FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Disentanglement and Assessment of Shortcuts on Ophthalmological Retinal Imaging Exams

Leonor Justo dos Santos Fernandes

Master in Bioengineering - Biomedical Engineering

Supervisor: Jaime dos Santos Cardoso

Co-Supervisor: Luis Filipe Nakayama

Co-Supervisor: João Carlos Gonçalves Matos

Co-Supervisor: Tiago Filipe Sousa Gonçalves

July 31, 2025

Disentanglement and Assessment of Shortcuts on Ophthalmological Retinal Imaging Exams

Leonor Justo dos Santos Fernandes

Master in Bioengineering - Biomedical Engineering

July 31, 2025

Resumo

A retinopatia diabética (RD) é a principal causa de perda de visão em pessoas em idade ativa nos países desenvolvidos. Os programas de rastreio são fundamentais para reduzir o risco de cegueira, mas as técnicas tradicionais de imagiologia continuam a ser dispendiosas e inacessíveis em locais com recursos limitados, em parte devido à escassez de oftalmologistas. Os algoritmos de inteligência artificial (IA) apresentam uma solução promissora para simplificar o processo de diagnóstico, melhorando a acessibilidade. No entanto, persistem preocupações quanto à equidade e generalização desses modelos, dada a falta de dados oftalmológicos de países de baixo e médio rendimento (PBMR). Este trabalho tem como objetivo avaliar a equidade e o desempenho dos modelos treinados por imagem na previsão da RD e avaliar o *disentanglement* como uma abordagem de mitigação da desigualdade.

Utilizando duas bases de dados, mBRSET e BRSET, três modelos do estado da arte, incluindo o ConvNeXt V2 e os *vision transformers* DINOv2 e Swin V2, foram treinados em imagens da mácula para prever a RD e atributos sensíveis (AS) (e.g., idade e género/sexo). Além disso, a regressão logística, o *extreme gradient boosting* (XGBoost) e as árvores de decisão foram treinados em dados demográficos e clínicos para prever a RD. A equidade foi avaliada entre subgrupos de AS. O *disentanglement* foi efetuado e o seu impacto no desempenho e na equidade foi avaliado utilizando *area under the receiver operating characteristic curve* (AUROC), análise da curva de decisão e gráficos de distribuição de risco.

Os modelos de *deep learning* (DL) treinados em imagens da mácula da retina alcançaram um elevado desempenho no diagnóstico da RD (94% e 99% AUROC, para o mBRSET e BRSET, respetivamente). Os modelos DL também foram capazes de prever razoavelmente a idade e o género/sexo, produzindo AUROCs de 91% e 77%, respetivamente. Os modelos treinados para prever DR em dados de pacientes exibiram um desempenho substancialmente inferior para o mBRSET (72% AUROC) e um ligeiro declínio para o BRSET (92% AUROC). A avaliação da equidade sugere potenciais disparidades no mBRSET, com uma diferença de 26% e 10% AUROC observada entre pacientes com menos e mais de 50 anos de idade, para dados não aumentados e aumentados, respetivamente, enquanto o BRSET teve variações mínimas.

O *disentanglement* de AS da previsão da retinopatia diabética no mBRSET teve resultados variáveis, dependendo do modelo selecionado. O *disentanglement* utilizando o DINOv2 mostrou um aumento de desempenho AUROC de 23% e 2%, para dados não aumentados e aumentados, respetivamente. Para os modelos ConvNeXt V2 e Swin V2, usando dados aumentados, houve uma diminuição de 7% e 3% do AUROC, respetivamente. Embora o *disentanglement* não tenha levado a melhorias na justiça e no desempenho na maioria dos modelos, o DINOv2 mostrou resultados muito promissores. Dada a complexidade do *disentanglement* de caraterísticas finas em imagens de fundo de olho, são necessárias mais adaptações à arquitetura para conseguir um *disentanglement* eficaz.

Em suma, os nossos resultados sublinham a importância de priorizar à equidade em IA de imagem médica, bem como os desafios envolvidos na remoção de viés. É essencial continuar a desenvolver estratégias para garantir sistemas de IA equitativos e fiáveis nos cuidados de saúde.

Abstract

Diabetic retinopathy (DR) is the leading cause of vision loss in people of working age in developed countries. Screening programs are pivotal in reducing the risk of blindness, but traditional imaging techniques remain costly and inaccessible in resource-limited settings, partly due to a scarcity of ophthalmologists. Artificial intelligence (AI) algorithms present a promising solution to streamline the diagnostic process, improving accessibility. However, concerns persist regarding the fairness and generalization of these models, given the lack of ophthalmological data from low- and middle-income countries (LMICs). This work aims to evaluate the fairness and performance of image-trained models in predicting DR and assess disentanglement as a fairness mitigation approach.

Using two diverse fundus image datasets, mBRSET and BRSET, three state-of-the-art pretrained models, including the convolutional network ConvNeXt V2 and the vision transformers DINOv2 and Swin V2, were trained on macula images to predict DR and sensitive attributes (SAs) (e.g., age and gender/sex). Additionally, logistic regression, extreme gradient boosting (XGBoost), and decision trees were trained on demographic and clinical data to predict DR. Fairness was assessed between subgroups of SAs. Disentanglement was performed, and its impact on performance and fairness was assessed using area under the receiver operating characteristic curve (AUROC), decision curve analysis, and risk distribution plots.

Deep learning (DL) models trained on retinal macula images achieved high performance in diagnosing DR (94% and 99% AUROC, for mBRSET and BRSET, respectively). DL models were also able to reasonably predict age and gender/sex, yielding AUROCs of 91% and 77%, respectively. Models trained to predict DR on patient data exhibited a substantially lower performance for mBRSET (72% AUROC) and a slight decline for BRSET (92% AUROC). Fairness assessment suggests potential disparities in mBRSET, with a 26% and 10% AUROC difference observed between patients under and over 50 years old, in DINOv2, for non-augmented and augmented data, respectively, while BRSET had minimal variations.

Disentangling SA from the diabetic retinopathy prediction in mBRSET had varying results, depending on the model selected. Disentanglement using DINOv2 showed an AUROC performance increase of 23% and 2%, for non-augmented and augmented data, respectively. For ConvNeXt V2 and Swin V2 models, using augmented data, there was a 7% and 3% AUROC decrease, respectively. While disentanglement did not lead to improvements in fairness and performance for most models, DINOv2 showed very promising results. Given the complexity of disentangling fine-grained features in fundus imaging, further architectural adaptations are necessary to achieve effective disentanglement.

To conclude, our findings emphasize the importance of prioritizing fairness in medical imaging AI, as well as the challenges involved in addressing bias. It is essential to continue developing strategies to ensure equitable and reliable AI systems in healthcare.

Agradecimentos

Primeiro, queria agradecer ao meu orientador, Professor Jaime Cardoso, e co-orientadores Doutor Luis Nakayama, João Matos e Tiago Gonçalves, por toda a partilha de ideias, conselhos e acompanhamento dados, que foram essenciais para conseguir concretizar este trabalho. Um agradecimento especial ao João e Tiago por toda a disponibilidade, seja para debater ideias nas reuniões semanais, ou para responder a uma mensagem a qualquer hora e nos mais diversos cantos do mundo.

À APPEJ, que desde cedo me deu oportunidades únicas, e que me ensinou a ser líder. Criei memórias por Portugal inteiro, pude ensinar várias "gerações" de APPEJers que agora vejo a assegurar o futuro da organização, e acima de tudo, assisti a assembleias gerais verdadeiramente únicas, sobre coca-cola de marca branca, garrafões de azeite e batalhas de Aljubarrota. Nunca irei conseguir expressar o quanto está por trás de um "Alele" e de um "See you somewhere in Europe".

À TUNAFE, na qual entrei meramente pela música, e lentamente fui descobrindo outras razões para ficar. Cresci imensamente por fazer parte deste grupo, e fiz o meu melhor para aproveitar cada momento. Fosse com autopraxes, trotinetes, rissóis de camarão, caos ou diagramas de parentesco excessivamente complicados à mistura, foram momentos cheios de amizade na Faculdade de Engenharia.

Ao meu ano, por me ter acompanhado nesta aventura, a partir de um 1º ano de máscara, até à cartola erguida na cabeça. Um carinho especial à Ana, Constança e Joana pelos lanches do bar de minas, mesmo quando muitas vezes abria o PC para adiantar trabalho.

À Maria Mota e Ana Padrão, que me conseguem superar nas indecisões crónicas, e me dão motivos para sorrir todos os dias.

À minha família, por me apoiarem ao longo deste percurso, mesmo sem perceberem exatamente o que é a Bioengenharia. Um enorme obrigada por todos os conselhos e por sempre acreditarem em mim. É muito graças ao vosso apoio e compreensão que pude sonhar alto e ter tantas aventuras, mesmo quando significava chegar tarde ao jantar.

A todos os que me acompanharam durante o meu percurso académico, fosse de longe ou de perto, e que me estenderam a mão nas muitas vezes que caí. Um obrigada muito especial ao professor Paulo Ramos, à Ana Leite, Inês Graça, Helena Abelha, Inês Martins, Inês Calmeiro, Ana Santos e Maria Beatriz Calçada.

Finalmente a ti, Leonor, por teres recheado estes cinco anos de memórias, "cultos", ambição e amizades, ficando aquém apenas de um bom horário de sono.

Leonor Fernandes

"O homem primeiro tropeça, depois anda, depois corre,
um dia voará."

José Saramago

Contents

List of Figures

List of Tables

Abbreviations and Symbols

AI	Artificial Intelligence
AMD	Age-related Macular Degeneration
AUROC	Area Under the Receiver Operating Characteristic Curve
CAM	Class Activation Mapping
CNNs	Convolutional Neural Networks
DCA	Decision Curve Analysis
Deep SHAP	Deep SHapley Additive exPlanations
DL	Deep Learning
DME	Diabetic Macular Edema
DNNs	Deep Neural Networks
DR	Diabetic Retinopathy
DT	Decision Tree
ETDRS	Early Treatment Diabetic Retinopathy Study
FAF	Fundus Autofluorescence
FAS	Fair Adaptive Scaling
FIS	Fair Identity Scaling
FP	False Positive
Grad-CAM	Gradient-weighted Class Activation Mapping
ICDR	International Clinical Diabetic Retinopathy
LIME	Local Interpretable Model-agnostic Explanations
LMICs	Low- and Middle-Income Countries
LR	Logistic Regression
LRP	Layer-wise Relevance Propagation
ML	Machine Learning
NPDR	Non-Proliferative Diabetic Retinopathy
OCT	Optical Coherence Tomography
PSD	Performance-Scaled Disparity
PDR	Proliferative Diabetic Retinopathy
SA	Sensitive Attribute
SGD	Sustainable Development Goal
TCAV	Testing with Concept Activation Vectors
TP	True Positive
UWF	Ultra-Widefield Imaging
VAE	Adversarial Variational Autoencoder
VEGF	Vascular Endothelial Growth Factor
ViT	Vision Transformer
XAI	eXplainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

Chapter 1

Introduction

1.1 Context

1.1.1 Diabetic Retinopathy

Diabetic retinopathy (DR) is one of the most common complications of diabetes and a leading cause of visual impairment and blindness worldwide. It affects the retina, which is the light-sensitive tissue at the back of the eye responsible for vision. The disease develops when chronically high blood glucose levels damage the small blood vessels that supply the retina, resulting in swelling, leakage or bleeding. In response, the eye may attempt to grow new, but often abnormal and fragile blood vessels that can further impair vision or lead to retinal detachment [1].

In its early stages, the condition is typically asymptomatic. As DR advances, individuals may experience symptoms such as blurry vision, floating spots in vision, or even sudden vision loss. If left untreated, DR can lead to irreversible blindness [2]. The risk of developing DR increases with the duration of diabetes and can be further elevated by poor glycemic control and hypertension. Regular eye examinations are essential for people with diabetes, as early detection and timely treatment, such as laser therapy, anti-angiogenic therapy, or surgical procedures, can prevent or slow the progression of diabetic retinopathy and help preserve vision [3, 4].

DR is estimated to affect 22.27% of diabetic individuals. Its global prevalence is expected to rise due to increasing obesity rates, aging populations, and longer life expectancy of people with diabetes [5, 6]. However, this rise will not affect all regions equally. Disparities in disease prevalence are driven by factors such as poor diabetes management and uneven access to eye care services, particularly affecting resource-limited areas [7, 8].

1.1.2 Bias and fairness in ophthalmic AI

Bias in machine learning refers to the tendency of an algorithm to systematically favor certain outcomes over others, resulting in unfair advantages or disadvantages for specific groups. In healthcare, these

biases can have significant consequences, as AI and machine learning models are increasingly used to guide clinical decision-making, diagnose diseases, and allocate resources [9].

In the field of ophthalmology, particularly with retinal fundus imaging, bias can emerge due to the complex nature of the data. Fundus images encode not only explicit clinical features but also demographic attributes such as sex and age, along with visual artifacts related to imaging conditions [10]. The presence of this additional information introduces additional sources of variation that can influence model behavior. Bias in ophthalmic AI systems is often exacerbated by imbalanced datasets, where certain populations are underrepresented, particularly those from low- and middle-income countries. This lack of diversity can lead to disparities in model performance across demographic groups, raising serious concerns about algorithmic fairness, which may translate to increased diagnostic errors for underrepresented populations [11].

1.2 Motivation

Deep learning (DL) holds great potential to improve clinical outcomes and expand access to healthcare, particularly in the early detection and management of DR. By enabling AI-based screening services, DL systems can help reduce unnecessary referrals to eye care professionals and assist ophthalmologists in making more accurate and timely diagnoses. However, to minimize patient harm, address existing healthcare disparities, and foster trust in clinical deployment, it is crucial that DL systems not only achieve high predictive performance but also operate fairly across diverse patient populations.

1.3 Objectives

This dissertation has three primary objectives. First, it explores a group fairness methodology for identifying bias in the classification of diabetic retinopathy from fundus photos. Second, it implements a disentanglement network to mitigate the effects of confounding information. Third, it evaluates its impact on bias and performance, compared with the baseline models.

1.4 Alignment with Sustainable Development Goals

The Sustainable Development Goals (SDGs) detail 17 principles, with 169 associated targets, to be reached by 2030. Adopted in 2015 by all countries in the United Nations, SDGs address global challenges, such as ending poverty, reducing inequality, and protecting the planet, among many others. This dissertation directly aligns with “SDG 3 - Ensure healthy lives and promote well-being for all at all ages” and “SDG 10 - Reduce inequality within and among countries” [12]. Table 1.1 outlines the contributions of this dissertation to the SDGs.

Table 1.1: Contributions of this dissertation to the SDGs.

SDG	**Target**	**Contribution**	**Performance Indicators and Metrics**
3	Ensure healthy lives and promote well-being for all at all ages.	Contributes to the early diagnosis of DR, a major cause of preventable blindness by developing a DL system for automated detection.	Increase in early treatment; Decrease in the number of vision-threatening DR cases prevented;
10	Reduce inequality within and among countries.	Addresses global inequality in access to eye care by promoting DL-based screening in low-resource settings via portable devices and telemedicine; Tackles algorithmic bias through fairness analysis and bias mitigation.	Reduction in time-to-diagnosis in resource-limited settings; Estimated number of vision-threatening DR cases prevented in LMICS; Subgroup performance metrics.

1.5 Main Contributions

This work evaluates the impact of disentanglement on performance and fairness mitigation in diabetic retinopathy prediction, using mBRSET and BRSET, two recently proposed datasets consisting of handheld fundus camera images and standard tabletop fundus camera images, using area under the receiver operating characteristic curve (AUROC), decision curve analysis, and risk distribution plots. To our knowledge, this is one of the first analyses of a handheld fundus camera dataset in the context of fairness. A disentangled autoencoder architecture was developed and tested as a fairness mitigation technique, with performance and fairness improvements in one of the three models studied. An oral presentation of this work was delivered at the "Encontro de Investigação Jovem da U.Porto" (IJUP), organized by the University of Porto. A paper was submitted for the Fairness & Medical Imaging Workshop at the 28th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2025), titled "Disentanglement and Assessment of Shortcuts in Ophthalmological Retinal Imaging Exams".

1.6 Structure of the Dissertation

This dissertation is structured into six chapters. Chapter 1 introduces the topic and objectives of this thesis, as well as the main contributions of this work. Chapter 2 explains the clinical background, beginning with an overview of retinal imaging and diseases and then focusing specifically on diabetic retinopathy. Chapter 3 explores deep learning in ophthalmology, emphasizing the current state of the art in DR prediction. Afterwards, it covers key concepts such as fairness, bias, explainable artificial intelligence, shortcut learning, and disentanglement. Chapter 4 details the methodology adopted in this

work, and the results and discussion are described in Chapter 5. Chapter 6 concludes the dissertation, summarizing results and outlining potential directions for future work.

Chapter 2

Understanding Diabetic Retinopathy

Vision loss or impairment can significantly impact quality of life. Ophthalmology, the medical speciality dedicated to eye health, focuses on diagnosing and treating conditions that affect vision. Many eye pathologies originate or manifest in the retina. Despite being located in the eye, the retina is an extension of the central nervous system and the only part of the body where blood vessels and neural tissue can be observed directly and non-invasively. Due to this, it offers a unique window into the brain and circulation, making it highly informative for diagnosing and monitoring many systemic diseases, such as hypertension, diabetes, and neurological disorders [13].

2.1 Retinal Imaging

Retinal imaging is a diagnostic exam that obtains high-resolution digital images of the back of the eye. The technique enables the identification of many eye pathologies and is also used to monitor disease progression and treatment response.

2.1.1 Eye anatomy

The human eye is a complex organ, composed of multiple structures that play a role in the processing of visual information (see Figure 2.1).

Externally, the white part of the eye is called the **sclera**. It provides structure to the eyeball and maintains its general shape. The **iris**, the colored part of the eye, surrounds the **pupil**, the black circle inside, that widens and narrows to control how much light enters the eye. Covering both the iris and the pupil is the **cornea**, a transparent surface that allows light to enter. **Extraocular muscles**, attached to the outside of the eye, control eye position, alignment, and movement.

Internally, just behind the pupil is a transparent, curved disc, the **lens**. It focuses the incoming light, creating an image on the retina. The cavity situated between the lens and the retina is known as the **vitreous chamber**. It is filled with a clear gel-like substance named **vitreous humor**, which helps maintain eye shape. In the back inner wall of the eye lies the **retina**, a multilayered, light-sensitive

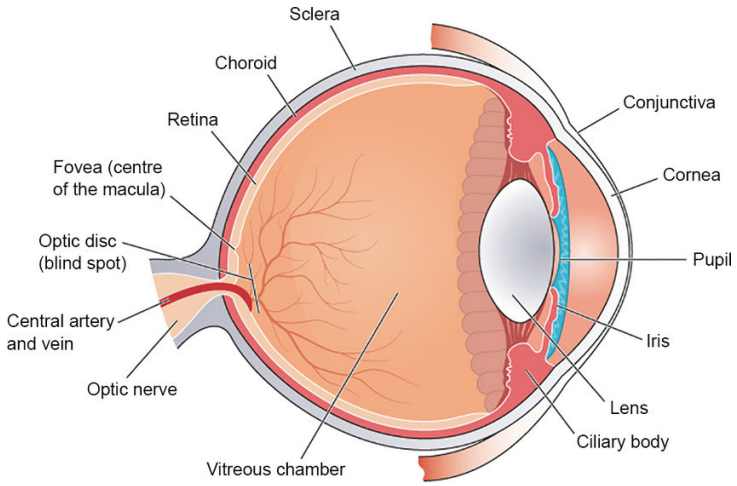


Figure 2.1: Anatomy of the eye (from [15]).

tissue. It converts the incoming light into electrical signals through specialized photoreceptor cells. These signals are then transmitted via the optic disc to the optic nerve, which takes the electrical signals and delivers them to the brain [14]. The **macula** is the central part of the retina. It is responsible for central vision and for seeing color and fine details. Within the macula is the **fovea**, the region of sharpest vision. The **eye fundus** refers to the inner, back surface of the eye, visible during an eye examination. It includes the retina, macula, optic disc, fovea, and the retinal blood vessels.

