had the highest ratings.

Method	**Median rating (IQR)**			
	CNV	**DME**	**Drusen**	**Total**
DcNet	2.17 (1.71-2.61)	2.47 (1.74-3.09)	2.32 (1.71-2.61)	2.32 (1.71-2.82)
DTaylor	**3.80 (3.22-4.05)**	**3.48(3.09-3.99)**	**3.99 (3.58-4.56)**	**3.85 (3.23-4.07)**
DLift-Res	2.44 (1.85-2.72)	2.44 (1.96-2.53)	2.53 (2.32-3.09)	2.47 (2.06-2.82)
Grad	2.32 (1.77-2.53)	2.47 (2.19-2.95)	2.44 (2.03-2.61)	2.44 (1.96-2.72)
GBP	3.23 (3.09-3.80)	3.26 (3.07-3.80)	3.71 (3.22-3.99)	3.29 (3.09-3.97)
I*Grad	2.50 (2.32-2.95)	2.47 (2.28-2.82)	2.53(2.44-3.04)	2.50 (2.32-2.95)
IG	2.50 (2.32-2.95)	2.47 (2.19-2.82)	2.57 (2.44-3.20)	2.50 (2.32-2.95)
LRP.E	2.50 (2.32-2.95)	2.50 (2.32-2.95)	2.53 (2.41-3.04)	2.50 (2.32-2.95)
LRP.Z	2.50 (2.32-2.95)	2.50 (2.32-2.95)	2.53 (2.41-3.04)	2.50 (2.32-2.95)
Occ64	1.71 (1.55-1.96)	1.71 (1.42-1.85)	1.71 (1.42-1.96)	1.71 (1.52-1.96)
Saliency	2.47 (1.74-3.29)	2.72 (1.74-3.29)	2.61 (1.74-3.29)	2.61 (1.74-3.29)
SHAP-R	3.23 (2.53-3.85)	3.23 (2.53-3.85)	3.58 (2.89-3.96)	3.23 (2.53-3.85)
SHAP-S	3.23 (2.53-3.85)	3.23 (2.53-3.85)	3.53 (2.61-3.96)	3.26 (2.53-3.96)
SmoothGrad	2.45 (1.85-2.95)	2.47 (1.96-3.09)	2.47 (1.85-3.04)	2.47 (1.93-3.04)

IG which is commonly employed in the literature for generating heatmaps for retinal diagnosis [52], [102] received a median score of only 2.5. It is known to be strongly related and in some cases mathematically equivalent to LRP - ϵ [84] and this was also reflected in similar ratings. DeepLIFT could not be tested in its newer Reveal Cancel rule due to

compatibility issues with the model architecture and the older Rescale rule had a below par performance. As expected, the baseline occlusion which used sliding window of size 64 to cover the pixel and then compute significance performed worse than the attribution-based methods.

Most of the methods have the majority of the values around the median indicating consistent ratings across images and raters. Both cases of SHAP and Saliency have particularly elongated distributions. For SHAP, the curve is widest around 4 indicating good ratings for many cases. However, the values around 2.5 due to lower coverage of pathology drive the overall median lower. In the case of Saliency, the ratings are spread from about 4.5 to 1.5 with many of them around 3.25 and 1.75 marks. The former is due to larger coverage of the pathological region and the latter is due to the fact that it missed regions frequently. Hence, despite better median value, it is not as suitable as lower-rated methods such as IG where a bulk of the value is around the median.

4.4.2 Comparison between raters

The Spearman's rank correlation was used to compare the ratings of the clinicians with each other. This test is a non-parametric measure that assesses the relationship between two variables, in this case the ratings of images by two different clinicians. A correlation of +1 indicates a perfect positive correlation, 0 indicates no correlation, and -1 indicates perfect negative correlation. The correlations between the ratings of all 14 clinicians for the 60 images and 13 methods are shown in figure 4.4. P1 to P10 are ophthalmologists while P11 to P14 are optometrists.

Most of the values are around 0.5 indicating an overall moderate agreement between clinicians. The highest correlation was of 0.76 between P10 and P13 while two cases of slight negative correlation were between P1 and P11 and P2 and P11. The rater P11 had relatively less experience with OCT which could have resulted in a lower correlation with other clinicians. This indicates that the background and training (i.e., prior experience) of clinicians affected their ratings of the system.

4.4.3 Qualitative observations

The positively correlation between the ratings of the methods by the clinicians indicates similar preferences between different attribution methods in a quantitative way. In this subsection the qualitative feedback given by the clinicians regarding the performance of the system, potential use cases and other suggestions are summarized. A survey was collected