2.1.2 Imaging modalities for retinal assessment

Accurate and prompt assessment of the retina is important to improve retinal disease outcomes. Several imaging modalities are employed in clinical practice and research, each offering specific strengths and limitations:

- **Optical coherence tomography (OCT):** Non-invasive imaging technique that uses light waves to generate high-resolution cross-sectional images of the retina. It is used to quantify retinal thickening and is the standard of reference for detecting and monitoring diabetic macular edema (DME) [16].
- **Fundus photography:** High-resolution color images of the retina. Widely used and non-invasive. There are several variants of fundus photography.
  - **Standard fundus photography:** Captures 30° of the posterior pole of the eye, which includes the macula and the optic nerve. Common modality in clinical practice, due to being easy to use and highly available [17].
  - **Single-field fundus photography:** Frequently used in screening programs. It is fast and simple to use, although it has limited coverage [18].

  - **Seven-field standardized photography:** Standardized imaging protocol for detailed assessment of DR and DME. It is considered the gold standard for clinical trials and large-scale epidemiological studies [18].
  - **Stereoscopic fundus photography:** Images with depth perception, useful for distinguishing between extraretinal neovascularization and intraretinal microvascular abnormalities [17].

- **Fluorescein angiography:** Fluorescent dye is injected intravenously and circulates in the retinal vasculature, captured using rapid-sequence imaging. It is an invasive modality, which is especially valuable for identifying vascular pathologies, such as retinal ischemia, leakage, and neovascularization [19].

- **Fundus autofluorescence (FAF):** Non-invasive imaging modality. Utilises the fluorescent properties of pigments in the retina to create an image. Provides complementary information to structural imaging techniques [20].

- **Ultra-wide field (UWF) retinal imaging:** Modern, non-invasive technology that captures a significantly higher area of the retina than traditional fundus cameras. Capturing the peripheral retina in a single image enhances the detection of peripheral pathologies, commonly missed in standard imaging techniques. Although it allows for more precise grading of DR, widespread clinical use is limited by its high cost and restricted availability, particularly in resource-limited settings [21].

2.1.3 The role of retinal imaging in disease detection and monitoring

Retinal imaging is a crucial diagnostic tool for detecting and monitoring a wide range of eye conditions, as well as systemic health issues that manifest in the eye.

2.1.3.1 Eye conditions

Retinal imaging plays an essential role in the diagnosis and management of various eye diseases, many of which can lead to vision loss if not detected early. These conditions include:

- **Diabetic retinopathy (DR):** The focus of this dissertation. It is a microvascular disease, caused by chronic hyperglycemia that damages the small blood vessels that supply the retina, resulting in leakage, blockage or abnormal growth of retinal vessels, leading to impaired vision.

- **Diabetic macular edema:** Retinal thickening within two disc diameters of the fovea. It is generally caused by leakage of fluid from damaged blood vessels into the macula, the central part of the retina responsible for sharp vision, resulting in blurred vision. It is a complication of DR and is estimated to affect 1 in 15 people with diabetes [22].

- **Glaucoma:** Group of eye disorders characterized by progressive deterioration of the optic nerve that can lead to vision loss. It is caused by elevated pressure of the fluid inside the eye, pressing

down on the optic disc and damaging the nerve fibers. It is the most common cause of irreversible blindness worldwide. The most common type of glaucoma is primary open-angle glaucoma [23].

- **Cataracts:** Loss of lens transparency due to opacification of the lens. It is the leading cause of blindness worldwide and one of the major causes of visual impairment in people with diabetes. Diabetic patients are reported to be up to five times more likely to develop cataracts, in particular at earlier ages [24].
- **Retinal detachment:** Diabetic retinopathy can cause scars to form in the back of your eye. When the scars pull your retina away from the back of your eye, it is called tractional retinal detachment [25].
- **Age-related macular degeneration (AMD)**
  - **Dry AMD (or atrophic AMD):** The most common form, characterized by thinning of the macula with age. It has 3 stages: early, intermediate and late, and typically progresses slowly over several years.
  - **Wet AMD (or advanced neovascular AMD):** Less common type of AMD that causes faster vision loss. It is caused by the growth of abnormal blood vessels under the retina that leak blood or other fluids, scarring the macula [26].

2.1.3.2 Systemic health conditions

Beyond ocular diseases, retinal imaging also provides insights into systemic health conditions. Changes in the retina often reflect broader vascular or neurological disorders, making retinal assessment a valuable non-invasive tool for early detection and monitoring of systemic diseases, such as:

- **Hypertension:** Chronic high blood pressure can cause changes in the retinal blood vessels, such as narrowing, leaking, or hemorrhages, which can be detected through retinal imaging. Blood pressure control has been shown to decrease the onset and development of microvascular complications, including diabetic retinopathy [27].
- **Multiple sclerosis:** Chronic autoimmune neuroinflammatory condition, usually accompanied by optic neuritis and retinal neuro-axonal damage, causing visual disturbances. OCT has been shown to be a reliable marker in detecting alterations in the optic nerve that may be indicative of multiple sclerosis [28].

2.2 Diabetic retinopathy

Diabetic retinopathy is clinically characterized by the presence of typical retinal microvascular lesions in individuals with diabetes. It is the most common complication of diabetes mellitus and a leading

cause of vision loss among working-age adults. DR is not an isolated condition; it reflects the broader vascular complications of diabetes, which are generally categorized into macrovascular and microvascular complications. Macrovascular complications include cardiovascular disease, stroke, and peripheral artery disease. Microvascular complications affect the kidneys (nephropathy), eyes (retinopathy), and peripheral nerves (neuropathy) [29].

2.2.1 Clinical features

DR is described by a variety of retinal lesions, including:

- **Microaneurysms:** Small outpouchings of the retinal capillaries, often appearing as tiny red dots on the retina. Result from weakening of the capillary walls due to loss of pericytes and endothelial cell damage [30].
- **Hemorrhages:** Leakage of blood from damaged retinal vessels, which appear as dark red spots or splotches on the retina [30].
- **Hard exudates:** Yellow-white deposits composed of lipids and proteins leaking plasma. They accumulate in the outer retina and often form clusters or rings, indicating previous or ongoing vascular leakage [30].
- **Cotton wool spots (soft exudates):** White, fluffy patches on the retina caused by small infarctions (blockages) of the retinal nerve fiber layer. They represent areas of nerve fiber swelling due to interrupted blood supply [30].
- **Intraretinal microvascular abnormalities:** Abnormal or dilated small blood vessels within the retina, representing remodeling of the retinal vasculature due to ischemia. It indicates more advanced non-proliferative diabetic retinopathy [31].
- **Venous beading:** Localized increase in the diameter of retinal veins, forming “beads”. It is a sign of severe non-proliferative diabetic retinopathy and indicates worsening retinal ischemia [31].
- **Neovascularization:** Proliferation of new abnormal blood vessels, which are fragile and prone to leaking or bleed easily, further exacerbating retinal damage. Stimulated to grow in response to retinal ischemia [32].
- **Fibrous tissue proliferation:** Occurs as a reparative responsive to retinal injury and hypoxia. Formation of fibrous tissue is typically associated with neovascularization. Contributes to tractional retinal detachment in advanced stages [33].

2.2.2 Disease cause

Diabetic retinopathy is primarily caused by chronic hyperglycemia, which induces a cascade of pathological changes in the retinal microvasculature, including biochemical, hemodynamic, and structural alterations. Chronic hyperglycemia imposes sustained metabolic stress on the retina, initiating the following sequence of events:

1. **Dilation of retinal blood vessels and altered blood flow:** These are autoregulation mechanisms to increase metabolism in diabetic subjects. Initially, these changes are compensatory, but prolonged hyperglycemia disrupts normal vascular function [1].

2. **Pericyte loss:** Pericytes provide structural support to capillaries. Their loss increases capillary permeability and leads to the formation of microaneurysms, the earliest clinical sign of diabetic retinopathy. It can further progress to the apoptosis of endothelial cells and thickening of the basement membrane, contributing to the breakdown of the blood-retina barrier [1].

3. **Capillary occlusion and retinal ischemia:** The pronounced loss of pericytes and endothelial cells results in capillary closure, depriving retinal tissue of oxygen and nutrients. In response, the retina activates the vascular endothelial growth factor (VEGF) pathways to stimulate the formation of new blood vessels, in an attempt to restore oxygen and nutrient supply [1].

2.2.3 Classification systems

There are three main classification systems for diabetic retinopathy:

2.2.3.1 Early Treatment Diabetic Retinopathy Study Classification

The Early Treatment Diabetic Retinopathy Study (ETDRS) classification system [34] is widely regarded as the gold standard for grading diabetic retinopathy in the context of clinical trials and research. It consists of a 14-level severity scale that enables detailed quantification of retinal lesions, making it the preferred method for standardized grading. Fundus lesions are graded individually from standard 7-field 30° fundus photographs and the assessments are integrated to determine an overall DR severity level. However, it has limited daily clinical use, as it requires specialized imaging and trained graders [35].

2.2.3.2 International Clinical Diabetic Retinopathy Disease Severity Scale

The international clinical diabetic retinopathy disease (ICDR) severity scale [36] was developed to address the need for a unified global classification system that facilitates communication and clinical decision-making. It is a simplified version of the ETDRS classification system, condensing the 14 ETDRS levels into 5 clinical stages. Due to its convenience and ease of use by medical practitioners in

daily clinical use, it is the most commonly used classification system worldwide. It is endorsed by international guidelines for DR management, including those from International Council of Ophthalmology and the American Academy of Ophthalmology [35, 37].

2.2.3.3 Scottish Diabetic Retinopathy Grading Scheme

Hierarchical, feature-based system for assessment of DR severity, typically via retinal photography. The system grades DR severity from R0 to R4, and separately also grades maculopathy from M0 to M2. It is widely used in Scotland's national diabetic retinopathy screening programme [38].

2.2.4 Disease types

The ICDR scale, outlined in Table 2.1, classifies DR into two main stages, based on the level of microvascular degeneration and related ischemic damage: non-proliferative DR (NPDR) and proliferative DR (PDR).

Non-proliferative diabetic retinopathy (NPDR) represents the early stages of diabetic retinopathy. It is further classified into mild, moderate and severe. It is characterized by increased vascular permeability and capillary occlusion. Retinal pathologies include microaneurysms, hemorrhages, and hard exudates, which can be detected by fundus photography, although the patients may be asymptomatic.

Proliferative diabetic retinopathy (PDR) represents the most advanced stage, characterized by neovascularization and subsequent fibrous proliferation in response to retinal ischaemia [36].

2.2.5 Symptoms

- **Early Symptoms:** None;
- **Later Symptoms:** Blurry vision, floating spots in vision, blind spots, low vision, blindness [2, 39].

DR is often asymptomatic in its early stages. Despite being preventable, without timely detection and management, it can progress to advanced stages and lead to permanent vision loss or blindness in diabetic people. Screening for this disease is essential for early treatment, significantly reducing the risk of vision loss. The risk of developing DR increases with the duration of diabetes [2].

Over time, it is estimated that more than half of people with diabetes will develop diabetic retinopathy.

2.2.6 Treatment options

In the early stages of DR, the recommended treatment consists of regular screening and systemic control of blood glucose, blood pressure, and lipids, via intensive insulin therapy. It significantly reduces the risk of retinopathy prevalence and progression of the disease [3, 4].

For advanced stages, the main treatment options are:

Table 2.1: International classification of diabetic retinopathy (ICDR) severity scale [36].

ICDR Level	**Observable findings upon dilated ophthalmoscopy**
No apparent DR	No abnormalities
Mild NPDR	Microaneurysms only
Moderate NPDR	More than just microaneurysms but less than severe NPDR, such as microaneurysms with other signs like intraretinal hemorrhages, hard exudates, or cotton wool spots
Severe NPDR	Any of the following (4:2:1 rule): • More than 20 intraretinal hemorrhages in each of 4 quadrants • Definite venous beading in 2 or more quadrants • Prominent intraretinal microvascular abnormalities (IRMA) in 1 or more quadrants (with no signs of PDR)
PDR	One or more of the following: • Neovascularization • Vitreous or preretinal hemorrhage

- **Anti-vascular endothelial growth factor (anti-VEGF) therapy:** Currently, one of the main therapies used in DR, particularly in cases involving DME and PDR. Anti-VEGF agents inhibit VEGF, reducing abnormal blood vessel growth, leakage, and swelling in the retina [40].
- **Laser treatments:**
  - **Pan-retinal photocoagulation:** The previous gold standard for treating DME and PDR, before the surge of anti-VEGF therapy. Currently, it plays an adjuvant treatment role. The procedure involves applying over 1000 laser burns in the peripheral retina, while sparing the central macula. This reduces retinal ischemia, lowers VEGF production, and promotes regression of neovascularization [40].
  - **Subthreshold laser (STL):** Newer laser modality that reduces the adverse side effects of conventional laser photocoagulation. STL uses laser energy with lower intensities, reducing retinal tissue damage [41].
- **Vitrectomy:** Surgical procedure that involves removing and replacing the vitreous humor of the eye. It is commonly used in the management of advanced diabetic retinopathy, particularly in

cases of vitreous hemorrhage or tractional retinal detachment. It helps remove blood and scar tissue caused by DR-related damage, thereby preserving vision [42].

- **Corticosteroids:** Potent anti-inflammatory agents, targeting mechanisms associated with DME pathogenesis. Corticosteroids are gaining increased importance in DME treatment, particularly in cases of lack of response to anti-VEGF treatment. Corticosteroids have shown promising efficacy, with the additional advantage of a lower injection frequency compared to anti-VEGF agents [43].

2.2.7 Incidence, burden and future projection

2.2.7.1 Incidence and Prevalence

According to the International Diabetes Federation, 463 million people were living with diabetes in 2019, and this number is projected to reach 700 million by 2045 [44]. As a consequence, the number of people with DR is expected to increase from 103 million in 2020 to 130 million in 2030, and 161 million by 2045 [5].

Globally, the prevalence of DR among people with diabetes is estimated at 22.27%, with risk increasing significantly with the duration of diabetes [5]. Nearly 99% of individuals with type 1 diabetes and about 60% with type 2 diabetes will develop some degree of retinopathy after 20 years of living with the disease [45].

2.2.7.2 Disease burden and impact

DR is a significant cause of visual impairment and blindness. From 1990 to 2020, the age-standardized global prevalence of blindness due to diabetic eye disease increased from 14.9% to 18.5%. DR is the fifth leading cause of blindness and moderate-to-severe vision impairment in adults over 50 years of age and is the leading cause of blindness in the working-age population [46, 47].

The economic burden includes both direct medical costs and indirect costs such as reduced productivity, increased hospital admissions, and diminished quality of life. Vision loss contributes to poverty and hinders national economic development [48].

2.2.7.3 Future projections

Much of the visual loss from DR is preventable. DR vision loss rates have steadily declined in recent decades, particularly in developed countries, largely due to improved awareness, early screening, and better systemic control of diabetic patients [5].

Despite positive trends in some regions, projections indicate a continued global increase in DR cases, as outlined in Table 2.2. This increase parallels global diabetes trends, which are driven by rapidly increasing obesity rates, sedentary lifestyles, and aging populations. As the life expectancy of diabetic patients improves due to improved systemic care, their cumulative risk of developing DR also increases [5, 6].

Table 2.2: Estimated number of individuals with diabetic retinopathy and vision-threatening diabetic retinopathy by world region (2020–2045) (from [5]).

World Region	**Diabetic Retinopathy (in Millions)**			**Vision-Threatening Diabetic Retinopathy (in Millions)**		
	2020	**2030**	**2045**	**2020**	**2030**	**2045**
South-East Asia (SEA)	14.95	19.62	26.06	3.15	4.13	5.48
Africa	6.99	10.29	16.93	2.83	4.16	6.84
Europe	11.25	12.46	12.89	3.28	3.64	3.76
Middle East and North Africa (MENA)	18.07	25.05	35.47	4.59	6.36	9.01
North America and Caribbean (NAC)	15.89	18.75	21.11	3.78	4.46	5.02
South and Central America (SACA)	4.47	5.69	6.96	1.87	2.37	2.90
Western Pacific (WP)	31.50	37.98	41.08	9.06	10.93	11.81
Global	103.12	129.84	160.50	28.54	36.05	44.82

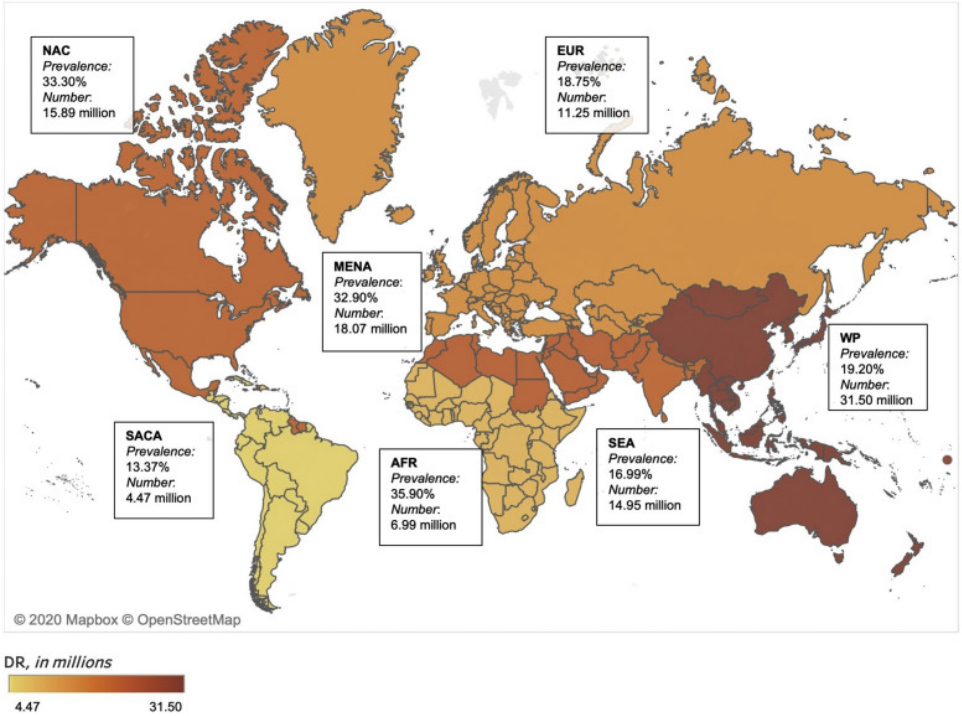


Figure 2.2: Global map showing diabetic retinopathy (DR) prevalence and numbers by International Diabetes Foundation world regions in 2020. Regions include: AFR = Africa; EUR = Europe; MENA = Middle East and North Africa; NAC = North America and Caribbean; SACA = South and Central America; SEA = South East Asia; WP = Western Pacific (from [5]).

2.2.8 Disparities in disease prevalence

The global burden of DR is rising, but it will not affect all regions equally, as shown in Figure 2.2. It is estimated that 75% of people with diabetes live in low and middle-income countries, where healthcare facilities for managing diabetes are poor and public awareness is low [7].

Additionally, the distribution of eye care professionals is highly uneven, with two-thirds of global ophthalmologists concentrated in just thirteen countries (China, USA, India, Brazil, Russia, Germany, Italy, Egypt, France, Mexico, Spain, and Poland). National income levels are strongly correlated with ophthalmologist availability, and within-country disparities also exist, reflecting broader socioeconomic inequalities [8].

Projections indicate that the global burden will remain high through 2045, with disproportionate effects in regions like the Middle East and North Africa, and the Western Pacific. These areas are expected to see the steepest increases in disease burden due to both demographic trends and healthcare access limitations [5].

2.2.9 Summary

In this chapter, we explored the clinical background and epidemiology of DR, a leading cause of vision impairment and blindness globally. The asymptomatic nature of early-stage disease, combined with the limited ability of current treatments to reverse retinal damage, makes early detection critical. Additionally, the global incidence of DR is projected to rise, with a disproportionate effect on low and middle-income countries. This is a worrying trend, that reinforces the importance of accessible diagnostic tools and sustainable screening programs. It is vital that we develop targeted health strategies that effectively address the growing burden of DR-related vision loss.

Chapter 3

Background and Related Work

This chapter explores the application of deep learning in healthcare, focusing particularly on ophthalmology, and presents the current state-of-the-art methods for DR prediction. Afterwards, it covers key concepts such as fairness and bias, explainable artificial intelligence (XAI), shortcut learning, and disentanglement.

3.1 Deep Learning in healthcare

Artificial intelligence (AI) is considered the fourth industrial revolution in human history. AI encompasses technology that enables computers and machines to simulate human-like behavior. Machine Learning (ML) is a subset of AI where algorithms have the ability to learn without being explicitly programmed. Deep Learning (DL) is an ML field that takes inspiration from the structure of the human brain, using artificial neural networks with multiple layers to learn complex data patterns. In recent years, Deep Learning has attracted great interest and been widely adopted in several fields, such as speech recognition, image recognition and natural language processing [49]. Its capacity to learn from large datasets and complex data makes it very promising for healthcare applications.

3.1.1 Deep Learning in ophthalmology

Ophthalmology stands to benefit greatly from DL technologies. Compared to other medical specialities, such as obstetrics, ophthalmology has access to extensive, high-quality imaging datasets, positioning the field at the forefront of AI-driven health care [50].

The main advantages of DL applications for ophthalmology are:

- **Improving diagnostic accuracy:** DL models have achieved disease prediction with accuracy challenging that of specialists, allowing for earlier detection of diseases, thereby improving patient outcomes and reducing the growing burden of preventable vision loss [51]. Earlier diagnosis and decreased diagnostic errors also lower the clinical burden and financial costs of healthcare systems [52]. Moreover, with clinical, genetic and lifestyle data of the patient, DL models can tailor

treatment to the needs of each patient, minimizing adverse effects while optimizing therapeutic efficacy [53].

- **Improving ophthalmology outreach:** Resource-limited settings have a shortage of ophthalmologists and technical equipment that severely impacts patient eye care [54]. Integrating DL algorithms in screening and diagnostic processes can help mend the gap in healthcare access, as they enable large-scale screening and monitoring. Such tasks traditionally require substantial financial and human resources, limiting their effect. By facilitating early detection and intervention, these technologies can have a very positive impact on public health.

- **Reduce resource strain:** Traditional diagnosing and monitoring require substantial time. DL models can act as support tools, accelerating decision-making, freeing up clinicians to devote more time to patient care, and decreasing patient wait times. Furthermore, setting up AI-based screening services could reduce the number of direct referrals and the workload of specialized tertiary centers [55, 56].

3.1.2 Barriers to application

- **Economic challenges:** Long-term DL applications in ophthalmology can be cost-effective, reducing costs [57]. However, the initial economic barrier of acquiring hardware equipment, AI software, and operating costs, as opposed to direct ophthalmologic evaluation, can pose a limitation to AI implementation in developing countries. Investing in affordable technology access is necessary to facilitate worldwide integration of DL-based systems [58]. Smartphone-based handheld devices offer a low-cost and accessible alternative, with potential to increase the outreach of eye care services [59].

- **Patient and clinician acceptance:** Resistance to AI implementation prevails among both clinicians and patients. Clinicians hold concerns over the "black box" nature of deep learning, which complicates clinical oversight and could strain patient-doctor relationships, in cases where doctors are unable to explain DL decisions to their patients. Among the public, limited technological literacy, insufficient AI education, along with fear of potential harm, given the high stakes of medical decisions, also fuel concern [60]. Enhancing model interpretability is an important step toward building trust in these systems. Additionally, improving AI literacy among both healthcare professionals and patients is also essential for promoting acceptance of AI-assisted healthcare.

- **Ethical and privacy concerns:** The use of retinal images in deep learning systems raises significant privacy concerns due to their highly distinctive biometric data. The vascular pattern in the retina is unique and remains unchanged throughout an individual's lifetime, making it susceptible to re-identification even after de-identification procedures [61]. To mitigate these risks, it is crucial to implement robust data protection strategies, such as secure storage and advanced de-identification techniques.

- **Lack of generalizability:** DL models are susceptible to discriminatory bias, due to the underrepresentation of sensitive features, such as ethnicity, gender, age, and socioeconomic status, in training datasets. For example, open-access datasets are primarily sourced from developed countries, which means many ethnic populations are underrepresented. This lack of diversity can lead to models that perform well only for the majority group, leading to increased diagnostic error in real-world implementations [62, 63]. Addressing this issue requires the development of diverse, representative datasets, alongside fairness studies to monitor and reduce algorithmic bias.

3.1.3 Diabetic retinopathy prediction

Several ophthalmological medical techniques have been successfully integrated with DL systems, including fundus photography, OCT, slit-lamp imaging, and visual field analysis. These techniques have shown promising results in the detection and management of conditions such as diabetic retinopathy, retinopathy of prematurity, glaucoma, macular edema, and age-related macular degeneration.

DL has revolutionized the diagnostic performance in detecting DR. Table 3.1 summarizes the state of the art of DL algorithms for DR prediction tasks, from fundus photographs.

Vasireddi et al. [64] introduce **DR-XAI**, an explainable deep learning framework developed for five-class DR prediction, from fundus photographs. The preprocessing pipeline employs a combination of adaptive histogram equalization (AHE), gamma correction, and contrast-limited adaptive histogram equalization (CLAHE) to denoise and enhance the fundus images. Additionally, the system performs automatic optic disk localization and removal to avoid misclassification between the optic disk and hard exudates. It uses a deep feedforward neural network (DFNN) architecture, with a Lion Optimization Algorithm (LOA) for optimization. Explainability is provided using Local Interpretable Model-agnostic Explanations (LIME). The model is trained using the MESSIDOR dataset, obtaining an accuracy of 98.04%, sensitivity of 99.69%, specificity of 96.37% and F1-score of 96.99%.

Almas et al. [65] propose an enhanced stacked auto-encoder–based model for five-class DR detection. It performs data augmentation techniques to underrepresented classes, artificially balancing the dataset. During the training process, dropout is performed to avoid overfitting. Using the KAGGLE fundus image dataset (n=35,126), a 93% training and 88% testing accuracy is achieved.

Table 3.1: Summary of SOTA DR prediction algorithms, from fundus photographs.

Ref.	Year	Datasets	Contributions
[66]	2016	EyePACS-1 [66], Messidor-2 [67]	Used Inception-v3 Convolutional Neural Network (CNN) for referable DR detection from color fundus photographs. AUC = 0.991 (EyePACS-1), 0.990 (Messidor-2).
[68]	2017	Multiethnic cohorts [68]	Used a CNN-based architecture to detect referable DR, among other eye pathologies. Primary validation set: sensitivity = 90.5%, specificity = 91.6%.
[69]	2019	Beijing Tongren cohort [69]	Used the Adaboost algorithm to construct an integrated model that ensembled the performance of Inception V3, Resnet152 and Inception-Resnet-V2. Added weighted class activation maps (CAMs) for increased explainability.
[70]	2021	EyePACS [70], Thailand [71]	Proposed an Inception-v3–based architecture to predict the 2-year risk of developing diabetic retinopathy. AUC = 0.79 (3-field, internal), 0.70 (1-field, external). Combining with risk factors improved AUC to 0.81 (internal) and 0.71 (external).
[72]	2022	APTOS 2019 [73], DIARETDB1 [74].	Proposed an automated ensemble combining modified DenseNet101 and ResNeXt architectures. GAN-based data augmentation used. accuracy = 96.98% (2-class task), accuracy = 86.08% (5-class task).
[75]	2022	DeepDRiD [76], IDRiD [77], Messidor [67],	Proposed a framework for joint grading of DR and DME, using dynamic difficulty-aware weighted loss and dual-stream disentangled learning architecture.
[78]	2023	EyePACS [70], APTOS 2019 [73]	Combined Mask R-CNN for lesion detection and a CART-based decision tree model for interpretable 5-stage DR classification. Detected and segmented five lesion types. Accuracy = 0.94.
[64]	2024	Messidor [67]	Used a DFNN with a Lion Optimization Algorithm. Performed optic disk localization/removal. Accuracy = 98.04%, sensitivity = 99.69%.
[79]	2024	KAGGLE [80]	Developed a Sequential CNN with Adam optimizer to predict cataract, DR and glaucoma. Accuracy=76.34%.
[65]	2025	EyePACS [70], APTOS 2019 [73]	Proposed a stacked auto-encoder–based model for five-class DR detection. Achieved 93% training and 88% testing accuracy.

3.2 Fairness and biases

Bias against a certain individual or group means that such individual or group is consistently disadvantaged. Bias in machine learning is the tendency of an algorithm to systematically favor certain outcomes over others, resulting in unfair advantages or disadvantages for specific groups [9].

Fairness is the property of prediction models that do not discriminate against individuals or groups of individuals based on attributes such as age, race/ethnicity, sex/gender, or socioeconomic status [81].

3.2.1 Types of bias

Bias in machine learning can arise from multiple sources and affect model performance, fairness, and generalizability. The most relevant types of bias for this work are sample selection bias, labelling bias, and algorithmic bias:

Sample selection bias arises when the training data is not representative of the true population, often due to selection bias or systematic measurement errors. Even if samples perfectly represent the true population, minority groups are, by definition, unbalanced, leading to lower model accuracy for that minority, given that it is less significant on a larger scale [82].

Labelling bias refers to systematic favor or disfavor towards particular groups at the time of creating the target variable (from which the model is going to learn). This is common when labels are derived from subjective human decisions. For example, if past decisions were influenced by biased attitudes, the model trained on such labels may perpetuate those biases. In clinical datasets, this is less likely if the labels are derived from objective measurements, although diagnostic biases can still play a role [82].

Algorithmic bias occurs when the model itself reinforces or amplifies existing disparities. Even with unbiased data, a model can learn spurious associations due to the way it optimizes for performance. In medical imaging, algorithmic bias is particularly challenging to detect and mitigate, given the complex nature of learned feature representations. Medical images contain subtle and overlapping signals related to pathology, as well as to patient demographics and technical factors. This information may be implicitly encoded in the image, even if the attributes are not explicitly labeled, creating largely uninterpretable and untraceable sources of bias [83].

3.2.2 Types of fairness

- **Individual fairness:** requires that similar individuals receive similar outcomes [82].
  - *Advantages:* prevents both direct and indirect forms of discrimination and can identify unfair practices that group fairness metrics might miss.
  - *Disadvantages:* defining and measuring similarity can be difficult, especially in high-dimensional or complex data. Additionally, treating similar individuals similarly does not guarantee overall fairness.
  - *Metric example:* distance-based fairness [84].

- **Group fairness:** ensures that groups defined by sensitive attributes have similar outcomes. Most fairness metrics focus on group fairness, as these are easier to measure.
  - *Advantages:* well-defined, easy to implement metrics.
  - *Disadvantages:* focuses on predefined groups, overlooking intersectional sources of unfairness or high individual variability [85].
  - *Metric examples:* demographic parity, conditional demographic parity, equality of opportunity, equality of odds [82].
- **Causal fairness:** evaluates how sensitive attributes influence outcomes through direct or indirect pathways, aiming to prevent discrimination from these causal effects. It acknowledges that changing a sensitive attribute may also require adjusting related variables, if these are causally linked.
  - *Advantages:* as it focuses on causal relationships, it is capable of addressing multiple types of discrimination simultaneously.
  - *Disadvantages:* it requires accurate causal models, which are difficult to obtain from observational data. This approach requires extensive domain knowledge and richer datasets than those typically available in real-world scenarios.
  - *Metric examples:* counterfactual fairness, path-specific fairness [86].

3.2.3 Fairness metrics

The most commonly referenced fairness metrics in the literature are summarized in Table 3.2.

Table 3.2: Summary of fairness metrics applied in medical imaging.

Metric	Metric Type	Description
Demographic parity	Performance-independent and threshold-dependent	Ensures that the probability of a positive outcome as predicted by the model is equal across different demographic groups, regardless of ground truth labels.
Conditional demographic parity	Performance-independent and threshold-dependent	Allows demographic parity to hold when conditioned on relevant attributes (e.g., risk scores).
Equalized odds	Performance-independent and threshold-dependent	Requires that both true positive and false positive rates are equal across groups [87, 88].
Equal opportunity	Performance-independent and threshold-dependent	A relaxation of Equalized Odds; requires equal true positive rates across groups [89].
Equity-scaled AUROC (ES-AUROC)	Performance-independent and probability-based	Modifies the traditional performance metric AUROC to incorporate fairness considerations across subgroups [87, 90].
Performance-scaled disparity AUROC (PSD AUROC)	Performance-independent and probability-based	PSD AUROC is calculated as the standard deviation of group performance or the absolute maximum group performance difference divided by overall performance [87].
AUROC parity	Performance-dependent and probability-based	Quantifies disparity in discrimination using AUROC [87, 91, 92].
Subgroup net benefit	Performance-independent and threshold-dependent	Quantifies disparity in both discrimination and calibration [93].

3.2.4 Bias mitigation

Mitigating unfairness in medical DL models can be broadly categorized into three types, pre-processing methods, in-processing methods, and post-processing methods:

3.2.4.1 Pre-processing methods

Methods that aim to mitigate bias by transforming the data, to remove underlying sources of discrimination, before inputting it into the model. Common strategies include re-weighting, synthetic data

generation and data augmentation.

Shi et al. [90] propose **Fair Adaptive Scaling (FAS)**, a system that dynamically adjusts the significance of samples during model training for DR detection, minimizing performance disparities across identity groups in DR detection. It evaluated models on both AUROC and equity scaled-AUROC (ES-AUC). Baseline DL models showed significant performance differences between demographic groups. Integrating FAS led to improvements in both overall AUC and ES-AUC, especially for underrepresented groups. The approach was effective across different datasets and imaging modalities (2D and 3D) of DR, highlighting its generalizability.

Luo et al. [87] introduce **Fair Identity Scaling (FIS)**. This method aims to improve disparities at the individual and subgroup level, by incorporating both individual scaling and group scaling to determine dynamic loss weight during training. FIS improved overall performance and fairness in eye disease detection tasks across diverse model architectures and imaging modalities, including 2D and 3D models.

3.2.4.2 In-processing methods

In-processing methods aim to improve fairness by modifying the model's architecture or training process. This includes techniques such as disentanglement learning, adversarial debiasing, and integration of fairness-related penalties [94].

Adding fairness constraints or regularization terms to the loss function is a common approach. However, applying this approach to deep neural networks (DNNs) is challenging, due to their overparameterized nature. As DNNs are prone to overfitting, the fairness achieved during training may not generalize to the test data [95].

Du et al. [96] introduce **FairDisCo**, a disentanglement framework with contrastive learning, developed to improve fairness in automated skin lesion diagnosis. The architecture is divided into three main branches:

1. **Feature extractor:** Encodes the input images into a disease prediction feature vector;
2. **Sensitive attribute (SA) branch:** Forces the feature extractor to decouple SA information from the image.
3. **Contrastive branch:** Applies supervised contrastive learning. Uses a contrastive loss to promote intra-class cohesion and inter-class diversity, in order to prevent disentanglement from discarding important features for feature extraction.

The authors adapted traditional fairness metrics, such as demographic parity and equality of opportunity, for multi-class. Compared against standard fairness methods like resampling, reweighting, and attribute-aware modeling, FairDisCo achieved higher overall classification accuracy and significantly improved fairness across skin types.

3.2.4.3 Post-processing methods

Methods that utilize the model's outputs and sensitive attributes to calibrate the model's prediction during inference. These methods do not require retraining or modifying model architecture. Common strategies include:

- **Calibration:** Adjusting prediction thresholds for different subgroups to improve fairness.
- **Pruning:** Removing parameters or features strongly associated with sensitive information, reducing their influence on predictions [94].

3.3 Explainable AI in ophthalmology

For the successful integration of deep learning models into a medical context, it is essential for clinicians to trust and understand the decision-making processes of models. XAI plays a key role in medical domains, as it encompasses any method that tries to understand or explain how a model makes predictions.

To facilitate discussion, it is first necessary to distinguish between two interconnected but distinct concepts, interpretability and explainability. **Interpretability** refers to the field of techniques that aim to interpret the underlying functioning of an AI system. Meanwhile, **explainability** focuses on analysing the model's decision, to build trust that it is making correct, unbiased predictions. It does not necessarily clarify how the model works; instead, it helps understand the role that features play in obtaining the outcome.

Table 3.3 represents the four main types of XAI techniques: visual explanation, textual explanation, example-based explanation and latent space explanation.

Table 3.3: Explainable AI techniques and applications in eye fundus photography.

Technique type	**Description**	**Applications in eye fundus photography**
Visual explanation	Highlights important regions in images (e.g., saliency maps).	CAM [69, 97], Grad-CAM [98, 99], LIME [64], Deep SHAP, LRP
Concept-based explanation	Provides textual descriptions or reports (e.g., 'speculated masses', medical reports).	TCAV [100]
Example-based explanation	Shows similar data points or cases for comparison.	Rarely used in eye fundus photography.
Latent space explanation	Uses disentangled representations for compact, explainable data features.	Adversarial VAEs for disentangled representations [101]

3.3.1 Visual explanation (or saliency mapping)

Visual explanation is the most common form of XAI. It is a technique that visualizes the regions of an input image that are most important for the prediction. The main methods are:

- **Class activation mapping (CAM):** Uses global average pooling after the last convolutional layer, combining the activations with the learned importance, to determine the most important regions.
- **Gradient-weighted class activation mapping (Grad-CAM):** Extension of CAM. Uses the gradients of the output class with respect to the feature maps of the last convolutional layer. It is more flexible, as it does not require global average pooling right before the output.
- **Local interpretable model-agnostic explanations (LIME):** Explains individual predictions by perturbing the input data and observing changes in the output. It fits a simple, interpretable model (like linear regression), highlighting features that are most important for the model's decision locally.
- **Deep shapley additive explanations (Deep SHAP):** Adapts Shapley values from cooperative game theory to deep learning models.
- **Layer-wise relevance propagation (LRP):** The output is backwards propagated through the network, following the neural pathways involved in the final prediction.

3.3.2 Concept-based explanation

Form of XAI that provides textual descriptions to clarify model predictions. In medical imaging, such descriptions can range from relatively simple characteristics to more comprehensive summaries.

One notable method is **Testing with Concept Activation Vectors (TCAV)** [100]. TCAV quantifies the influence of high-level concepts defined by the user on predictions. A concept (e.g., "microaneurysm") is defined by providing example data. A linear classifier is then trained to distinguish the internal activations of the neural network associated with this concept. The TCAV score quantifies the degree of influence of a concept on the model's predictions for a specific class. This method was applied to a DL model trained to predict DR from fundus images. Using TCAV, clinical concepts such as "microaneurysm", "hemorrhage" and "exudate" were evaluated to assess their influence on model predictions. The analysis revealed that the model relied on clinically relevant features, but that it occasionally relied on concepts not emphasized by clinicians. TCAV analysis allowed for a more transparent understanding of the model's decision-making, important for model refinement and clinical validation.

3.3.3 Example-based explanations

XAI technique that clarifies a model's prediction by providing representative examples from the dataset that are the most similar to the input being analyzed. Although this approach has been explored in

various medical imaging domains [102, 103], there is currently limited application to eye fundus photography.

3.3.4 Latent space explanation

XAI technique aimed at interpreting how a model's internal feature representations correspond to meaningful factors in the data. It achieves this by learning disentangled feature representations, where each latent dimension encodes a distinct and interpretable factor of variation. This leads to more compact, explainable representations of the data.