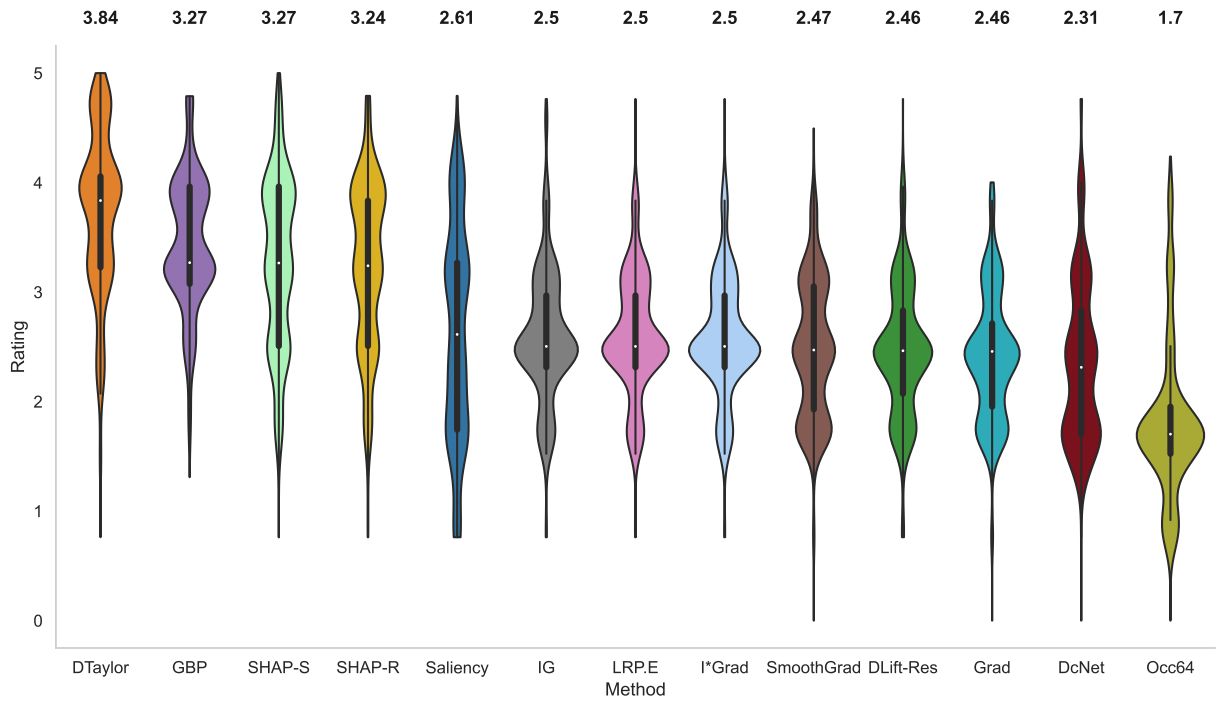


Figure 4.3: Violin plots of normalized ratings of all methods. The breadth of the plot shows the probability density of the data and the median value is reported on top of the plots. Deep Taylor was rated the highest overall followed by GBP and SHAP.

from the clinicians to seek their opinion post study. This would help understand the observations of clinicians in a more nuanced manner.

It is notable that 79% (11/14) clinicians who participated in the study would prefer to have an explainable system assisting them in practice, reaffirming the need for such system to the clinical community. One of the ophthalmologists gave their feedback on the system as – "It is a definite boon to the armamentarium as far as screening and diagnosis is concerned on a mass scale or in a telemedicine facility".

The clinicians noted an overall better coverage of the pathology by Deep Taylor as the reason for higher ratings, however it and other methods except SHAP were found to be detecting the boundaries of the regions better. SHAP was observed to be identifying regions inside the edema also, though the partial coverage of the region and to some extent

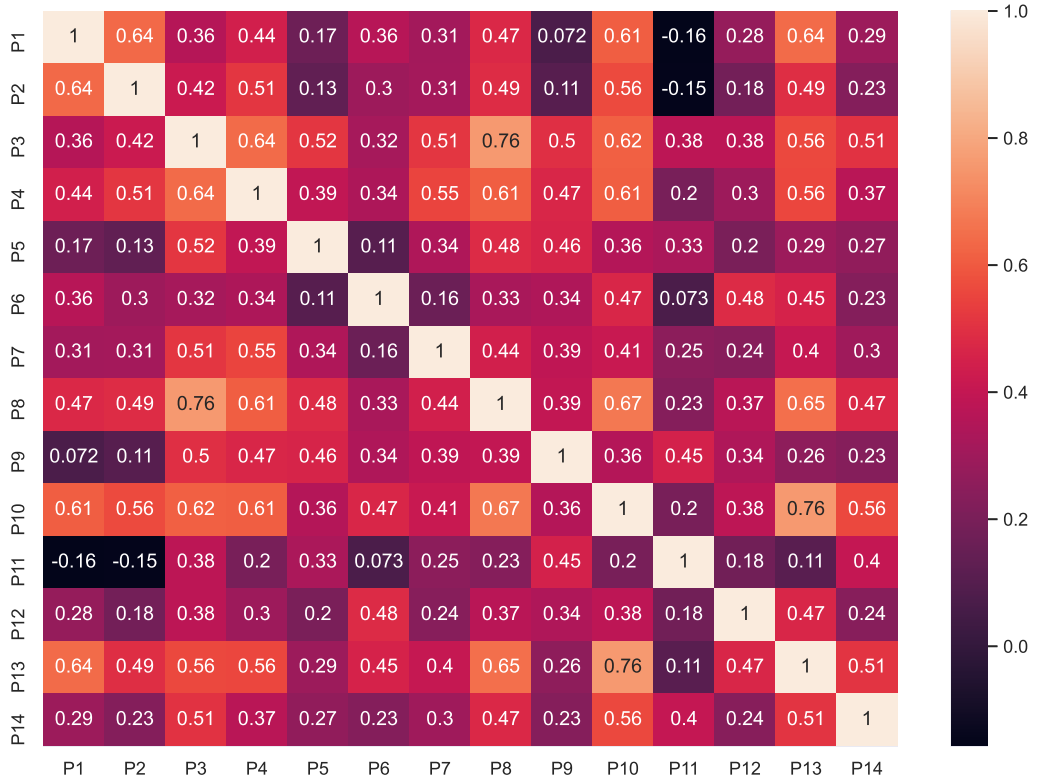


Figure 4.4: Spearman's correlation for clinician's ratings.

the noise (represented in blue) from negative attributions led to a lower score. The noise, especially in case of LRP was found to be a distraction by some clinicians. It was kept in this study to not alter the attributions in any way and compare the methods in their original form. However, it is easy to remove it by keeping only positive valued attributions.