Sarhan et al. [101] introduce a novel approach for learning interpretable and disentangled latent representations, using an **adversarial variational autoencoder (VAE)** with a total correlation constraint. It encourages the model to learn disentangled representations, while preserving high reconstruction fidelity. It is evaluated on a skin lesion dataset, demonstrating its ability to separate latent factors that correspond to meaningful clinical attributes. This leads to increased explainability, and is of particular interest for fairness and bias mitigation, as it helps in identifying and isolating confounding factors.

3.4 Shortcut learning in ophthalmology

Shortcut learning occurs when two conditions are met: (1) a machine learning model relies on spurious features, such as sensitive attributes, instead of learning the underlying medical features, and (2) this effect does not considerably improve performance but harms fairness [88].

3.4.1 Sensitive attributes encoded in medical images

This issue is particularly relevant in medical imaging, and especially in the context of ophthalmology, as retinal images can encode a wide range of information beyond explicit clinical features. Deep learning models trained on fundus photographs have been shown to extract sensitive attributes, such as sex [10, 104], age [10], and cardiovascular risk factors from fundus photographs [105], even when these attributes are not explicitly included as inputs. Additionally, variations in how medical images are captured (e.g., lighting and camera type) can introduce systematic biases into the images, leading to spurious feature correlations of these factors with sensitive patient attributes, such as ethnicity [106].

3.4.1.1 Is this information inherently bad?

Not all correlations between sensitive attributes and disease prediction are problematic. In some scenarios, such relationships are clinically meaningful and contribute to increased diagnostic accuracy. For example, androgenetic alopecia is more prevalent in men, breast cancer predominantly affects women, keloid scarring occurs more frequently in skin of color, and melanoma is more common in lighter skin tones [88]. In these cases, ignoring or removing these attributes could decrease model performance.

However, problems arise when the effect of sensitive attributes contributes to the model performing poorly in different domains, failing to generalize across populations or clinical settings, and producing biased predictions.

Additionally, DR datasets are often imbalanced, with uneven representation of disease stages or demographic groups. This can exacerbate shortcut learning, as models may learn dataset-specific associations that do not generalize well to new populations or clinical settings [107].

Moreover, even training on large, well-balanced datasets is not enough to guarantee fairness or generalization. Disparate impact can still occur through confounding variables - third variables related to both the dependent variable and the independent variable. These variables can distort the true relationship between the predictors and the target, making it difficult to assess a potential causal effect [108]. For example, age is a common confounder, as it is associated with retinal thinning and increased disease risk in diabetic retinopathy. A model might learn to associate retinal thinning directly with disease, without recognizing age as the underlying cause.

3.4.2 Relationship between shortcut learning and fairness

There is a strong connection between shortcut learning and algorithmic fairness, although their scopes differ slightly:

- **Shortcut learning** encompasses the reliance of the model on any spurious feature that may be used as a shortcut. It may not lead to fairness concerns.
- **Algorithmic fairness** specifically addresses the issue of unfair model behavior linked to sensitive attributes, such as age, sex, or ethnicity. A model is considered unfair if its predictions lead to unequal performance across different demographic groups.

Algorithmic bias can be a cause of shortcut learning. This is especially true in the medical imaging domain, as sensitive attributes are often encoded in the data. In these cases, the model may use sensitive features as predictive shortcuts, leading to both unfairness and decreased domain generalization. By reducing the model's dependence on sensitive attributes, we can simultaneously promote fairness and reduce the risk of shortcut learning [109].

3.4.3 Detecting shortcut learning

Identifying if unfairness is caused by shortcut learning or other factors is challenging. Brown et al. address this issue with a multitask learning framework called **ShorT (Shortcut Testing)**, which directly tests the presence and impact of shortcut learning in clinical ML systems [88].

Firstly, ShorT controls the degree to which a model encodes a sensitive attribute. This is achieved by adding a demographic prediction head (a neural network branch trained to predict the sensitive attribute) and using variable gradient scaling to control how much the model encodes this information.

By systematically increasing or decreasing the encoding strength, the authors assess its effect on model fairness. If a stronger encoding of a sensitive attribute increases unfairness, this is considered a signature of shortcut learning.

Secondly, the authors also investigate the role of dataset bias in fairness. With this aim, two training datasets are constructed via subsampling. In the biased dataset, the majority of cases were over-represented (e.g., older patients with effusion, younger patients without effusion) and minority cases were under-represented. In the balanced dataset, prevalence differences between groups were removed. Even with similar task performance, the biased dataset led to much lower fairness, while the balanced dataset improved fairness.

The first finding of the paper was that shortcut learning is not always responsible for unfairness. In some cases, reducing shortcut encoding does not improve fairness, indicating that unfairness can arise from other sources. The second finding was that holistic fairness mitigation is needed, as addressing shortcut learning alone may not be sufficient; a broader approach to fairness is necessary in clinical ML.

3.5 Disentanglement in Representation Learning

Disentanglement refers to the process of separating the underlying independent factors of variation within data. A disentangled representation encodes the true underlying factors, for example, medical information, demographic information and technical imaging characteristics, into distinct dimensions of a model's latent space. Disentangled representations can offer many advantages, including improving predictive performance [110], facilitating model interpretability [101, 111], improving fairness [96, 112] and mitigating shortcut learning [11].

In the ophthalmology domain, disentanglement has been applied to several tasks, including anonymization and de-identification of retinal images [113, 114] and domain generalization across imaging devices and clinical sites [106, 115, 116]. While the impact of disentanglement in enhancing fairness has been explored in other medical imaging domains [11, 96], it remains underexplored in ophthalmology.

Xia et al. [115] present **DECO**, a framework designed to improve domain generalization for DR detection by disentangling retinal image features into medical and domain-specific features. It employs a ResNet-50 backbone (pre-trained on ImageNet), followed by a fully connected classifier. It uses data-aware weights to emphasize rare classes and domains. Moreover, it uses an adversarial loss to enforce disentanglement, and a custom pixel-level alignment loss to encourage pixel-level consistency while preserving medically relevant details. DECO also introduces a novel augmentation strategy. By decoupling and recombining semantic and domain features from different images, the model generates new, diverse representations. This augmentation method outperforms traditional domain generalization techniques.

Montenegro et al. [113] propose a novel disentanglement-based method for anonymization of medical images. The approach uses an autoencoder architecture that decomposes images into three independent latent vectors that encode medical, identity, and remaining features, respectively. To achieve

disentanglement, the model includes a disease classifier, an identity classifier, and a disentanglement loss that enforces independence between the features. In the anonymization task, the identity vector is replaced with a synthetic identity generated by a VAE decoder. The modified latent representation is then used to produce an anonymized version of the original image. The method is evaluated on multiple medical imaging datasets, including chest radiographs, iris images, and facial images.

3.6 Summary

Although some research has been conducted on detection and mitigation of fairness issues in DR prediction with DL models, there is still ample room for growth. Factors deterring fairness progress include the lack of a widely accepted definition of fairness, as numerous fairness metrics have been proposed. Furthermore, most publicly available DR datasets lack annotations for sensitive attributes, limiting the scope of fairness-focused studies. Beyond fairness mitigation techniques, developing diverse datasets is also essential to address the issue of algorithmic bias. Models may exhibit bias toward Western populations, due to data availabity. This imbalance can decrease the generalizability of models and introduce or amplify biases against underrepresented groups. Disentanglement techniques have emerged as a promising method to address fairness issues, although their application in DR prediction remains underexplored.

Chapter 4

Methodologies

4.1 Overview of workflow

The framework for this work consisted of four main stages:

1. **Training of tabular models:** Machine learning models were developed using only tabular demographic and clinical features to perform binary classification of DR. This was done to understand the predictive power of the tabular features.

2. **Training and fairness evaluation of image-based models:** Three deep learning models were trained on retinal images to perform binary classification for DR. Afterwards, the models were evaluated on performance and fairness across predefined groups. Additionally, the models were also trained to predict sensitive attributes.

3. **Investigating the impact of feature disentanglement in bias mitigation:** Using a disentanglement network, models were trained to separate medical features from specified sensitive attributes. These models were evaluated both in terms of fairness and performance, and were compared against baseline image-based models.

4. **Validating the disentanglement architecture:** To understand the effectiveness of the disentanglement architecture, a series of validation studies were conducted. These included assessments of the model's ability to predict sensitive attributes after disentanglement, ablation tests on the latent space dimensionality, and experiments involving architectural variations.

4.2 Datasets

4.2.1 BRSET

The Brazilian Multilabel Ophthalmological Dataset (BRSET), introduced in [117], was developed with the goal of improving the representation of underrepresented populations in ophthalmological datasets.

BRSET consists of 16,266 fundus photos images from 8,524 Brazilian patients, captured with a Nikon NF505 and a Canon CR-2. The dataset includes demographic and clinical information, such as age, sex, and duration of diabetes. For each patient, it includes a macula-centered paired exam. The attributes relevant to this work are summarized in Table 4.1. We refer the reader to [117] for an in-depth analysis of the remaining attributes. Example images from BRSET are shown in Figure 4.1 and Figure 4.2.

4.2.2 mBRSET

The mBRSET dataset, introduced in [118], is the first publicly available dataset of images captured using handheld retinal cameras. The images were acquired using Phelcom Eyer, a portable, handheld retinal fundus camera. It contains 5,164 images from 1,291 patients of diverse ethnic backgrounds in Brazil, addressing the underrepresentation of populations from low- and middle-income countries in current ophthalmological datasets. It includes a macula-centered paired exam (one image per eye) and an additional optic disc-centered exam for each patient.

In addition to retinal images, mBRSET includes demographic information, such as sex, age and insurance status, for fairness and generalizability of analysis. The attributes relevant to this work are represented in Table 4.2. We refer the reader to [118] for an in-depth analysis of the remaining attributes. Example images from mBRSET are shown in Figure 4.3 and Figure 4.4.

Table 4.1: Main attributes available in the BRSET Dataset.

Field	**Values**	**Description**
image_id	String	Image identifier
patient_id	Integer	Patient identifier
camera	String	Retinal camera used (Canon CR or NIKON NF5050)
patient_age	Integer	Age of the patient in years
diabetes_time	Integer	Self-referred duration of diabetes diagnosis in years
insulin_use	0 = No, 1 = Yes	Use of insulin
patient_sex	1 = Male, 2 = Female	Gender of the patient
exam_eye	1 = 'right eye', 2 = 'left eye'	Eye being examined
DR_ICDR	0 = No retinopathy, 1 = Mild non-proliferative, 2 = Moderate non-proliferative, 3 = Severe non-proliferative, 4 = Proliferative diabetic retinopathy and post-laser status	International Clinic Diabetic Retinopathy classification
quality	'Adequate' or 'Inadequate'	Final quality assessment of the retinal image

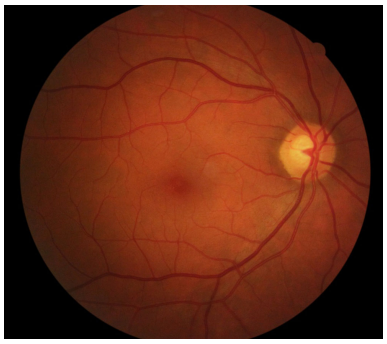


Figure 4.1: Example of healthy eye from BRSET (ICDR=0).

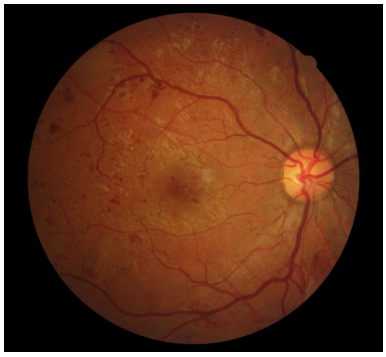


Figure 4.2: Example of eye with diabetic retinopathy from BRSET (ICDR=4).

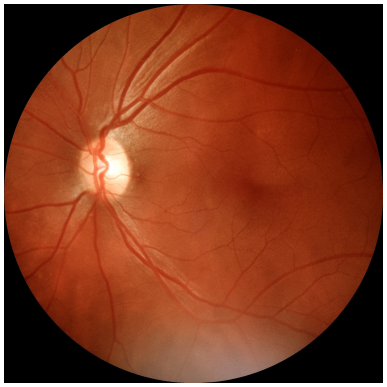


Figure 4.3: Example of healthy eye from mBRSET (ICDR=0).

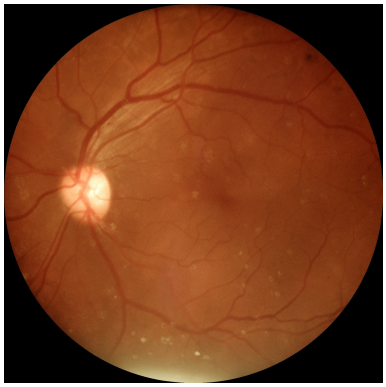


Figure 4.4: Example of eye with diabetic retinopathy from mBRSET (ICDR=4).

4.3 Exploratory data analysis

We conducted an exploratory analysis to investigate the relationships between the patient attributes in the BRSET and mBRSET datasets, to identify highly correlated features that could influence model behavior.

To ensure consistency, samples with missing values in any of the selected attributes were removed. For both datasets, DR severity was binarized into two classes **Normal** (ICDR levels 0 or 1) and **Referable** (levels 2, 3 or 4), aligning with the clinical referral threshold. Continuous variables such as patient age, duration of diabetes diagnosis, and insulin treatment duration were scaled to the range [0, 1] using min–max normalization. Age values reported as "$\geq$ 90" were replaced with 91 before scaling. In the mBRSET dataset, educational level was binarized, distinguishing individuals with at least completed primary education from those with incomplete or no formal primary education. A Pearson correlation matrix, which quantifies linear associations between all pairs of selected variables, was then computed and visualized using a heatmap.

Table 4.2: Main attributes available in the mBRSET Dataset.

Field	**Values**	**Description**
file	String	Image identifier
patient	Integer	Patient identifier
age	Integer	Age of the patient in years
sex	0 = Female, 1 = Male	Gender of the patient
dm_time	Integer	Duration of diabetes diagnosis in years
insulin_time	Integer	Duration of insulin use in years
oraltreatment_dm	0 = No, 1 = Yes	Use of oral drugs for diabetes
systemic_hypertension	0 = No, 1 = Yes	Diagnosis of systemic arterial hypertension
insurance	0 = No, 1 = Yes	Health insurance status
educational_level	1 = Illiterate, 2 = Incomplete Primary, 3 = Complete Primary, 4 = Incomplete Secondary, 5 = Complete Secondary, 6 = Incomplete Tertiary, 7 = Complete Tertiary	Highest level of education achieved
alcohol_consumption	0 = No, 1 = Yes	Regular alcohol consumption
smoking	0 = No, 1 = Yes	Active smoking status
obesity	0 = No, 1 = Yes	Diagnosis of obesity
laterality	'Left' or 'Right'	Laterality of retinal fundus photo
final_quality	'Yes' or 'No'	Final quality assessment of the retinal image
final_icdr	0 = No retinopathy, 1 = Mild non-proliferative, 2 = Moderate non-proliferative, 3 = Severe non-proliferative, 4 = Proliferative or post-laser status	ICDR score for retinopathy assessment

4.4 Tabular models for DR prediction

Tabular models were employed to assess the predictive value of sensitive attributes for DR classification. Using the models, we investigated whether DR could be accurately predicted using demographic and clinical features alone, without relying on retinal images.

4.4.1 Architectures

Logistic Regression (LR) is a supervised machine learning algorithm used for classification tasks. It models the relationship between input features and a binary outcome by applying a sigmoid function, which transforms the linear combination of input features into a probability value between 0 and 1. This probability indicates the likelihood that a given instance belongs to the positive class.

Decision Trees (DT) are a non-parametric, supervised learning method used for classification and regression tasks. They learn simple decision rules inferred from the data features. The term non-parametric indicates that the model does not assume a predefined functional form that is fixed, nor a fixed number of parameters, to define the relationship between the input data and the output labels. A decision tree follows a hierarchical structure, where each internal node represents a decision based on a specific feature, each branch represents a possible outcome of that decision, and each leaf node represents the prediction outcome (see Figure 4.5).

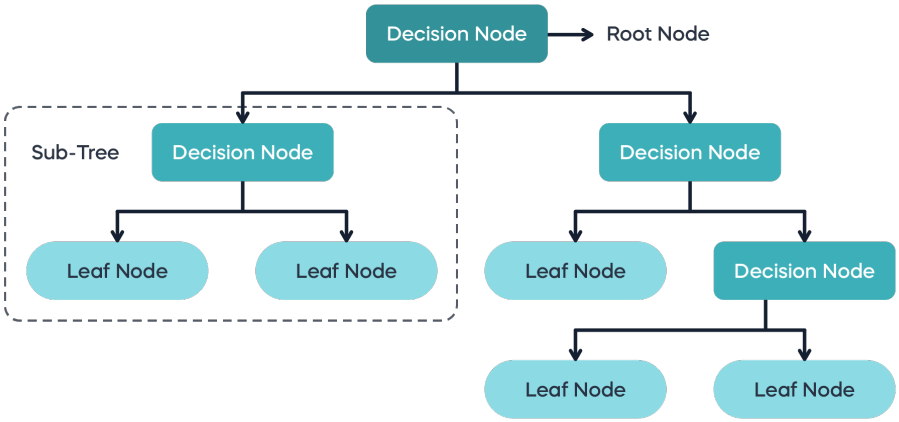


Figure 4.5: Decision tree structure (from [119]).

eXtreme Gradient Boosting (XGBoost) is a boosting algorithm that combines multiple weak decision trees sequentially. Each new tree is trained to correct the residual errors made by the previous models. Following hundreds of iterations, the weighted combination of weak learners forms a strong learner. It is a leading machine learning model for classification and regression tasks (see Figure 4.6).

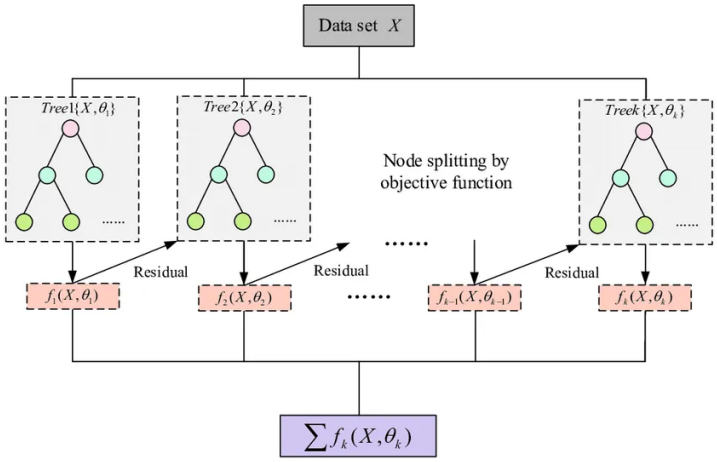


Figure 4.6: eXtreme Gradient Boosting (from [120]).

4.4.2 Pipeline

All features were preprocessed as described in Section 4.8.1. Models were implemented using the `scikit-learn` library (for DT and LR) and the `xgboost` library (for XGBoost). Each model was trained on three datasets: mBRSET, BRSET and a restricted mBRSET, containing only the subset of attributes shared with BRSET. Hyperparameter tuning was performed using grid search to determine the optimal configurations of each model.

4.5 Image models

4.5.1 Image selection and preprocessing

For the binary classification task, DR severity was categorized based on the ICDR scale. Labels were binarized into two classes: **Normal** (ICDR levels 0 or 1) and **Referable** (levels 2, 3 or 4). In the mBRSET dataset, each patient has both macula-centered and optic disc-centered fundus images. To ensure consistency in the anatomical region being analyzed, only the macula-centered images were used. For both mBRSET and BRSET datasets, only high-quality fundus images were selected. All images were resized to 224×224 pixels and normalized using the mean and standard deviation values of the ImageNet dataset, to align with pretrained model expectations.

The dataset was split into training (70%), validation (10%), and testing (20%) subsets. The data was stratified to preserve the original distribution of DR across subsets. Additionally, all images from the same patient were assigned to the same subset to prevent data leakage. The number of samples and the DR prevalence for each subset, across both the mBRSET and BRSET datasets, are summarized in Table 4.3.

Table 4.3: Number of samples and DR prevalence for each data subset, for both mBRSET and BRSET datasets.

Dataset	**Subset**	**N**	**DR Prevalence (%)**
BRSET	Train	4884	4.85
	Validation	701	3.28
	Testing	1368	4.68
mBRSET	Train	1685	18.40
	Validation	247	29.96
	Testing	498	17.07

4.5.2 Data augmentation

To improve model generalization and reduce overfitting, the following transformations were randomly applied during training:

- Random rotations up to 20º;
- Vertical flips;
- Color jitter, randomly changing brightness, contrast, saturation, and hue;
- Gaussian blur.

4.5.3 Image Model Architectures

Three image-based deep learning architectures, including a convolutional neural network, ConvNeXt V2, and vision transformer-based models, DINOv2 and Swin V2, were selected for binary classification of DR and trained with the mBRSET and BRSET datasets.

4.5.3.1 ConvNeXt V2

Introduced in [121], ConvNeXt V2 is a fully convolutional neural network that modernizes traditional CNNs, by incorporating design principles inspired by vision transformers (see Figure 4.7). The architecture is co-designed with a fully convolutional masked autoencoder (FCMAE) for self-supervised pretraining. This enables the model to learn robust feature representations by reconstructing masked regions of input images, a similar approach to masked autoencoding in transformer models. Additionally, it adds a new Global Response Normalization (GRN) layer, which enhances inter-channel feature competition and prevents feature collapse during masked pretraining.

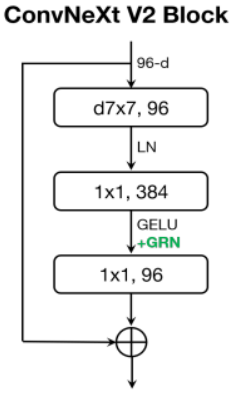


Figure 4.7: ConvNeXt V2 Block Design.

4.5.3.2 DINOv2

DINOv2 is a self-supervised learning framework for visual representation. It is built upon the vision transformer (ViT) architecture and uses a self-supervised approach based on knowledge distillation. The training process includes multi-crop augmentation and is conducted on large datasets. DINOv2 produces highly generalizable and robust visual features, and surpasses the performance of ResNet and the original ViT [122].

4.5.3.3 Swin V2 Transformer

Swin V2 is a state-of-the-art hierarchical transformer-based model that builds upon the original Swin Transformer architecture. Its main innovation is the use of a shifted window attention mechanism, which applies self-attention within local windows that shift across layers. This design efficiently captures both local and global image features, addressing key limitations of vision transformers, such as the high computational cost of global attention on high-resolution images. Swin V2 achieved superior performance to previous state-of-the-art models, making it a powerful backbone for computer vision tasks [123].

4.5.4 Training strategy and hyperparameters

To address the class imbalance present in the datasets, we employed the focal loss function (see Equation 4.1), a modified version of the binary cross-entropy loss that emphasizes hard-to-classify cases. In the focal loss formula, p_t represents the model's estimated probability for the true class. The parameter $\alpha_t \in [0,1]$ is a weighting factor that balances the importance of classes, typically assigning higher weight to the minority class to mitigate imbalance. The focusing parameter $\gamma \geq 0$ controls the rate at which easy examples are down-weighted: increasing γ gives more focus to hard-to-classify cases by reducing the loss contribution of well-classified examples [124].

Class weights were computed from the training data distribution and incorporated into the loss function to further reduce the influence of dominant classes.

$$\mathscr{L}_{\text{focal}}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{4.1}$$

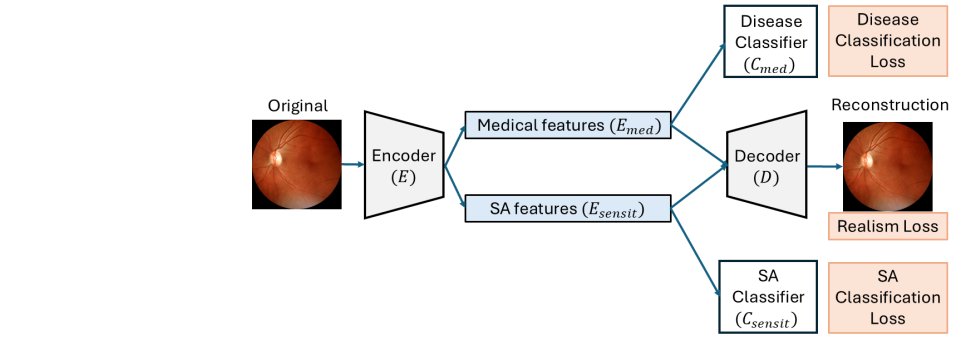
Model training was performed using the Adam optimizer [125]. To prevent overfitting, early stopping based on F1-score validation was implemented. The key hyperparameters used in the task are summarized in Table 4.4. The training process was similar for the two datasets, with small adjustments to the batch size to accommodate GPU memory usage.

Table 4.4: Hyperparameters used in training of baseline image models.

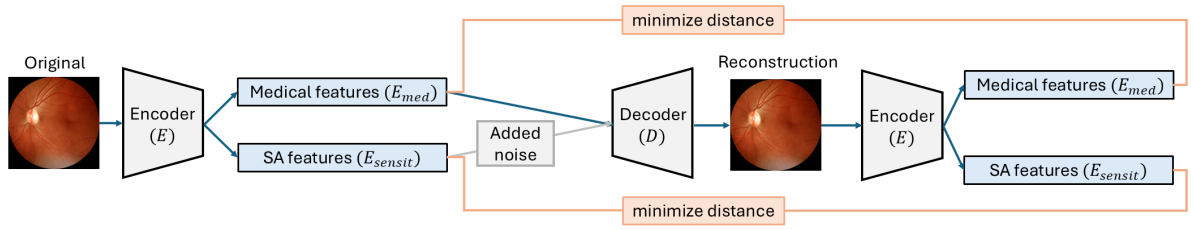
Dataset	Arch.	Hidden Dim.	LR	Batch size	Num workers
mBRSET	ConvNeXt V2	128	1e-5	4	4
	DINOv2	128	1e-5	4	4
	Swin V2	128	1e-5	2	4
BRSET	ConvNeXt V2	128	1e-5	8	4
	DINOv2	128	1e-5	4	4
	Swin V2	128	1e-5	2	4

4.6 Disentanglement architecture

The next step was to remove possible confounding information from the image via disentanglement, with the goal of decreasing model bias. To achieve this, we developed a disentangling architecture that could separate the medical information from the information of a specified sensitive attribute. This work is based on the architecture proposed in [113], which uses a generative model for disentangling identity and medical characteristics in images. The adapted architecture is represented in Figure 4.8.



(a)



(b)

Figure 4.8: Overview of the disentanglement network. (a) Representation of the components of the network, without the disentanglement loss. (b) Representation of the disentanglement loss.

The four main components of the network are:

- **Encoder** (E): Transforms the input image (I_{ori}) into two independent vectors, (E_{med}), that encodes the medical features, and (E_{sensit}), that encodes the sensitive attribute features.
- **Diabetic Retinopathy Classifier** (C_{med}): Guides E_{med} to capture medically relevant information by classifying diabetic retinopathy severity.
- **Sensitive Attribute Classifier** (C_{sensit}): The sensitive attribute classification task is used to approximate the E_{sensit} vector to the geometric space of the sensitive attribute features.
- **Decoder** (D): Trained to reconstruct the original image from the vectors E_{med} and E_{sensit}, promoting realistic reconstruction and enforcing the utility of both vectors.

The model is trained using a total loss function that combines three loss terms, each with a specific objective. These components are summed to form the total training objective, which balances the contributions of each term through weighting factors (see Table 4.5 for the hyperparameters used in training). The loss terms are:

- **Disentanglement Loss** $(\mathscr{L}\text{dis})$: Represented in Equation (4.2), promotes independence between latent vectors E_{med} and E_{sensit}. The loss penalizes scenarios where modifying a latent vector affects

the other. Gaussian noise is added to only one of the latent vectors to simulate perturbations, and a new, altered image (I_i) is generated. If the image is generated by altering medical features, the image is denoted as (I_{med}); if the image is generated by altering the sensitive attribute vector, the image is denoted as (I_{sensit}). $\{I \sim p_d(I)\}$ indicates that the image I is sampled from the data distribution $p_d(I)$. $\mathbb{E}_{I\sim p_d(I)}$ indicates the expected value, i.e., the average computed by sampling images I from the distribution $p_d(I)$.

The loss consists of two components:

- **Self-consistency term:** ensures that the altered latent feature (e.g., E_{med}) is preserved when re-extracted from the generated image.
- **Independence term:** ensures that the unaltered feature (e.g., E_{sensit}) remains unaltered, promoting disentanglement.

$$\begin{aligned}\mathscr{L}_{\text{disent}} = \mathbb{E}_{I\sim p_d(I)}\Big[\textstyle\sum_{i\in\{\text{med,sensit}\}}\Big(\left(E_i\left(I_{\text{ori}}\right)-E_i\left(I_i\right)\right)^2 \\ +\textstyle\sum_{\substack{j\in\{\text{med,sensit}\}\\ j\neq i}}\left(E_j\left(I_{\text{ori}}\right)-E_j\left(I_i\right)\right)^2\Big)\Big]\end{aligned} \tag{4.2}$$

- **Classification Loss** ($\mathscr{L}_{\text{cls}}$): Ensures that E_{med} and E_{sensit} encode the intended features by evaluating their effectiveness on disease and sensitive attribute classification tasks, respectively. Represented in Equation (4.3).

$$\begin{aligned}\mathscr{L}_{\text{classifier}} = -\lambda_{\text{med}}\textstyle\sum_c y_{\text{med}}(c)\log\left(p_{\text{med}}(c)\right) \\ -\lambda_{sensit}\textstyle\sum_k y_{sensit}(k)\log\left(p_{sensit}(k)\right)\end{aligned} \tag{4.3}$$

- **Realism Loss** ($\mathscr{L}_{\text{real}}$): Promotes high-quality image reconstruction by optimizing pixel-level similarity. This loss combines the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR), normalized with a threshold parameter $\alpha = 48$ for PSNR. Represented in Equation (4.4). $D\left(E\left(I_{\text{ori}}\right)\right)$ represents the reconstructed image (I_{reconst}).

$$\mathscr{L}_{\text{realism}} = \mathbb{E}_{I\sim p_d(I)}\left[\left(1-\text{SSIM}\left(I_{\text{ori}}, D\left(E\left(I_{\text{ori}}\right)\right)\right)\right)+\left(1-\frac{1}{\alpha}\text{PSNR}\left(I_{\text{ori}}, D\left(E\left(I_{\text{ori}}\right)\right)\right)\right)\right] \tag{4.4}$$

- **Total Loss** ($\mathscr{L}_{\text{total}}$): Represented in Equation (4.5), combines the three loss terms into a total training objective. The relative contribution of the realism and disentanglement losses is controlled by the weighting factors λ_r and λ_d, respectively.

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{classifier}} + \lambda_r\mathscr{L}_{\text{realism}} + \lambda_d\mathscr{L}_{\text{disent}} \tag{4.5}$$

Dataset	Arch.	Latent Dim.	Batch size	Learning Rate	Decay	λ_r	λ_d
mBRSET	All	256	32	5e-5	1e-6	1	5

Table 4.5: Hyperparameters used in training of the disentangled architecture.

4.6.1 Validating the disentanglement architecture

We developed a method to evaluate the effectiveness of the disentanglement architecture in separating medically relevant information from sensitive attributes.

This evaluation is based on analyzing AUROC scores across four tasks:

1. **Med→Med:** Using only medical features extracted from the DR prediction branch to predict DR;
2. **Med→SA:** Using only medical features from the DR branch to predict sensitive attributes;
3. **SA→SA:** Using only SA features extracted from the SA prediction branch to predict sensitive attributes;
4. **SA→Med:** Using only SA features to predict DR.

Effective disentanglement is indicated when the DR classifier cannot predict DR from SA-related features (SA $\rightarrow$ Med), and the SA classifier cannot predict sensitive attributes from medical features (Med $\rightarrow$ SA), resulting in AUROC scores close to 50% (random chance). At the same time, strong performance is expected when predicting DR from medical features (Med $\rightarrow$ Med) and SA from SA features (SA $\rightarrow$ SA).

4.6.2 Alterations to the disentanglement architecture

To help validate the proposed disentanglement architecture and better understand the impact of each loss individually, we explored three variants (illustrated in Figure 4.9). For each variation, we evaluated the AUROC achieved by the DR classifier when using only the medical-related features. These results were compared against the baseline model and the full disentanglement architecture.

4.7 Performance Evaluation

4.7.0.1 Discrimination

Discrimination refers to the ability of a model to distinguish between individuals with and without the target condition. In the context of diabetic retinopathy, a model with perfect discrimination consistently assigns higher predicted probabilities to patients with referable DR than to those without. In other words, all true positive cases are ranked above negative cases.

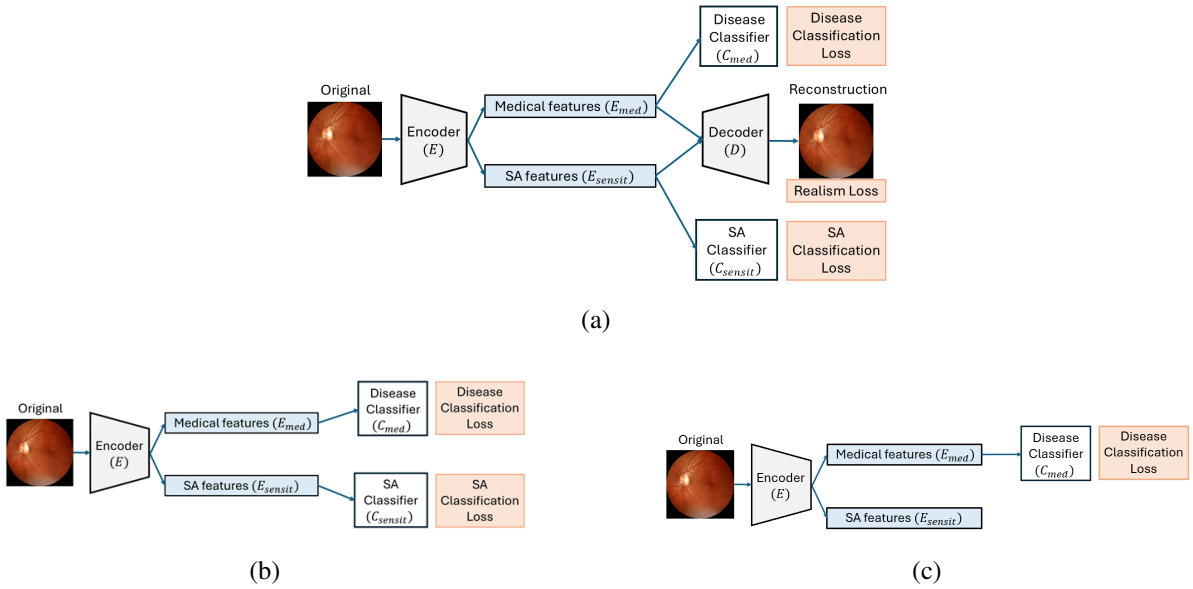


Figure 4.9: Variants of the proposed disentanglement architecture used to assess the individual contribution of each loss component. (a) Removal of the disentanglement loss while retaining reconstruction and classification losses. (b) Removal of both disentanglement and reconstruction losses, relying solely on the classification loss. (c) Use of only the DR classification loss with separated latent spaces for DR and SA information, but without applying SA classification, reconstruction, or disentanglement losses.

Area Under the Receiver Operating Characteristic Curve (AUROC) was used to quantify discrimination. An AUROC of 0.5 means a model has no discriminatory ability, and 1 indicates a model with perfect discrimination. The ROC curve, which plots the true positive rate against the false positive rate at various threshold settings, may also be used to assess discrimination.

4.7.0.2 Calibration

Calibration reflects how well a model's predicted probabilities reflect the observed target condition proportions. For instance, in a well-calibrated model, if a group of individuals is assigned a 70% risk of DR, then approximately 70% of them have a DR diagnosis.

Calibration plots are typically used to assess this performance domain. This metric reflects the relationship between predicted and observed probabilities. Predictions that fall above the diagonal line indicate that the model is underconfident, while those below the diagonal line indicate an overconfident model. However, in this work, calibration plots proved difficult to interpret due to variability and limited subgroup sample sizes. As an alternative, we relied on risk distribution plots to assess calibration, as these offered a more intuitive visualization.

4.7.0.3 Overall performance

Overall performance assesses how closely the estimated probabilities of the model (ranging from 0 to 1) align with the actual binary outcomes (0 or 1).

To assess this parameter, risk distribution plots were employed. These plots visually represent the distribution of predicted probabilities for each outcome class, intuitively assessing both calibration and discrimination. Ideally, the risk distribution plot of a well-performing model demonstrates both a clear separation between the predicted probability distributions between outcome groups (discrimination) and a close alignment between predicted probabilities and true outcomes (calibration).

4.7.0.4 Clinical utility

Clinical utility evaluates how effectively a predictive model improves real-world healthcare decisions. It incorporates misclassification costs.

It is commonly assessed using Decision Curve Analysis (DCA). DCA plots the net benefit of the model against a range of threshold probabilities, comparing it to two default strategies: treating all patients as if they are at risk and treating none of the patients as if they are at risk [126].

A predictive model has clinical utility if its net benefit curve lies above both strategies for a meaningful threshold range. Net benefit quantifies the trade-off between true positive and false positive classifications, and is computed as follows (see Equation 4.6):

$$\text{Net Benefit} = \frac{TP}{N} - \frac{FP}{N} \cdot \frac{pt}{1-pt} \tag{4.6}$$

Here, TP denotes the number of true positive cases, FP the number of false positives, N the total patient count, and p_t the decision threshold probability.

4.8 Fairness and Bias Assessment

We evaluate the fairness of the predictive models, using a group fairness approach, which assesses whether the models perform equitably across different population subgroups. This approach was chosen for its clinical relevance, ease of interpretation, and the availability of well-established evaluation metrics. To assess subgroup performance, we employed AUROC, DCA, and risk distribution plots (see Section 4.7).

4.8.1 Subgroups of mBRSET

The fairness analysis for the mBRSET dataset focused on detecting performance disparities across five sensitive attributes: age, sex, educational level, insurance status, and obesity. Obesity was also chosen to include a clinical attribute in the analysis, as it is recognized as a risk factor for diabetes. To facilitate subgroup comparisons, educational level was simplified into two categories, literate (individual had at

least complete primary education) and illiterate (illiterate or incomplete primary education), and age was split into two groups, individuals aged ≤ 50 years (Group 0) and those aged > 50 years (Group 1). Table 4.6 indicates the selected subgroups within the mBRSET dataset, along with information about their respective sample size and prevalence rates of DR.

Table 4.6: Attributes and Subgroups of mBRSET (Test set).

Attributes of mBRSET	**Subgroup**	**N**	**DR prevalence(%)**
Age	Group0 (below or equal to 50)	81	16.05
	Group1 (above 50)	417	17.27
Sex	Group0 (female)	329	13.98
	Group1 (male)	169	23.08
Educational level	Group0 (literate)	222	14.86
	Group1 (illiterate)	268	19.40
Insurance	Group0 (no insurance)	455	17.80
	Group1 (insurance)	39	10.26
Obesity	Group0 (no obesity)	450	17.33
	Group1 (obesity)	40	12.50

4.8.2 Subgroups of BRSET

The fairness analysis for the BRSET dataset focused on detecting performance disparities across two sensitive attributes: age and sex. Similarly to mBRSET, age was split into two groups, individuals aged ≤ 50 years (Group 0) and those aged > 50 years (Group 1). Table 4.7 indicates the selected subgroups within the BRSET dataset, along with information about their respective sample size and prevalence rates of DR.