Most of the clinicians identified telemedicine and tertiary care centres as potential sites which can utilize this system. It was suggested that it can be used for screening in places with large number of patients without sufficient number of specialists (or a tertiary care center), which was one of the initial goals of this study. It helps clinicians by categorizing the scans with suspect conditions and thus allows clinicians to focus their attention on examining the areas of the images highlighted by algorithm and hence take necessary steps towards making final decision on the diagnosis. This can improve efficiency and help in saving time, resulting in more efficient patient care. Another application could be archival

and data management where the heatmaps could be used for separating images faster.

4.5 Discussion

In addition to a comparison of various available attribution methods to explain deep learning models, this study validated their results through ratings from a large panel of clinicians. Most of them were not involved in the design process at this stage, were in general positive about the utility of the system and were and receptive to using this methodology.

A method based on Taylor series expansion, known as Deep Taylor, received the highest ratings showing that methods with stronger or better theoretical backgrounds and high performance on standard datasets may not be the optimal methods in a practical medical imaging situation. It must be noted that the original goal of these attribution methods was to explain the model's decision-making process by generating a true representation of the features used by a model to perform a given task. Hence, the heatmaps generated are affected by both by the model and the attribution method used.

It must be noted that a significant issue with GBP, the second highest rated method is that it acts as an edge detector and not actually revealing the model's decision-making process [141], [142]. Despite this issue its attributions were rated highly due to a good coverage of the pathological region, especially its boundaries. However, we suggest using explainability methods which are both technically sound and generate heatmaps to highlight the pathology to be used in clinical deployment of explainable deep learning systems.

Chapter 5

Conclusion and future research

In this thesis, different approaches to explain deep learning models for retinal diagnosis were compared. The quantitative comparison involved metrics used in the literature to compare the explainability methods on standard computer vision datasets. The results from these metrics could not clearly distinguish the methods in line with the observed images. Hence, a qualitative comparison for coverage of the clinically relevant regions was performed. A direct comparison using cosine scores with markings of a clinician improved the distinction but favored the methods covering larger areas. A comparison using ratings of each method was found to be in line with the observations and favored a relatively simple yet effective method, namely Deep Taylor. The study revealed that despite performing reasonably well, the state-of-the-art explainability methods may not be the most suitable for a specific task such as retinal diagnosis. There is a need to explain the deep learning models to improve their acceptability and using the right approach to do so is the key.

Identifying a suitable method from the choices available is a single element of the problem of explaining deep learning models for retinal diagnosis using OCT images and making them more acceptable by the clinicians and the public. Several additional hurdles need to be overcome before such a method can be deemed suitable for wider deployment. Some such directions of future research are:

- **System enhancements:**
  - *Uncertainty aware explainable system:* Deep learning methods tend to give a high softmax value as output due to the nature of the objective function used in their training. This value is often misinterpreted as a probability or confidence in the decision. However, an uncertainty analysis using methods such as

Bayesian neural networks have to be undertaken in order to create a comprehensive system. This is currently being studied in a project which combines and relates uncertainty and explainability for retinal diagnosis [143].

- *On demand explainability method choice:* Even though Deep Taylor emerged as the highest-rated method it is not the truest to the actual model features as revealed in the sensitivity analysis. SHAP displayed better results for covering the drusen area while Deep Taylor and GBP covered the RPE better. Similarly, occlusion gave prominent negative heatmaps in cases with misdiagnosis. Hence, depending on the uncertainty and the class predicted, a specific attribution method or a combination of multiple methods can be used for generating the heatmaps for the end-user.
- *Screening tool with human-in-the-loop (HITL) design:* AI systems such as the one designed in this thesis are not meant to replace human experts but to augment their abilities for delivering better service. The decisions made by a deep learning model must be approved by a human expert before using them for patient care. An explainable system with suitably trained clinicians can be used to implement large screening programs with much higher patient throughput. For this, the trust of the clinicians must be calibrated using uncertainty metrics. [144].
- *Integration with portable OCT systems:* Currently, OCT systems are not used in eye camps since the devices are not only expensive but also are bulky. Given recent advances in low-cost portable OCT devices [145], it is possible to integrate an explainable diagnosis system on a laptop or mobile device for teleophthalmology purposes, which would be invaluable to the clinical community.

- **Data and model improvements:**
  - *Using a dataset labeled for secondary diagnosis:* The dataset used here labeled only primary diagnosis. However, the clinicians were able to identify a secondary diagnosis for some images from their evaluation.
  - *Using information about the orientation of scans:* Due to the nature of the dataset, the study is limited to a single orientation of the OCT scan which might differ between the images. Using volumetric scans could train more robust models and potentially better explanations for a complete view of the retina.
  - *Integrating patient records and fundus images:* This study used only OCT image data whereas in practice multiple sources of information such as fundus image,

age, past ocular and medical history, fellow eye status, etc. An integrated AI system with electronic medical records can be used to develop a clinical decision support system. This can make early-stage detection of disease or predict prognosis. Another application of the explainability system could be as a self-learning tool. All the clinicians in this study preferred having fundus images in addition to OCT, hence, a system that uses fundus, OCT, and patient data similar to [146] could be useful in practice. A multi-modal system with all available diagnostic information - patient reports, fundus, and OCT images can improve both accuracy and explainability of diagnosis.

– *More diseases and types of imaging equipment:* The system can be developed to encompass other diseases and finetuned for the specific imaging modality, taking into account variables such as noise, illumination, field position, etc. A deep learning model for retinal OCT usually does not generalize well to the images from a device from a different manufacturer than the one it was trained on [147].

Bibliography

[1] H. A. Leopold, A. Singh, S. Sengupta, *et al.*, "Recent Advances in Deep Learning Applications for Retinal Diagnosis using OCT," in *State of the Art in Neural Networks, A. S. El-Baz (ed.)* Elsevier, NY, in press, 2020.

[2] S. Sengupta, A. Singh, H. A. Leopold, *et al.*, "Ophthalmic diagnosis using deep learning with fundus images – A critical review," *Artificial Intelligence in Medicine*, vol. 102, p. 101 758, 2020, ISSN: 18732860. DOI: 10.1016/j.artmed.2019.101758.

[3] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020. DOI: 10.3390/jimaging6060052.

[4] A. Denniston and P. Murray, *Oxford Handbook of Ophthalmology*. OUP Oxford, 2014.

[5] V Lakshminarayanan, "Ibn al Haytham: Founder of Physiological Optics?" *Light-Based Science: Technology and Sustainable Development: The Legacy of Ibn Al-Haytham*, 63–108, eds. R. Rashed and A. Bourdiroua and V. Lakshminarayanan, Chapter 6, CRC Press, Boca Raton, FL, 2017.

[6] S. E. Sherman, "The history of the ophthalmoscope," in *History of Ophthalmology*, Springer, Cham, 1989, http://doi.org/10.1007/978-94-009-2387-4_10, pp. 221–228.

[7] J Fujimoto and E Swanson, "The development, commercialization, and impact of Optical Coherence Tomography.," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 9, OCT1–OCT13, https://doi.org/10.1167/iovs.16\bibrangedash19963, 2016.

[8] C. Costagliola, R. Dell'Omo, M. R. Romano, *et al.*, "Pharmacotherapy of intraocular pressure: part I. parasympathomimetic, sympathomimetic and sympatholytics," *Expert Opinion on Pharmacotherapy*, vol. 10, no. 16, pp. 2663–2677, 2009, https://doi.org/10.1517/14656560903300103.

[9] A. Krolewski, J. Warram, L. Rand, *et al.*, "Risk of proliferative diabetic retinopathy in juvenile-onset type I diabetes: a 40-yr Follow-up Study," *Diabetes Care*, vol. 9, no. 5, pp. 443–452, 1986.

[10] H. A. Leopold, J. S. Zelek, and V. Lakshminarayanan, "Deep Learning Methods for Retinal Image Analysis," in *Biomedical signal processing in big data.* E. Sejdić and T. H. Falk, Eds. CRC Press, 2018, pp. 329–365.

[11] A. Agarwal and D. A. Kumar, *Essentials of OCT in Ocular Disease*. Thieme, 2015, https://doi.org/10.1097/OPX.0000000000000875.

[12] D. Huang, E. A. Swanson, C. P. Lin, *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[13] F. C. Delori, E. S. Gragoudas, R. Francisco, *et al.*, "Monochromatic ophthalmoscopy and fundus photography: The normal fundus," *Archives of ophthalmology*, vol. 95, no. 5, pp. 861–868, 1977.

[14] R. H. Webb and G. W. Hughes, "Scanning laser ophthalmoscope," *IEEE Transactions on Biomedical Engineering*, no. 7, pp. 488–492, 1981.

[15] J. I. Morgan, "The fundus photo has met its match: Optical coherence tomography and adaptive optics ophthalmoscopy are here to stay," *Ophthalmic and Physiological Optics*, vol. 36, no. 3, pp. 218–239, 2016, https://doi.org/10.1111/opo.12289.

[16] H. G. Bezerra, M. A. Costa, G. Guagliumi, *et al.*, "Intracoronary optical coherence tomography: A comprehensive review: Clinical and research applications," *JACC: Cardiovascular Interventions*, vol. 2, no. 11, pp. 1035–1046, 2009.

[17] R Leitgeb, C. Hitzenberger, and A. F. Fercher, "Performance of fourier domain vs. time domain optical coherence tomography," *Optics Express*, vol. 11, no. 8, pp. 889–894, 2003.

[18] J. F. De Boer, C. K. Hitzenberger, and Y. Yasuno, "Polarization sensitive optical coherence tomography–a review," *Biomedical Optics Express*, vol. 8, no. 3, pp. 1838–1873, 2017.

[19] B. Baumann, "Polarization sensitive optical coherence tomography: A review of technology and applications," *Applied Sciences*, vol. 7, no. 5, p. 474, 2017, https://doi.org/10.3390/app7050474.

[20] S. Kishi, "Impact of swept source optical coherence tomography on ophthalmology," *Taiwan Journal of Ophthalmology*, vol. 6, no. 2, pp. 58–68, 2016.

[21] M. Bhende, S. Shetty, M. K. Parthasarathy, *et al.*, "Optical coherence tomography: A guide to interpretation of common macular diseases," *Indian Journal of Ophthalmology*, vol. 66, no. 1, pp. 20–35, 2018.

[22] M. Pircher, C. K. Hitzenberger, and U. Schmidt-Erfurth, "Polarization sensitive optical coherence tomography in the human eye," *Progress in Retinal and Eye Research*, vol. 30, no. 6, pp. 431–451, 2011.

[23] M. T. Nicolela and J. R. Vianna, "Optic nerve: Clinical examination," in *Pearls of Glaucoma Management*, Springer, Belin, 2016, pp. 17–26.

[24] R. D. Jager, W. F. Mieler, and J. W. Miller, "Age-related macular degeneration," *New England Journal of Medicine*, vol. 358, no. 24, pp. 2606–2617, 2008.

[25] D. S. Friedman, B. J. O'Colmain, B. Munoz, *et al.*, "Prevalence of age-related macular degeneration in the united states," *Arch ophthalmol*, vol. 122, no. 4, pp. 564–572, 2004.