Table 4.7: Attributes and Subgroups of BRSET (Test set).

Attributes of BRSET	**Subgroup**	**N**	**DR prevalence(%)**
Age	Group0 (below or equal to 50)	597	4.02
	Group1 (above 50)	1395	6.09
Sex	Group0 (female)	1802	4.00
	Group1 (male)	1095	8.95

Based on [127], we chose to evaluate the model across four key performance domains: discrimination, calibration, overall performance, and clinical utility. Each domain reflects a distinct aspect of model performance, providing a comprehensive assessment.

Chapter 5

Results and Discussion

5.1 Exploratory data analysis

First, correlation analyses were conducted in order to detect possible dependencies between attributes. In the mBRSET dataset (see Figure 5.1), the most significant correlations were:

- 0.52 between diabetes duration and insulin usage duration, which is intuitive given that patients with a longer history of diabetes are more likely to have had insulin treatment over extended periods;
- 0.29 between hypertension and age, consistent with the role of age as a risk factor for hypertension;

In the BRSET dataset (see Figure 5.2), the most significant correlations were:

- 0.59 between insulin use and patient age, indicating that older patients are less likely to be on insulin therapy. This aligns with existing literature that suggests older adults may be less frequently treated with insulin, due to decreased functional ability and comorbidities [128];
- 0.42 between ICDR score and duration of diabetes, which is expected given that longer diabetes duration increases the risk of developing diabetic retinopathy;
- 0.36 between diabetes duration and patient age, an expected correlation since age is a risk factor for diabetes;
- 0.35 between ICDR score and patient age, which is consistent with increasing age being associated with a higher risk of DR.

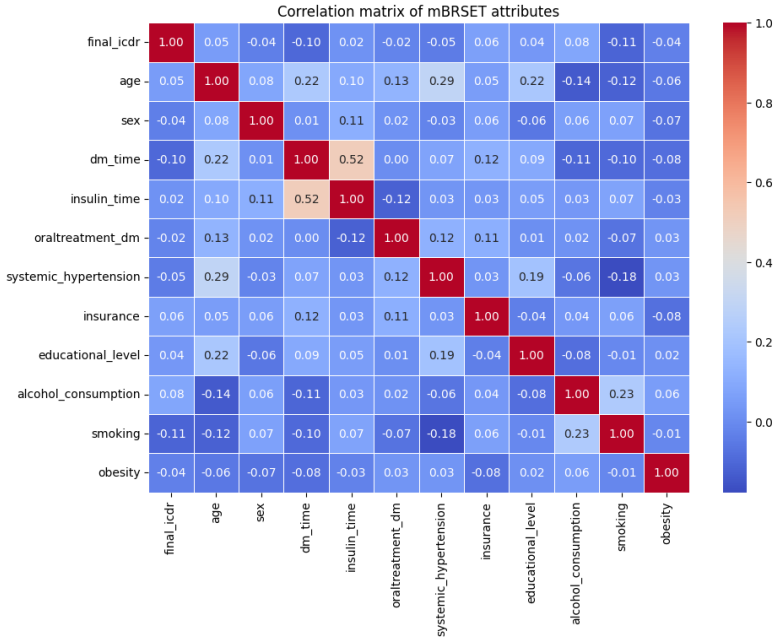


Figure 5.1: Correlation matrix between attributes of mBRSET.

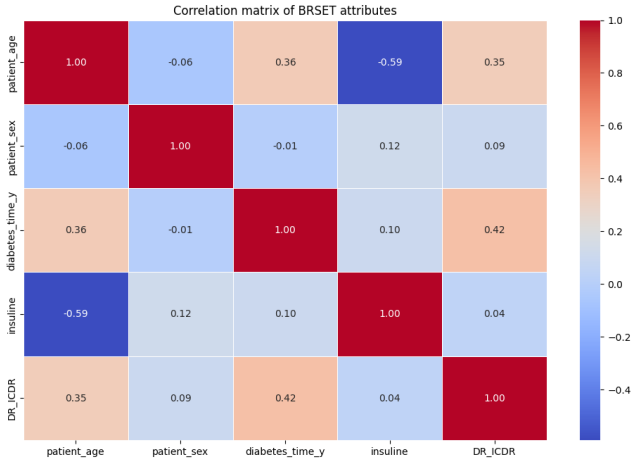


Figure 5.2: Correlation matrix between attributes of BRSET.

5.2 Classification of tabular models

To further understand the relevance of patient attributes in the prediction of DR, we trained three classification models: Decision Tree, Logistic Regression, and XGBoost. These models were applied to three

different datasets: mBRSET, BRSET and a restricted version of mBRSET, containing only the subset of attributes shared with BRSET.

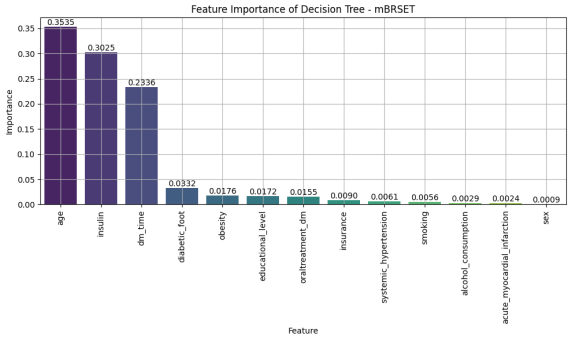
5.2.1 mBRSET

Two versions were trained: one using all features of interest in mBRSET and another, restricted mBRSET, using only features in common with BRSET (i.e., insulin use, duration of diabetes diagnosis, age and sex). The feature importance of the DT, LR and XGBoost models was determined and plotted (see Figure 5.3).

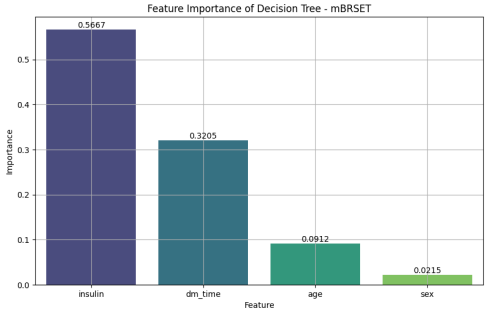
For the Decision Tree model, the three most important features were, by order, age, insulin and duration of diabetes diagnosis. Interestingly, for the restricted mBRSET, although the three most important features remained the same, their importance level changed, with insulin becoming the most important, followed by duration of diabetes diagnosis and then age.

For the Logistic Regression model, duration of diabetes diagnosis, insulin and oral treatment were considered the most important features for mBRSET, with age being considered the sixth most important, although with a similar feature importance score to oral treatment. In the restricted mBRSET, the order of importance remained the same.

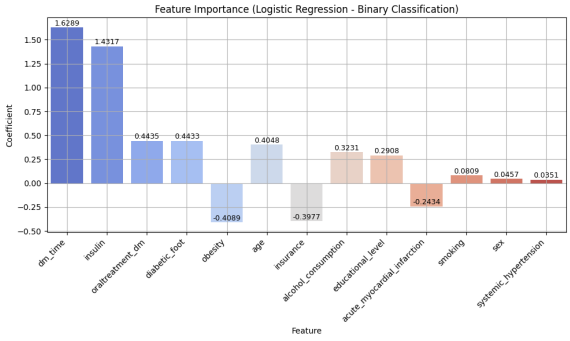
For the XGBoost model, age, duration of diabetes diagnosis, and sex were the most important attributes. In the restricted mBRSET, the order of importance remained the same. In both versions of the dataset, insulin also had some importance, although significantly less than the two most important attributes.



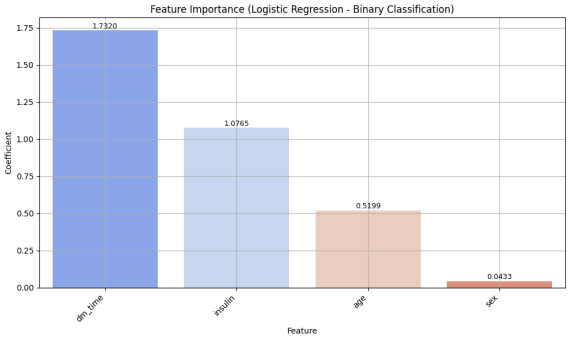
(a) Feature importance of Decision Tree model for mBRSET dataset.



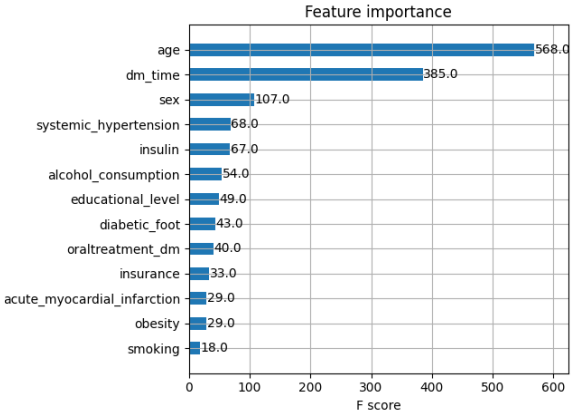
(b) Feature importance of Decision Tree model for restricted mBRSET dataset.



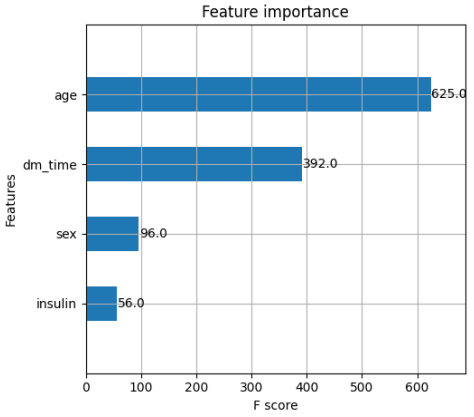
(c) Feature importance of Logistic Regression model for mBRSET dataset.



(d) Feature importance of Logistic Regression model for restricted mBRSET dataset.



(e) Feature importance of XGBoost model for mBRSET dataset.



(f) Feature importance of XGBoost model for restricted mBRSET dataset.

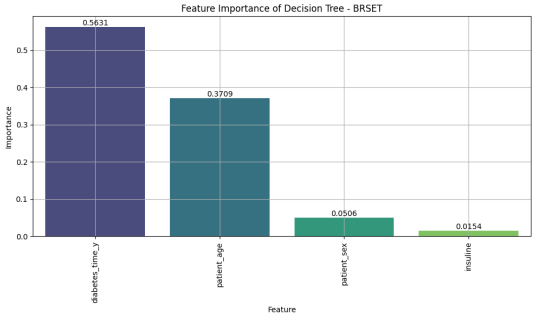
Figure 5.3: Feature importance plots for three models (DT, LR, XGBoost) comparing full and restricted mBRSET datasets.

5.2.2 BRSET

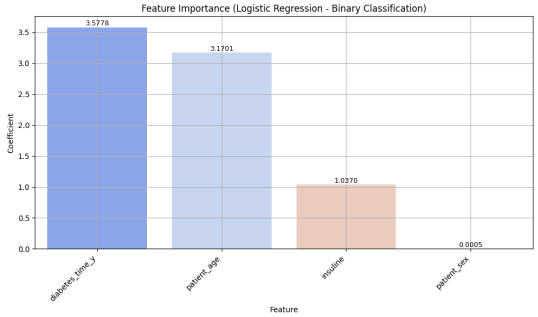
Figure 5.4 shows the feature importance of the DT, LR and XGBoost models, respectively, trained with the BRSET dataset.

For the three models, the 2 most important features were the duration of diabetes diagnosis and patient age, with the duration of diabetes diagnosis having the highest importance in both DT and LR. Insulin and patient sex had significantly less importance, with patient sex having more importance in DT and XGBoost, and insulin having more importance in LR.

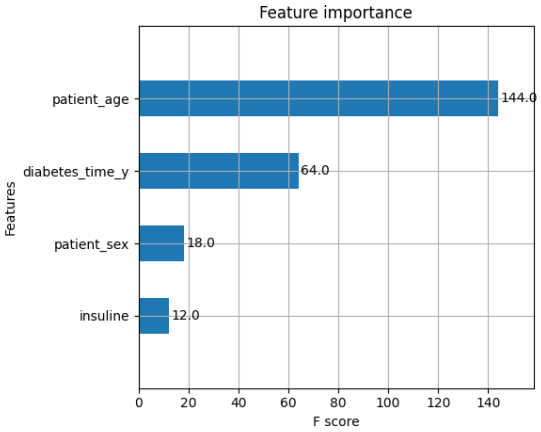
Comparing the results from restricted mBRSET with BRSET, duration of diabetes diagnosis time is an important attribute across all models. Insulin has a much higher importance for restricted mBRSET than for BRSET, while the opposite is true for age.



(a) Feature importance of Decision Tree model for BRSET dataset.



(b) Feature importance of Logistic Regression model for BRSET dataset.



(c) Feature importance of XGBoost model for BRSET dataset.

Figure 5.4: Feature importance plots for three models (DT, LR, XGBoost) for BRSET dataset.

5.2.3 Performance

As shown in Table 5.4, all models were able to reasonably predict the presence of DR for each dataset: mBRSET (72%), restricted mBRSET (75%) and BRSET (92%). XGBoost proved to be the model most prone to overfitting.

The presence of additional attributes in mBRSET does not contribute to an increase in predictive performance, as the restricted version consistently produces slightly better results, likely due to overfitting.

BRSET in particular has an extremely high discrimination performance, with all models achieving an AUROC greater than 89% in the test set. XGBoost was the best performer, with 92% AUROC.

For both datasets, sensitive attributes can be used to inform the prediction of DR. If these attributes are encoded in the image, then DL models may use this information for DR prediction. While this could be helpful in increasing model performance, it could also contribute to algorithmic bias and generalization issues.

Table 5.1: AUROC scores of the three tabular models, respectively, for mBRSET, restricted mBRSET and BRSET. The best result for each dataset is highlighted in bold.

Dataset	**Model**	**AUROC Train**	**AUROC Test**
mBRSET	DT	82.27	63.16
	LR	73.00	**72.05**
	XGBoost	87.65	70.91
Restricted mBRSET (only features in common with BRSET)	DT	71.19	69.03
	LR	70.41	**74.84**
	XGBoost	83.09	71.66
BRSET	DT	87.07	88.73
	LR	82.62	89.51
	XGBoost	88.43	**92.21**

5.3 Classification of image models

Following the analysis of feature importance and inter-attribute correlations, we evaluated the performance of models trained exclusively on fundus images. Initially, we assessed their effectiveness in predicting referable DR. Afterwards, we investigated the models' ability to decode sensitive attributes, to understand the extent to which SA information is encoded within fundus images.

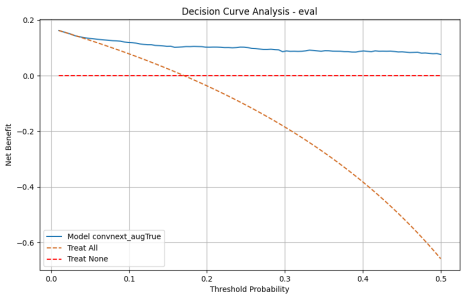
5.3.1 Binary diabetic retinopathy prediction

This section presents the performance of DL models in predicting referable DR. Three state-of-the-art architectures, ConvNeXt V2, DINOv2, and Swin V2, were trained separately on the mBRSET and

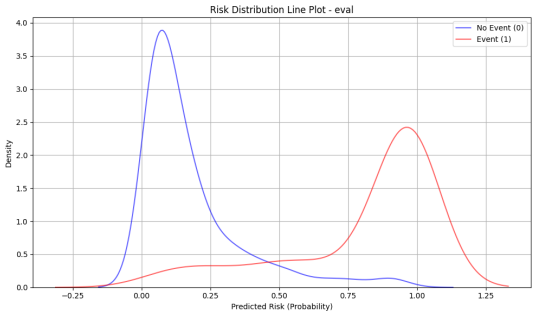
BRSET datasets. Model performance was evaluated using AUROC, DCA, and risk distribution plots (previously described in Section 4.7).

5.3.1.1 mBRSET performance

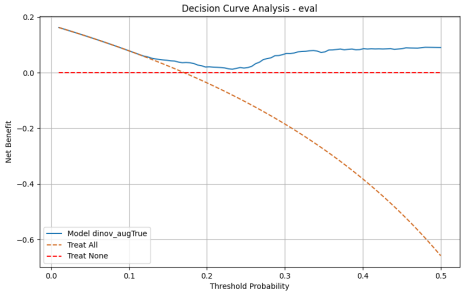
Figure 5.5 shows the DCA and risk distribution plots for the three evaluated models, while Table 5.2 shows their AUROC scores. On the mBRSET dataset, among the models trained, ConvNeXt V2 achieved the highest performance, with an AUROC of 94%. It had a low false positive rate, but a higher false negative rate. The decision curve for ConvNeXt V2 indicates strong clinical utility across a range of decision thresholds, with a slight decline at higher thresholds. Swin V2 achieved an AUROC of 91%, slightly lower than ConvNeXt V2. It had more false negatives, but fewer false positives. Clinical utility was very similar to that of ConvNeXt V2. DINOv2 had the weakest performance, with an AUROC of 88%. It yielded the lowest net benefit at low thresholds and the highest number of false negatives compared to all other models.



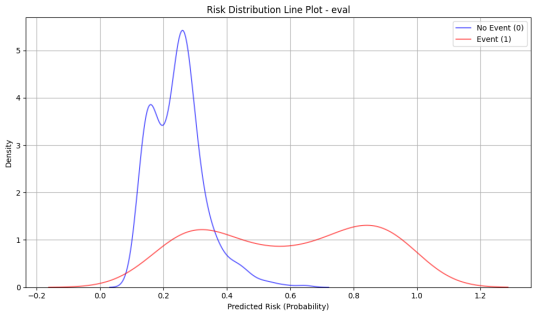
(a) ConvNeXt V2 - mBRSET.



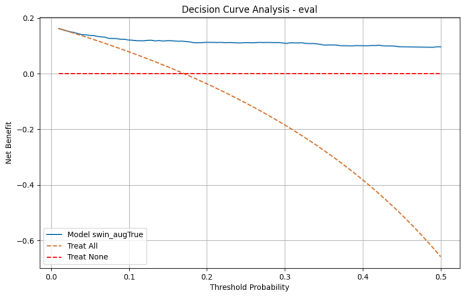
(b) ConvNeXt V2 - mBRSET.



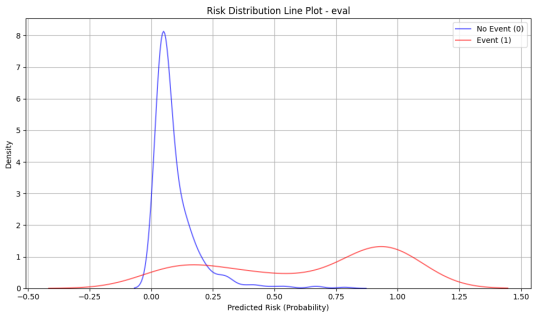
(c) DINOv2 - mBRSET.



(d) DINOv2 - mBRSET.



(e) Swin V2 - mBRSET.



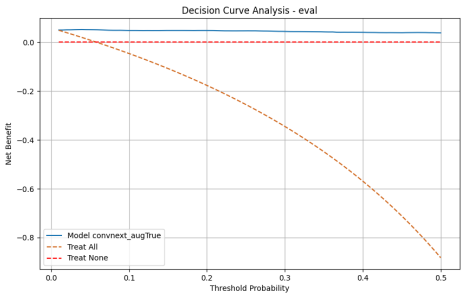
(f) Swin V2 - mBRSET.

Figure 5.5: Decision curve analysis and risk distribution plots for mBRSET by model.

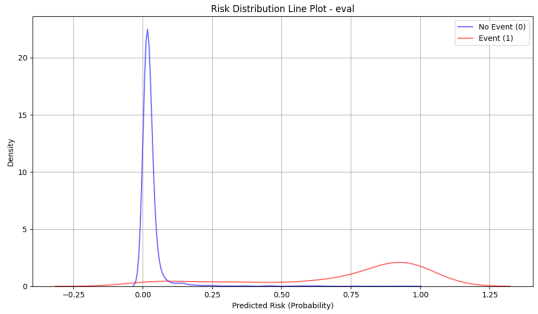
5.3.1.2 BRSET performance

As summarized in Table 5.2, ConvNeXt V2 achieved the highest predictive performance for BRSET, with an AUROC of 99%, followed closely by Swin V2 with an AUROC of 98%. Both models seemed to demonstrate clinical utility, with only a marginal drop in net benefit with increasing thresholds, reflecting a slight increase in false negatives. Similarly to mBRSET, DINOv2 was the lowest-performing model, achieving an AUROC of 90%. It exhibited the lowest net benefit and the highest false negative rate among the three models. All models had very strong capability in identifying non-referable DR

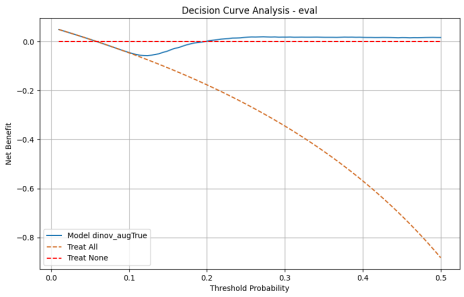
cases (non-DR). Due to this, lowering the decision threshold would be beneficial to model performance, reducing false negative predictions.



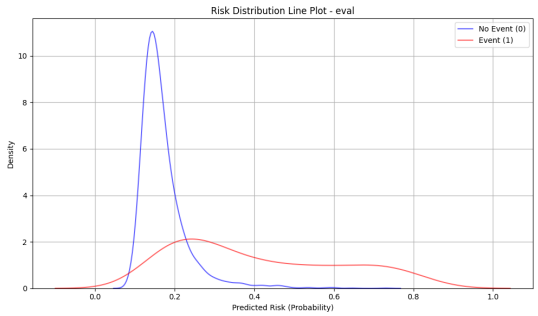
(a) ConvNeXt V2 - BRSET.



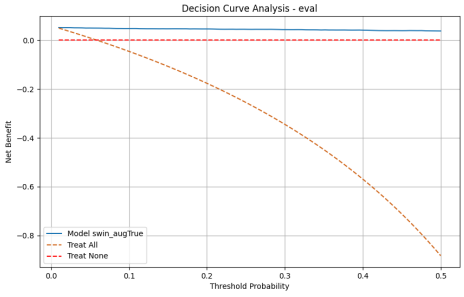
(b) ConvNeXt V2 - BRSET.



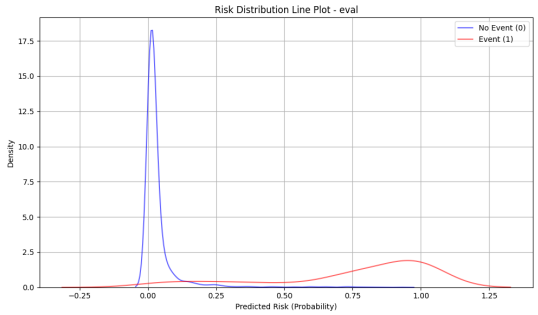
(c) DINOv2 - BRSET.



(d) DINOv2 - BRSET.



(e) Swin V2 - BRSET.



(f) Swin V2 - BRSET.

Figure 5.6: Decision Curve and Risk Distribution Analysis for BRSET by model.

5.3.1.3 Comparison

Models consistently performed better in identifying non-DR than referable DR. This is likely attributable to class imbalance, as referable DR cases only correspond to 17% and 5% of mBRSET and BRSET, respectively.

Models trained on BRSET achieved higher performance across all architectures. This is expected since BRSET is made up of images captured using conventional fundus cameras, while mBRSET images were acquired via portable fundus cameras, which have comparatively lower image quality. Among the models, ConvNeXt V2 exhibited the highest overall performance on both mBRSET and BRSET. Conversely, DINOv2 consistently underperformed across both datasets, achieving lower AUROC scores and net benefit, particularly in identifying referable DR cases.

Table 5.2: Performance of DR prediction using different image models for test sets, reported as AUROC (%).

Dataset	**Model**	**AUROC Test**
mBRSET	ConvNeXt V2	**94.33**
	DINOv2	88.31
	Swin V2	90.96
BRSET	ConvNeXt V2	**98.56**
	DINOv2	90.14
	Swin V2	98.25

5.3.2 Sensitive attribute group prediction

Next, we studied the performance of the three DL models in predicting SA information. For the mBRSET dataset, the predicted SA were age, sex, educational level, insurance status, and obesity. For the BRSET dataset, only age and sex were predicted. The results are presented in Table 5.3. Data augmentation techniques were applied during training (see Section 4.5.2).

For both datasets, all models demonstrated strong predictive performance for age. Predictive performance for sex was moderate overall and consistently lower than for age, with the exception was DINOv2, which had poor performance on mBRSET.

Specifically for the mBRSET dataset, all models struggled to accurately predict educational level, insurance status, and obesity, with prediction performance remaining low across the three models.

Table 5.3: Performance of sensitive attribute prediction using different models for test set, reported as AUROC (%) and balanced accuracy (BA) (%).

Dataset	**Model**	**Age**		**Sex**		**Educational Level**		**Insurance**		**Obesity**	
		AUROC	**BA**	**AUROC**	**BA**	**AUROC**	**BA**	**AUROC**	**BA**	**AUROC**	**BA**
mBRSET	ConvNeXt V2	90.50	69.09	76.60	68.29	56.96	53.57	46.65	50.00	63.14	50.00
	DINOv2	90.73	77.32	53.35	50.00	57.01	50.00	48.79	50.00	40.29	50.00
	Swin V2	87.85	78.64	76.79	69.89	63.02	52.64	41.74	50.00	58.50	50.00
BRSET	ConvNeXt V2	96.59	90.42	84.10	76.16	N/A					
	DINOv2	95.26	88.68	80.27	70.13	N/A					
	Swin V2	95.89	88.60	83.12	74.55	N/A					

5.3.3 Fairness analysis of DR prediction with mBRSET

Among the evaluated models, ConvNeXt V2 had the most balanced performance across SA groups, with AUROC differences typically within 1-3%, indicating strong generalizability and fairness.

Swin V2 also performed consistently, with most AUROC disparities remaining under 4%. Compared to ConvNeXt V2, Swin V2 exhibited slightly higher discrepancies, particularly between different educational levels and obesity groups.

In contrast, DINOv2 had the largest discrepancies across SA groups, raising fairness concerns. However, DINOv2's performance and fairness improved considerably with data augmentation, suggesting enhanced generalization. Notable improvements were observed across several attributes. For age, the AUROC difference reduced significantly from 26% (non-augmented) to 9% (augmented). Similarly, the disparity in educational level decreased from 16% to 7%. The difference in insurance status also narrowed slightly, from 15% and 12%.

Finally, it is also important to highlight that all models achieved a perfect AUROC (100%) for obese patients in the augmented setting. While this indicates excellent performance, it may also reflect data imbalance or overfitting, particularly given the small sample size of this group.

Table 5.4: AUROC (%) for DR prediction across SA groups in mBRSET, comparing performance of different models with and without data augmentation.

	Non-augmented			**Augmented**		
	ConvNeXt V2 (%)	**DINOv2 (%)**	**Swin V2 (%)**	**ConvNeXt V2 (%)**	**DINOv2 (%)**	**Swin V2 (%)**
Full	94.13	68.44	90.77	94.33	88.31	90.96
Age - group0	93.67	46.83	93.67	97.06	80.09	88.69
Age - group1	94.27	72.88	90.66	93.79	89.81	91.41
Sex - group0	94.55	69.80	90.31	95.22	88.95	91.10
Sex - group1	93.63	67.53	91.83	93.43	88.21	90.83
Educational level - group0	94.64	78.61	92.75	95.22	92.43	92.74
Educational level - group1	93.82	62.31	89.42	93.71	85.52	89.73
Insurance - group0	94.25	69.01	90.63	94.26	88.86	91.27
Insurance - group1	95.71	54.29	90.00	97.14	77.14	85.71
Obesity - group0	93.35	69.08	90.84	93.87	87.93	91.10
Obesity - group1	100.00	70.29	100.00	100.00	100.00	100.00

5.3.4 Fairness analysis of DR prediction with BRSET

Similarly to mBRSET, the best-performing model was ConvNeXt V2, and the worst-performing model was DINOv2. Overall, no significant AUROC discrepancies were detected in this dataset, with the

largest difference being 8% in DINOv2. Both ConvNeXt V2 and Swin V2 had minimal variations, with maximum AUROC differences of 1% and 2%, respectively.

Table 5.5: AUROC (%) for DR prediction across SA groups in BRSET, comparing performance of different models with and without data augmentation.

	Non-augmented			**Augmented**		
	ConvNeXt V2 (%)	**DINOv2 (%)**	**Swin V2 (%)**	**ConvNeXt V2 (%)**	**DINOv2 (%)**	**Swin V2 (%)**
Full	97.79	94.46	97.52	98.56	90.14	98.25
Age - group0	98.52	95.54	99.15	99.10	93.21	99.71
Age - group1	97.64	93.38	97.26	98.24	87.84	97.46
Sex - group0	97.66	93.07	97.96	98.06	85.26	97.66
Sex - group1	97.88	95.51	97.18	98.93	93.61	98.71

5.4 Disentanglement architecture

A disentanglement network was implemented to investigate the impact of separating sensitive attributes from medically relevant features. The effectiveness of this approach was assessed using the AUROC obtained by the DR classifier when relying exclusively on the medical-related features. Three representative cases from the mBRSET dataset were selected for analysis:

1. mBRSET - ConvNeXt V2 (age)
2. mBRSET - DINOv2 (sex)
3. mBRSET - Swin V2 (insurance)

All selected cases concern mBRSET, as it exhibited the most significant group disparities across sensitive attributes. Additional experimental results are presented in Appendix A.2. With additional time, it would be of interest to extend this analysis to the BRSET dataset as well.

5.4.1 ConvNeXt V2 - Disentangling medical from age information

The results of disentangling age-related information from medical features using the ConvNeXt V2 model are presented in Figure 5.7.

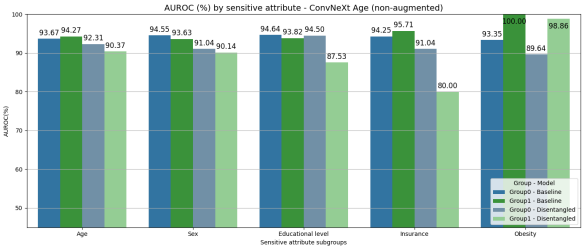
In terms of discriminatory performance, the AUROC decreased after performing disentanglement. On the non-augmented mBRSET dataset, the baseline model achieved an AUROC of 94%, while the disentangled model reached 91% on the DR-predicting task. On the augmented dataset, performance dropped from 94% (baseline) to 87% (disentangled).

Disentanglement had mixed effects on fairness across SA groups. In the non-augmented dataset, both age and obesity disparity increased by 2%, and sex disparity remained unchanged. The most

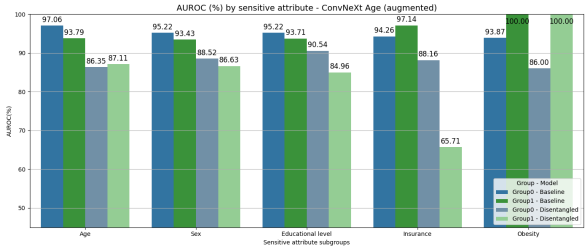
significant effect was in insurance status, which increased from 2% to 11% In the augmented dataset, age group disparity decreased from 4% to 1%, suggesting improved fairness with respect to age. However, nearly all remaining attributes observed increased disparities. The most pronounced effect observed was also in insurance status, where AUROC disparity increased from 3% to 21%, resulting in insured patients being at a greater risk of being misclassified.

Decision curve analysis revealed no meaningful difference in clinical utility between the age groups when comparing the baseline and disentangled models. Additionally, the risk distribution plots reveal that although disentanglement reduced the false positive rate, this was outweighed by a greater increase in the false negative rate, ultimately leading to a drop in overall performance.

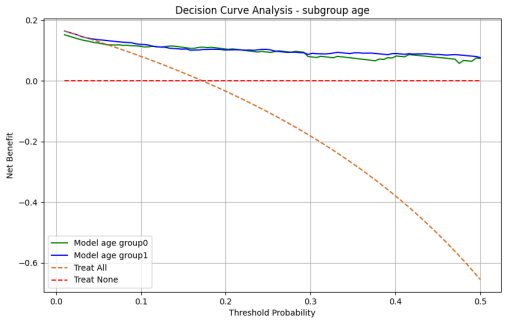
Disentangling age-related information reduced age-based disparities, at the expense of model performance and fairness across other SA. A possible explanation is that the model relies on age-encoded information for DR prediction, thereby hindering its performance when removing this information. This is further supported by the model's strong performance in the age prediction task (see Section 5.3.2). When comparing the effect of disentanglement and data augmentation for ConvNeXt V2, data augmentation yielded better overall performance than disentanglement.



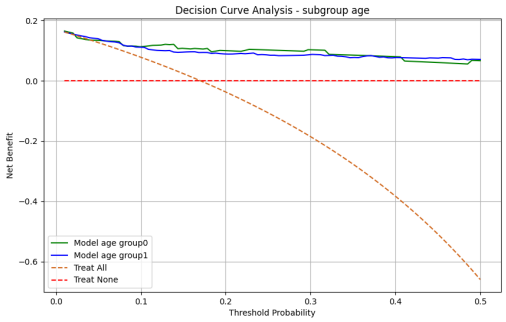
(a) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the non-augmented mBRSET dataset.



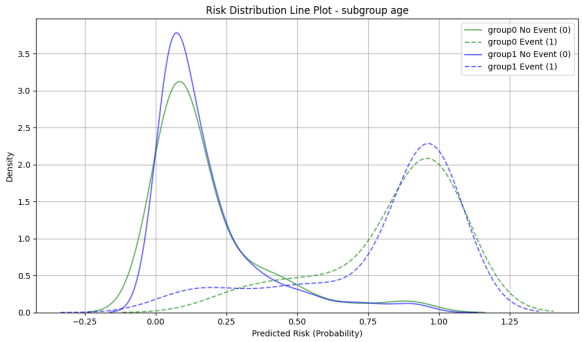
(b) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the augmented mBRSET dataset.



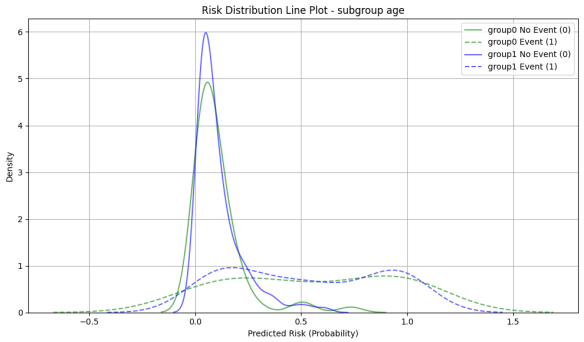
(c) Decision curve analysis of the baseline model on the augmented mBRSET dataset.



(d) Decision curve analysis of the disentangled model on the augmented mBRSET dataset.



(e) Risk distribution plot of the baseline model on the augmented mBRSET dataset.



(f) Risk distribution plot of the disentangled model on the augmented mBRSET dataset.

Figure 5.7: Comparison of baseline and disentangled model performance on the mBRSET dataset across AUROC, decision curve analysis, and risk distribution plots.

5.4.2 DINOv2 - Disentangling medical from sex information

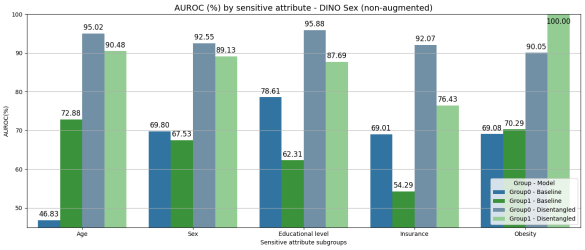
The results of disentangling sex-related information from medical features using the DINOv2 model are shown in Figure 5.7. In terms of discriminatory performance, disentanglement resulted in a substantial improvement. On the non-augmented mBRSET dataset, AUROC increased markedly from 68% (baseline) to 91% (disentangled). On the augmented dataset, performance had a modest increase, with AUROC improving from 88% to 90%.

Disentanglement contributed to mitigating unfairness across SA groups, particularly in the non-augmented dataset. All groups experienced an improvement in performance. Age and educational level reduced disparity gaps (by 22% and 8%, respectively), slight increases in disparity were observed for sex and insurance (both by 1%) and a 10% disparity increase was observed for obesity. In the augmented dataset, performance also improved across nearly all groups. Disparity gaps also decreased in age and educational level (by 8% and 1%, respectively), while sex and obesity had minimal increases (both by 1%), and insurance increased by 6%, placing insured patients at a disadvantage.

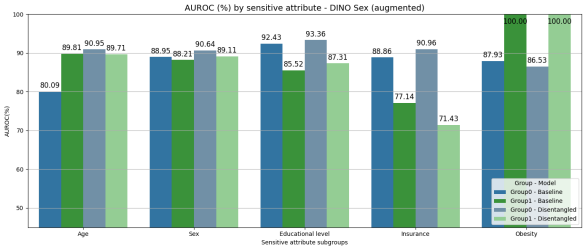
Decision curve analysis revealed improved clinical utility for the disentangled model between the age groups. Furthermore, the net benefit difference improved substantially at lower decision thresholds.

The risk distribution plots reveal that disentanglement reduced both false positive and false negative rates, contributing to an increase in overall performance. The results also suggest that lowering the decision threshold slightly could benefit the augmented model.

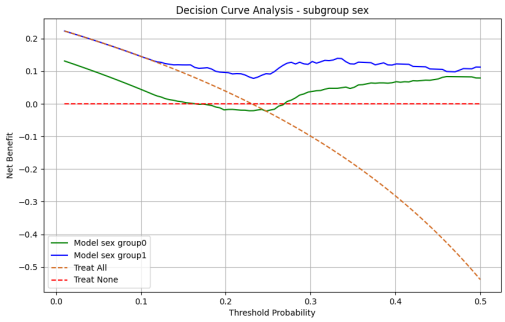
Overall, disentangling sex-related information improved model performance and partly improved model fairness, particularly on the non-augmented dataset. Comparison with the DINOv2 baseline suggests that data augmentation played a significant role in mitigating overfitting, consistent with known limitations of transformer-based models on small datasets [129]. This result is interesting given that the baseline model had a weak performance in the sex prediction task (see Section 5.3.2). Disentanglement appeared to produce a similar effect to data augmentation, with slightly greater performance gains. This supports the role of disentanglement as a bias mitigation strategy, but also as a technique that improves generalization.



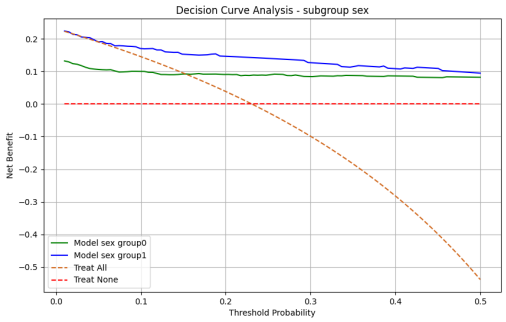
(a) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the non-augmented mBRSET dataset.



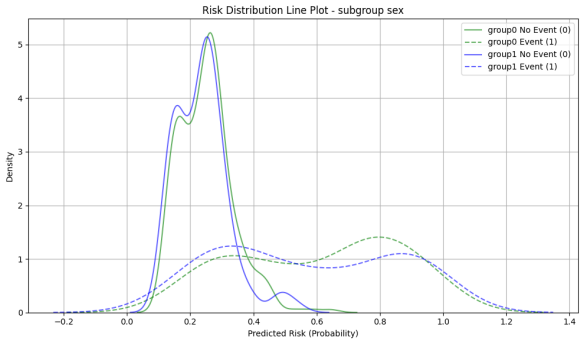
(b) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the augmented mBRSET dataset.