[26] P Gholami, P Roy, M. K. Parthasarathy, *et al.*, "OCTID: Optical Coherence Tomography Image Database," *arXiv preprint arXiv:1812.07056*, 2018.

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

[28] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[29] C. Szegedy, S. Ioffe, V. Vanhoucke, *et al.*, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[30] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Glaucoma diagnosis using transfer learning methods," in *In Proc. Applications of Machine Learning, SPIE*, International Society for Optics and Photonics (SPIE), vol. 11139, 2019, 111390U.

[31] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[32] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural Networks for Perception*, IEEE, Washington, DC, 1992, 65–93, doi: 10.1109/IJCNN.1989.118638.

[33] H. Muhammad, T. J. Fuchs, N. De Cuir, *et al.*, "Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects," *Journal of Glaucoma*, vol. 26, no. 12, pp. 1086–1094, 2017.

[34] H. Fu, Y. Xu, S. Lin, *et al.*, "Multi-context deep network for angle-closure glaucoma screening in anterior segment oct," in *International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2018. Lecture Notes in Computer Science eds. Frangi A., Schnabel J., Davatzikos C., Alberola-López C., Fichtinger G*, Springer, vol. 11071, 2018, pp. 356–363.

[35] O. Perdomo, H. Rios, R. Francisco, *et al.*, "Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography," *Computer Methods and Programs in Biomedicine*, vol. 178, 181–189, doi: 10.1016/j.cmpb.2019.06.016, 2019.

[36] V. Das, S. Dandapat, and P. K. Bora, "Multi-scale deep feature fusion for automated classification of macular pathologies from oct images," *Biomedical Signal Processing and Control*, vol. 54, p. 101 605, 2019.

[37] X. Liu, T. Fu, Z. Pan, *et al.*, "Automated layer segmentation of retinal optical coherence tomography images using a deep feature enhanced structured random forests classifier," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1404–1416, 2019, ISSN: 2168-2208. DOI: 10.1109/JBHI.2018.2856276.

[38] D. S. Kermany, M. Goldbaum, W. Cai, *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[39] M. Treder, J. L. Lauermann, and N. Eter, "Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 256, no. 2, pp. 259–265, 2018.

[40] S. Saha, M. Nassisi, M. Wang, *et al.*, "Automated detection and classification of early amd biomarkers using deep learning," *Scientific reports*, vol. 9, no. 1, p. 10 990, 2019.

[41] A. Varadarajan, P. Bavishi, P. Raumviboonsuk, *et al.*, "Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning," *arXiv preprint arXiv:1810.10342*, 2018.

[42] L. Huang, X. He, L. Fang, *et al.*, "Automatic classification of retinal optical coherence tomography images with layer guided convolutional neural network," *IEEE Signal Processing Letters*, vol. 26, no. 7, pp. 1026–1030, 2019.

[43] H. Fu, Y. Xu, S. Lin, *et al.*, "Angle-closure detection in anterior segment oct based on multilevel deep network," *IEEE Transactions on Cybernetics*, 1–9, doi: 10.1109/TCYB.2019.2897162, 2019.

[44] A. Vahadane, A. Joshi, K. Madan, *et al.*, "Detection of diabetic macular edema in optical coherence tomography scans using patch based deep learning," in *Biomedical Imaging (ISBI 2018), IEEE 15th International Symposium on*, IEEE, 2018, pp. 1427–1430.

[45] O. Perdomo, S. Otálora, F. A. González, *et al.*, "Oct-net: A convolutional network for automatic classification of normal and diabetic macular edema using sd-oct volumes," in *Biomedical Imaging (ISBI 2018), IEEE 15th International Symposium on*, IEEE, 2018, pp. 1423–1426.

[46] A. ElTanboly, M. Ghazaf, A. Khalil, *et al.*, "An integrated framework for automatic clinical assessment of diabetic retinopathy grade using spectral domain oct images," in *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, 2018, pp. 1431–1435.

[47] L. Fang, D. Cunefare, C. Wang, *et al.*, "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search," *Biomedical optics express*, vol. 8, no. 5, pp. 2732–2744, 2017.

[48] G. C. Chan, R. Kamble, H. Müller, *et al.*, "Fusing results of several deep learning architectures for automatic classification of normal and diabetic macular edema in optical coherence tomography," in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2018, pp. 670–673.

[49] R Rasti, H Rabbani, A Mehridehnavi, *et al.*, "Macular oct classification using a multi-scale convolutional neural network ensemble," *IEEE Transactions on Medical Imaging*, vol. 37, no. 4, pp. 1024–1034, 2018.

[50] R. Rasti, A. Mehridehnavi, H. Rabbani, *et al.*, "Convolutional mixture of experts model: A comparative study on automatic macular diagnosis in retinal optical coherence tomography imaging," *Journal of Medical Signals and Sensors*, vol. 9, no. 1-14, p. 1, 2019.

[51] S. Athar, A. Vahadane, A. Joshi, *et al.*, "Weakly supervised fluid filled region localization in retinal oct scans," in *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, 2018, pp. 1467–1470.

[52] H.-L. Yang, J. J. Kim, J. H. Kim, *et al.*, "Weakly supervised lesion localization for age-related macular degeneration detection using optical coherence tomography images," *PloS One*, vol. 14, no. 4, e0215076, 2019.

[53] S. Maetschke, B. Antony, H. Ishikawa, *et al.*, "A feature agnostic approach for glaucoma detection in oct volumes," *PloS One*, vol. 14, no. 7, e0219126, 2019.

[54] R. Xu, S. Niu, K. Gao, *et al.*, "Multi-path 3d convolution neural network for automated geographic atrophy segmentation in sd-oct images," in *International Conference on Intelligent Computing*, Springer, 2018, pp. 493–503.

[55] O. J. Perdomo, H. A. Rios, F. J. Rodríguez, *et al.*, "3d deep convolutional neural network for predicting neurosensory retinal thickness map from spectral domain optical coherence tomography volumes," in *Proc SPIE.*, International Society for Optics and Photonics, vol. 10975, 2018, p. 109750I.

[56] T. Kurmann, P. Márquez-Neila, S. Yu, *et al.*, "Fused detection of retinal biomarkers in oct volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 255–263.

[57] L. Fang, C. Wang, S. Li, *et al.*, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1959–1970, 2019.

[58] A. Ben-Cohen, D. Mark, I. Kovler, *et al.*, "Retinal layers segmentation using fully convolutional network in oct images," *RSIP Vision*, 2017.

[59] J. Kugelman, D. Alonso-Caneiro, S. A. Read, *et al.*, "Automatic choroidal segmentation in OCT images using supervised deep learning methods," *Scientific Reports*, vol. 9, no. 1, pp. 1–13, 2019.

[60] V. A. Dos Santos, L. Schmetterer, H. Stegmann, *et al.*, "Corneanet: Fast segmentation of cornea oct scans of healthy and keratoconic eyes using deep learning," *Biomedical Optics Express*, vol. 10, no. 2, pp. 622–641, 2019.

[61] M Pekala, N Joshi, D. E. Freund, *et al.*, "Deep Learning based Retinal OCT Segmentation," *arXiv preprint arXiv:1801.09749*, 2018.

[62] X. Liu, L. Bi, Y. Xu, *et al.*, "Robust deep learning method for choroidal vessel segmentation on swept source optical coherence tomography images," *Biomedical Optics Express*, vol. 10, no. 4, pp. 1601–1612, 2019.

[63] L. Varga, A. Kovács, T. Grósz, *et al.*, "Automatic segmentation of hyperreflective foci in oct images," *Computer Methods and Programs in Biomedicine*, vol. 178, pp. 91–103, 2019.

[64] D. Lu, M. Heisler, S. Lee, *et al.*, "Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network," *Medical Image Analysis*, vol. 54, pp. 100–110, 2019.

[65] R. Asgari, J. I. Orlando, S. Waldstein, *et al.*, "Multiclass segmentation as multitask learning for drusen segmentation in retinal optical coherence tomography," *arXiv preprint arXiv:1906.07679*, 2019.

[66] A. G. Roy, S. Conjeti, S. P. K. Karri, *et al.*, "Relaynet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical Optics Express*, vol. 8, no. 8, pp. 3627–3642, 2017.

[67] S. K. Devalla, G. Subramanian, T. H. Pham, *et al.*, "A deep learning approach to denoise optical coherence tomography images of the optic nerve head," *arXiv preprint arXiv:1809.10589*, 2018.

[68] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[69] *The discriminator — generative adversarial networks*. [Online]. Available: https://developers.google.com/machine-learning/gan/discriminator.

[70] Y Ma, X Chen, W Zhu, *et al.*, "Speckle noise reduction in optical coherence tomography images based on edge-sensitive cGAN," *Biomedical optics express*, vol. 9, no. 11, pp. 5129–5146, 2018.

[71] X Zha, F Shi, Y Ma, *et al.*, "Generation of retinal OCT images with diseases based on cGAN," in *proc. SPIE*, International Society for Optics and Photonics, vol. 10949, 2019, p. 1 094 924.

[72] R. Tennakoon, A. K. Gostar, R. Hoseinnezhad, *et al.*, "Retinal fluid segmentation in oct images using adversarial loss based convolutional neural networks," in *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, 2018, pp. 1436–1440.

[73] Y Huang, Z Lu, Z Shao, *et al.*, "Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network," *Optics express*, vol. 27, no. 9, pp. 12 289–12 307, 2019.

[74] T. Jo, K. Nho, and A. J. Saykin, "Deep learning in alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data," *Frontiers in aging neuroscience*, vol. 11, p. 220, 2019.

[75] K.-L. Hua, C.-H. Hsu, S. C. Hidayati, *et al.*, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," *OncoTargets and therapy*, vol. 8, 2015.

[76] A. Holzinger, C. Biemann, C. S. Pattichis, *et al.*, "What do we need to build explainable AI systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[77] M. Stano, W. Benesova, and L. S. Martak, "Explainable 3d convolutional neural network using gmm encoding," in *in Proc. of Twelfth International Conference on Machine Vision (ICMV 2019)*, Proc. SPIE, vol. 11433, 2020, 114331U.

[78] R. Meyes, C. W. de Puiseau, A. Posada-Moreno, *et al.*, "Under the hood of neural networks: Characterizing learned representations by functional neuron populations and network ablations," *arXiv preprint arXiv:2004.01254*, 2020.

[79] A. B. Arrieta, N. D′iaz-Rodr′iguez, J. Del Ser, *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[80] G. Stiglic, P. Kocbek, N. Fijacko, *et al.*, "Interpretability of machine learning based prediction models in healthcare," *arXiv preprint arXiv:2002.08596*, 2020.

[81] V. Arya, R. K. Bellamy, P.-Y. Chen, *et al.*, "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," *arXiv preprint arXiv:1909.03012*, 2019.

[82] M. Alber, S. Lapuschkin, P. Seegerer, *et al.*, "iNNvestigate neural networks," *Journal of Machine Learning Research*, vol. 20, no. 93, pp. 1–8, 2019.

[83] G. Montavon, S. Lapuschkin, A. Binder, *et al.*, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[84] M. Ancona, E. Ceolini, C. Öztireli, *et al.*, "Towards better understanding of gradient-based attribution methods for deep neural networks," *arXiv preprint arXiv:1711.06104*, 2017.

[85] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *in Proc. European Conference on Computer Vision*, Springer, Cham, 2014, pp. 818–833.

[86] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[87] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach," *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, pp. 319–330, 2001.

[88] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[89] J. T. Springenberg, A. Dosovitskiy, T. Brox, *et al.*, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[90] S. Bach, A. Binder, G. Montavon, *et al.*, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, 2015. DOI: 10.1371/journal.pone.0130140.

[91] A. Shrikumar, P. Greenside, A. Shcherbina, *et al.*, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.

[92] R. R. Selvaraju, M. Cogswell, A. Das, *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *in Proc. IEEE International Conference on Computer Vision*, 2017, pp. 618–626. DOI: 10.1007/s11263-019-01228-7.

[93] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *34th International Conference on Machine Learning, ICML 2017*, JMLR. org, vol. 70, 2017, pp. 5109–5118, ISBN: 9781510855144. arXiv: 1703.01365.

[94] P.-J. Kindermans, K. T. Schütt, M. Alber, *et al.*, "Learning how to explain neural networks: Patternnet and patternattribution," *arXiv preprint arXiv:1705.05598*, 2017.

[95] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th International Conference on Machine Learning*, JMLR. org, vol. 70, 2017, pp. 3145–3153.

[96] D. Smilkov, N. Thorat, B. Kim, *et al.*, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[97] H. Chen, S. Lundberg, and S.-I. Lee, "Explaining Models by Propagating Shapley Values of Local Components," *arXiv preprint arXiv:1911.11888*, 2019.

[98] J. Yosinski, J. Clune, Y. Bengio, *et al.*, "How transferable are features in deep neural networks?" In *In Proc. Advances in neural information processing systems*, 2014, pp. 3320–3328.

[99] J. Deng, W. Dong, R. Socher, *et al.*, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[100] F. Eitel, K. Ritter, A. D. N. I. (ADNI, *et al.*, "Testing the Robustness of Attribution Methods for Convolutional Neural Networks in MRI-Based Alzheimer's Disease Classification," in *In: Suzuki K. et al. (eds) Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support. ML-CDS 2019, IMIMIC 2019.. Springer, Cham, vol 11797.* DOI: 10.1007/978-3-030-33850-3_1.

[101] S. Pereira, R. Meier, V. Alves, *et al.*, "Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment," in *In Proc. Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, Springer, Cham, 2018, pp. 106–114.

[102] R. Sayres, A. Taly, E. Rahimy, *et al.*, "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, 2019.

[103] A. Singh, S. Sengupta, M. Abdul Rasheed, *et al.*, "Interpretation of deep learning using attributions : application to ophthalmic diagnosis," in *In Proc. Applications of Machine Learning, SPIE*, International Society for Optics and Photonics (SPIE), vol. 11511, 2020, 115110A.

[104] A. Singh, S. Sengupta, J. J. Balaji, *et al.*, "What is the optimal attribution method for explainable ophthalmic disease classification?" In *International Workshop on Ophthalmic Medical Image Analysis*, Springer, In press, 2020.

[105] Z. Papanastasopoulos, R. K. Samala, H.-P. Chan, *et al.*, "Explainable ai for medical imaging: Deep-learning cnn ensemble for classification of estrogen receptor status from breast mri," in *In proc. SPIE Medical Imaging 2020: Computer-Aided Diagnosis*, International Society for Optics and Photonics, vol. 11314, 2020, 113140Z.

[106] D. Lévy and A. Jain, "Breast mass classification from mammograms using deep convolutional neural networks," *arXiv preprint arXiv:1612.00542*, 2016.

[107] K. Young, G. Booth, B. Simpson, *et al.*, "Deep neural network or dermatologist?" In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, Springer, Cham, 2019, pp. 48–55.

[108] P. Van Molle, M. De Strooper, T. Verbelen, *et al.*, "Visualizing convolutional neural networks to improve decision support for skin lesion classification," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, Springer, Cham, 2018, pp. 115–123.

[109] V. Couteaux, O. Nempont, G. Pizaine, *et al.*, "Towards interpretability of segmentation networks by analyzing deepdreams," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, Springer, Cham, 2019, pp. 56–63.

[110] L. Wang and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images," *arXiv preprint arXiv:2003.09871*, 2020.

[111] Z. Q. Lin, M. J. Shafiee, S. Bochkarev, *et al.*, "Explaining with impact: A machine-centric strategy to quantify the performance of explainability algorithms," *arXiv preprint arXiv:1910.07387*, 2019.

[112] Y. Sha and M. D. Wang, "Interpretable predictions of clinical outcomes with an attention-based recurrent neural network," in *Proc. 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017, pp. 233–240.

[113] Z. Zhang, Y. Xie, F. Xing, *et al.*, "Mdnet: A semantically and visually interpretable medical image diagnosis network," in *Proc. IEEE conference on computer vision and pattern recognition*, 2017, pp. 6428–6436.

[114] J. Sun, F. Darbeha, M. Zaidi, *et al.*, "Saunet: Shape attentive u-net for interpretable medical image segmentation," *arXiv preprint arXiv:2001.07645*, 2020.

[115] B. Kim, M. Wattenberg, J. Gilmer, *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," *arXiv preprint arXiv:1711.11279*, 2017.

[116] M. Graziani, V. Andrearczyk, and H. Müller, "Regression concept vectors for bidirectional explanations in histopathology," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, Springer, Cham, 2018, pp. 124–132.

[117] H. Yeche, J. Harrison, and T. Berthier, "Ubs: A dimension-agnostic metric for concept vector interpretability applied to radiomics," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, Springer, Cham, 2019, pp. 12–20.

[118] M. Pisov, M. Goncharov, N. Kurochkina, *et al.*, "Incorporating task-specific structural knowledge into cnns for brain midline shift detection," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, Springer, Cham, 2019, pp. 30–38.

[119] P. Zhu and M. Ogino, "Guideline-based additive explanation for computer-aided diagnosis of lung nodules," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, Springer, Cham, 2019, pp. 39–47.

[120] N. C. Codella, C.-C. Lin, A. Halpern, *et al.*, "Collaborative human-ai (chai): Evidence-based interpretable melanoma classification in dermoscopic images," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, Springer, Cham, 2018, pp. 97–105.

[121] W. Silva, K. Fernandes, M. J. Cardoso, *et al.*, "Towards complementary explanations using deep neural networks," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, Springer, Cham, 2018, pp. 133–140.

[122] H. Lee, S. T. Kim, and Y. M. Ro, "Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, Springer, Cham, 2019, pp. 21–29.

[123] C. Biffi, J. J. Cerrolaza, G. Tarroni, *et al.*, "Explainable anatomical shape analysis through deep hierarchical generative models," *IEEE Transactions on Medical Imaging*, 2020. DOI: 10.1109/TMI.2020.2964499.

[124] T. Eslami, J. S. Raiker, and F. Saeed, "Explainable and scalable machine-learning algorithms for detection of autism spectrum disorder using fmri data," *arXiv preprint arXiv:2003.01541*, 2020.

[125] P. Romero-Aroca, "Managing diabetic macular edema: the leading cause of diabetes blindness," *World journal of diabetes*, vol. 2, no. 6, p. 98, 2011.

[126] J. R. Willis, Q. V. Doan, M. Gleeson, *et al.*, "Vision-Related Functional Burden of Diabetic Retinopathy Across Severity Levels in the United States," eng, *JAMA ophthalmology*, vol. 135, no. 9, pp. 926–932, 2017, ISSN: 2168-6173.

[127] V. Lakshminarayanan, "The global problem of blindness and visual dysfunction," *Photonic Innovations and Solutions for Complex Environments and Systems (PISCES)*, vol. 8482, 84820A, 2012.

[128] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[129] M. D. Abràmoff, Y Lou, A Erginay, *et al.*, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 13, pp. 5200–5206, 2016.

[130] P. Ruamviboonsuk, J. Krause, P. Chotcomwongse, *et al.*, "Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–9, 2019.

[131] H. Kaur, H. Nori, S. Jenkins, *et al.*, "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14. DOI: 10.1145/3313831.3376219.

[132] Z. Wang, P. Mardziel, A. Datta, *et al.*, "Interpreting interpretations: Organizing attribution methods by criteria," *arXiv preprint arXiv:2002.07985*, 2020.

[133] C Szegedy, W Liu, Y Jia, *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[134] D. Kermany and M. Goldbaum, "Labeled optical coherence tomography (OCT) and Chest X-Ray images for classification," *Mendeley Data*, vol. 2, 2018. DOI: 10.17632/RSCBJBR9SJ.2.

[135] A. Singh, J. J. Balaji, V. Jayakumar, *et al.*, "Quantitative and qualitative evaluation of explainable deep learning methods for ophthalmic diagnosis," *arXiv preprint arXiv:2009.12648*, 2020.

[136] S. Sengupta, A. Singh, J. Zelek, *et al.*, "Cross-domain diabetic retinopathy detection using deep learning," in *Applications of Machine Learning*, International Society for Optics and Photonics, vol. 11139, 2019, p. 111390V.

[137] M. Abadi, A. Agarwal, P. Barham, *et al.*, *Tensorflow: Large-scale machine learning on heterogeneous systems*.

[138] F. Chollet and Others, *Keras*, url https://keras.io, 2015.

[139] M. Ancona, C. Öztireli, and M. Gross, "Explaining deep neural networks with a polynomial time algorithm for shapley values approximation," *arXiv preprint arXiv:1903.10992*, 2019.

[140] W. Samek, A. Binder, G. Montavon, *et al.*, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.

[141] J. Adebayo, J. Gilmer, M. Muelly, *et al.*, *Sanity Checks for Saliency Maps*, 2018. arXiv: 1810.03292 [cs.CV].

[142] L. Sixt, M. Granz, and T. Landgraf, *When Explanations Lie: Why Many Modified BP Attributions Fail*, 2019. arXiv: 1912.09818.

[143] A. Singh, S. Sengupta, A. R. Mohammed, *et al.*, "Uncertainty aware and explainable diagnosis of retinal disease," in *In proc. SPIE Medical Imaging 2021*, International Society for Optics and Photonics, 2021, In press.

[144] R. Tomsett, A. Preece, D. Braines, *et al.*, "Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI," *Patterns*, vol. 1, no. 4, p. 100 049, 2020.

[145] G. Song, K. K. Chu, S. Kim, *et al.*, "First Clinical Application of Low-Cost OCT," *Translational Vision Science & Technology*, vol. 8, no. 3, p. 61, 2019, ISSN: 2164-2591. DOI: 10.1167/tvst.8.3.61.

[146] P. Mehta, C. Petersen, J. C. Wen, *et al.*, "Automated detection of glaucoma with interpretable machine learning using clinical data and multi-modal retinal images," *bioRxiv*, 2020. DOI: 10.1101/2020.02.26.967208.

[147] R. T. Yanagihara, C. S. Lee, D. S. W. Ting, *et al.*, "Methodological challenges of deep learning in optical coherence tomography for retinal diseases: A review," *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 11–11, 2020.