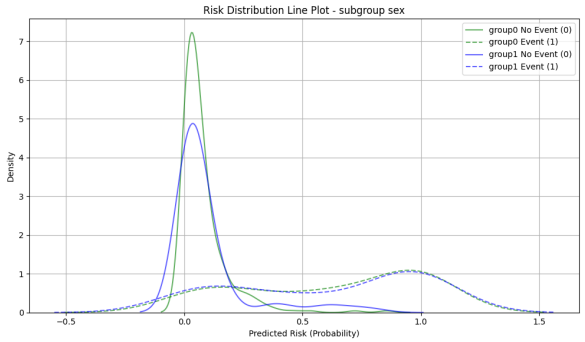
(c) Decision curve analysis of the baseline model on the augmented mBRSET dataset.



(d) Decision curve analysis of the disentangled model on the augmented mBRSET dataset.



(e) Risk distribution plot of the baseline model on the augmented mBRSET dataset.



(f) Risk distribution plot of the disentangled model on the augmented mBRSET dataset.

Figure 5.8: Comparison of baseline and disentangled model performance on the mBRSET dataset across AUROC, decision curve analysis, and risk distribution plots.

5.4.3 Swin V2 - Disentangling medical from insurance information

The results of disentangling insurance-related information from medical features using the Swin V2 model are illustrated in Figure 5.7.

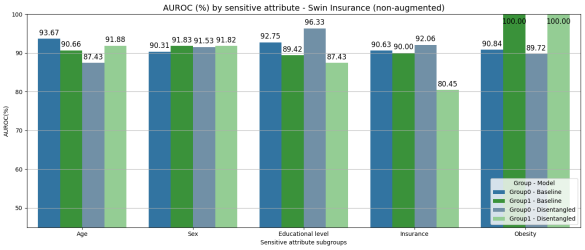
Regarding discriminatory performance, following disentanglement, a slight decrease in AUROC was observed. On the non-augmented mBRSET dataset, the baseline model achieved an AUROC of 91%,

which remained at 91% after disentanglement. On the augmented dataset, performance declined from 91% (baseline) to 88% (disentangled).

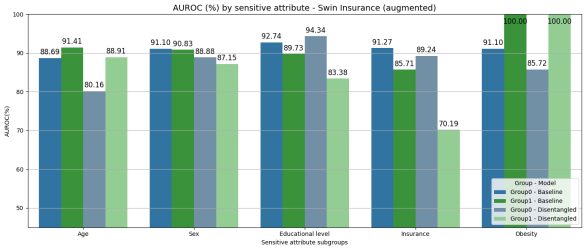
The impact of disentanglement on fairness across SA groups was mostly negative. In the non-augmented dataset, disparity in sex decreased by 1%, disparities in age and obesity increased marginally (both by 1%). For educational level and insurance, a more substantial increase of disparities (6% and 11%, respectively). For the augmented dataset, sex-based disparity remained unchanged, while disparities related to age, insurance, and obesity worsened by 5%, 13% and 6% respectively.

Decision curve analysis showed a decrease in disparity in clinical utility between insurance groups in the disentangled model when compared with the baseline models. Risk distribution plots revealed that disentanglement reduced the false positive rate of insured patients, although this was accompanied by a slight increase in the false negative rate of uninsured patients. Both groups could potentially benefit from lowering the decision threshold.

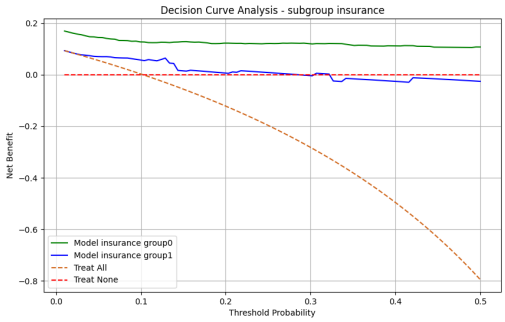
Disentangling insurance-related information did not significantly affect disparities, at the expense of model performance and fairness across other SA. Once again, a possible explanation is that the model may rely on insurance-encoded information for DR prediction, thereby hindering its performance when removing this information. However, this hypothesis is less convincing in this context, given that the baseline model performed poorly in the insurance prediction task (see Section 5.3.2). When comparing the effect of disentanglement and data augmentation for Swin V2, data augmentation yielded better overall performance than disentanglement.



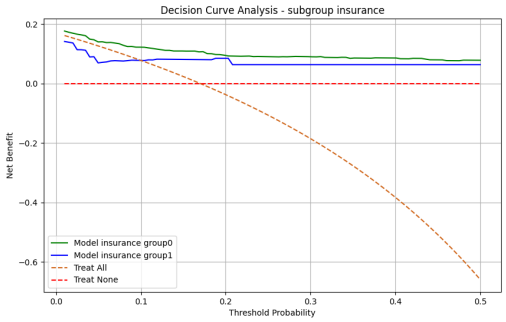
(a) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the non-augmented mBRSET dataset.



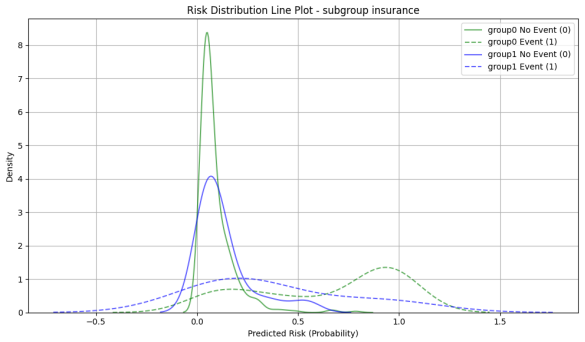
(b) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the augmented mBRSET dataset.



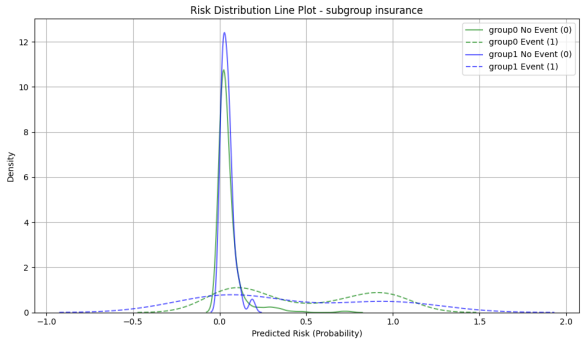
(c) Decision curve analysis of the baseline model on the augmented mBRSET dataset.



(d) Decision curve analysis of the disentangled model on the augmented mBRSET dataset.



(e) Risk distribution plot of the baseline model on the augmented mBRSET dataset.

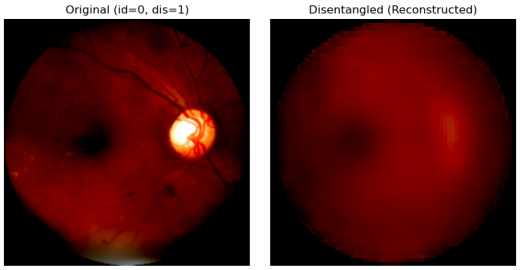


(f) Risk distribution plot of the disentangled model on the augmented mBRSET dataset.

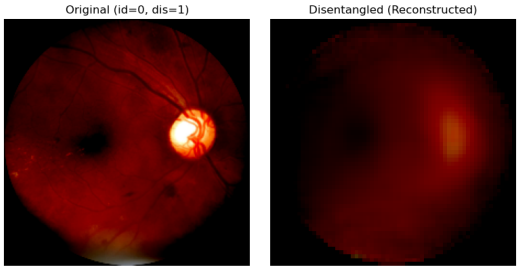
Figure 5.9: Comparison of baseline and disentangled model performance on the mBRSET dataset across AUROC, decision curve analysis, and risk distribution plots.

5.4.4 Visual results of disentanglement

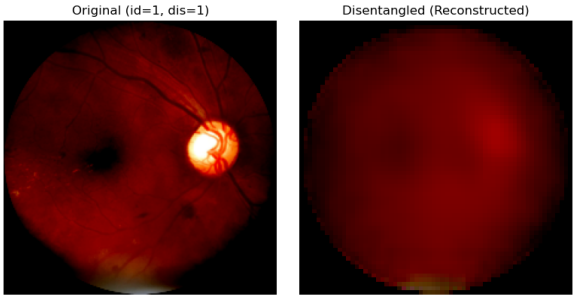
Figure 5.10 presents the original retinal images alongside their reconstructed disentangled versions for the three selected cases.



(a) mBRSET - ConvNeXt V2 (age).



(b) mBRSET - DINOv2 (sex).



(c) mBRSET - Swin V2 (insurance).

Figure 5.10: Original and reconstructed retinal images after disentanglement for three models: (a) ConvNeXt V2 (age), (b) DINOv2 (sex), and (c) Swin V2 (insurance).

The reconstructed images broadly preserve the overall retinal structure but emphasize distinct regions of the retina. These variations suggest that each model encodes information differently and that each feature may be internally represented and disentangled differently.

Additionally, the results highlight a key limitation of the generative reconstruction of disentangled retinal images. Retinal images contain fine vascular structures and subtle pathological features, which are often challenging for generative models to reproduce accurately [114]. Inaccuracies in reconstruction can negatively affect the quality of disentanglement, as the disentanglement loss depends on the quality of the reconstruction.

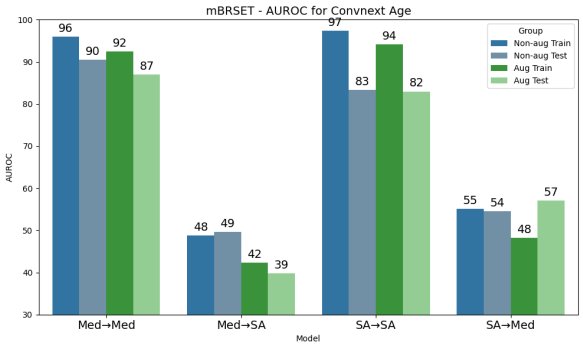
5.4.5 Validating the disentanglement architecture

This section evaluates the effectiveness of the disentanglement architecture in separating medically relevant information from sensitive attributes without compromising diagnostic performance.

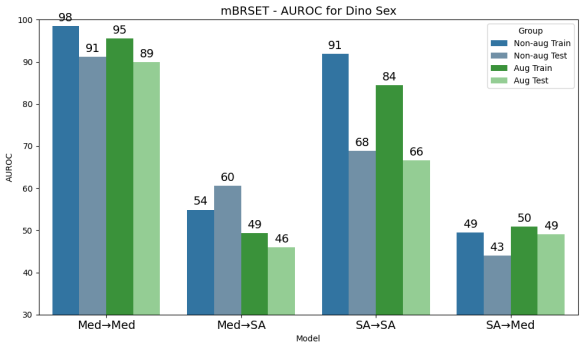
To assess this, we analyze AUROC scores across four tasks:

1. **Med→Med:** Using only medical features for the medical prediction task;
2. **Med→SA:** Using only medical features for SA prediction;
3. **SA→SA:** Using only SA features for SA prediction;
4. **SA→Med:** Using only SA features for medical prediction.

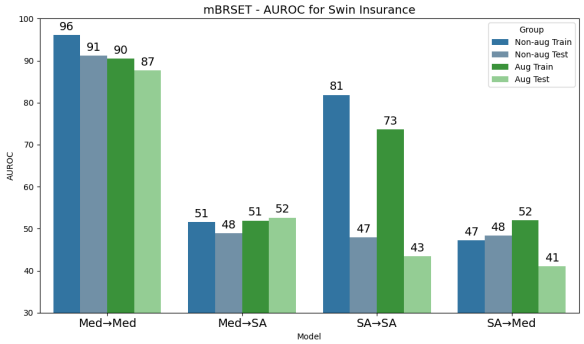
Results are presented for the three cases, including with and without data augmentation (see Figure 5.11).



(a) AUROC performance across prediction tasks – ConvNeXt V2 (Age).



(b) AUROC performance across prediction tasks – DINOv2 (Sex).



(c) AUROC performance across prediction tasks – Swin V2 (Insurance).

Figure 5.11: AUROC comparison across four prediction tasks after disentanglement: Med→Med, Med→SA, SA→SA, and SA→Med. Each subfigure shows results for a specific model–attribute pair. Models trained with and without data augmentation are compared to evaluate the impact on disentanglement performance.

All models had a strong performance on the Med→Med task after disentanglement. ConvNeXt V2 (Age) and DINOv2 (Sex) have good results on the SA→SA task, while both Med→SA and SA→Med performances remained close to random (AUROC$\approx$ 50%), indicating effective disentanglement. For Swin V2 (Insurance), the SA→SA task showed signs of overfitting, with high AUROC values on the training set and a drop to 50% on the test set (consistent with the results reported in Section 5.3.2). Despite this, the Med→SA and SA→Med performances remained equivalent to random, suggesting that overfitting did not significantly compromise disentanglement.

5.4.6 Alterations to the disentanglement architecture

We explored three variants of the proposed disentanglement architecture to assess the impact of each loss component.

1. Removal of the disentanglement loss while retaining reconstruction and classification losses;
2. Removal of both disentanglement and reconstruction losses, relying solely on the classification loss;
3. Use of only the DR classification loss with separated latent spaces for DR and SA information, but without applying SA classification, reconstruction, or disentanglement losses.

Table 5.6 presents the AUROC scores across these variants, achieved by the DR classifier when using only the medical-related features. The baseline model achieved the highest AUROC (94%). Introducing the complete disentanglement architecture results in a performance drop (88%). When the disentanglement loss was removed, performance remained similar (88%), suggesting that disentanglement loss had a positive, although small contribution. Removing both the disentanglement and reconstruction losses decreased the performance (86%), suggesting a positive contribution of the reconstruction loss.

The model trained using only the disease classification loss (with the SA prediction branch active) had the best performance after the baseline (92%), outperforming the complete disentangled architecture. Theoretically, maintaining the SA branch allowed the model to retain useful SA-related information.

The biggest drop in performance seemed to be related to the addition of the SA loss, contributing to the hypothesis that DR prediction uses SA-related features.

Trends are generally consistent across different SA groups. Notably, insured patients (Insurance Group1) have a very significant AUROC disparity. The addition of the SA loss greatly increases this disparity by 23%. Reconstruction and disentanglement losses partly mitigate this gap by 9%.

Interestingly, the relationship between sex and insurance status is not intuitive, especially since previous correlation analysis showed no significant association between these variables (see Section 5.1) and insurance status information could not be directly predicted from the image (see Section 5.3.2), raising the possibility of hidden confounding factors. However, due to the small sample size of insured patients, further investigation with larger datasets is necessary to draw definitive conclusions.

Table 5.6: AUROC scores across variants to the disentanglement architecture. The case presented is the disentanglement of medical from sex information in the ConvNeXt V2 model.

SA	**Group**	**N**	**Baseline**	**Dis-entangled**	**No dis loss**	**No dis + reconst loss**	**Only disease loss (with SA branch)**
Full	Full	-	94.33	87.76	87.55	85.49	91.78
Sex	Group0	81	97.06	88.46	91.40	86.54	90.48
	Group1	417	93.79	87.55	86.74	85.34	92.10
Age	Group0	329	95.22	89.35	90.27	87.70	94.16
	Group1	169	93.43	87.48	85.15	83.85	89.51
Educational level	Group0	222	95.22	91.37	91.55	90.24	96.24
	Group1	268	93.71	85.59	84.95	82.39	88.57
Insurance	Group0	455	94.26	89.07	89.01	86.99	92.83
	Group1	39	97.14	65.71	60.71	**56.43**	**79.17**
Obesity	Group0	450	93.87	86.53	86.35	84.52	90.20
	Group1	40	100.00	100.00	100.00	99.43	100.00

5.4.7 Hyperparameter testing

ConvNeXt V2 was selected for hyperparameter testing, as it demonstrated the best baseline performance among models trained on the mBRSET dataset. Sex was chosen as the SA to be disentangled, as its groups had the smallest AUROC disparity. The hyperparameter to be changed was the size of the latent dimension (*ld*) in the disentanglement architecture. The baseline model used a latent dimension of 256. Additional experiments were conducted with reduced latent sizes, including 128, 64, and 32.

Results showed a slight decrease in classification performance as the latent dimension decreased. However, model performance remained robust even for the smallest latent size.

Table 5.7: AUROC scores varying the latent dimension (ld). The case presented is the disentanglement of medical from sex information in the ConvNeXt V2 model.

SA	Group	N	Baseline	Id=256	Id=128	Id=64	Id=32
Full	Full	-	94.33	87.76	86.14	86.75	85.37
Age	Group0	81	97.06	88.46	84.28	86.31	82.69
	Group1	417	93.79	87.55	86.35	86.73	85.82
Sex	Group0	81	95.22	89.35	86.90	87.49	84.77
	Group1	417	93.43	87.48	86.21	86.61	87.53
Educational Level	Group0	222	95.22	91.37	89.75	90.27	89.93
	Group1	268	93.71	85.59	83.74	84.51	82.69
Insurance	Group0	455	94.26	89.07	87.59	88.04	86.30
	Group1	40	97.14	65.71	57.86	59.29	65.00
Obesity	Group0	450	93.87	86.53	84.96	85.69	84.26
	Group1	40	100.00	100.00	100.00	100.00	99.43

5.4.8 Summary

First, we used correlation matrices to detect possible correlations between attributes. Results revealed that age was correlated with ICDR score for BRSET.

Second, we analyzed the predictive power of patient attributes in the prediction of DR. Models trained to predict DR on patient data exhibited a substantially lower performance for mBRSET (72% AUROC) and a slight decline for BRSET (92%). Overall, duration of diabetes diagnosis, age, and insulin were considered the most important attributes across both mBRSET and BRSET.

Third, using the mBRSET and BRSET datasets, ConvNeXt V2, DINOv2, and Swin V2 models were trained on macula images to predict DR. DL models trained on BRSET achieved high diagnostic performance (AUROC of 99%), while performance declined when applied to mBRSET (AUROC of 94%). This decline may be attributed to images being captured using portable handheld cameras, as well as a smaller dataset size.

DL models were also evaluated for their ability to predict SA. Age and sex were predicted with reasonable accuracy (AUROC values of 91% and 77%, respectively), while other attributes in mBRSET, such as educational level, insurance status, and obesity, proved more challenging. These results suggest that retinal images can encode SA-related information, introducing fairness concerns. However, the absence of direct predictability does not guarantee an absence of shared latent factors (latent confounders). Shared underlying features between SA and DR-related patterns may still impact model behavior.

Next, we introduced a fairness assessment framework based on group fairness, comparing AUROC, decision curve analysis, and risk distribution across SA subgroups. Our fairness analysis revealed potential disparities in mBRSET. DINOv2 exhibited the most pronounced disparities, including a 26% AUROC gap between groups aged $\leq$50 and >50, which was reduced to 9% through data augmentation.

ConvNeXt V2 and Swin V2 both showed small variations in AUROC scores, remaining under 4% in most cases. BRSET had few disparities, with the most significant disparities also being for DINOv2.

Afterwards, a disentanglement architecture was developed to separate DR-related from SA-specific information. Three cases were analyzed in greater depth, one for each model. While disentanglement was not successful in improving fairness and performance in ConvNeXt V2 and Swin V2, it is worth noting the impact of disentanglement on DINOv2. Disentanglement led to significant improvements in DINOv2, surpassing even those achieved by data augmentation. Interestingly, we observed that disentanglement occurred successfully, even when the model struggled to predict the SA directly. This was evident in the disentangling of sex-related information in the DINOv2 model and insurance-related information in the Swin V2 model. In these cases, the model appeared to disentangle irrelevant features, focusing on DR-relevant features, even if this did not align perfectly with the targeted SA.

Several factors inherent to fundus imaging could have contributed to the difficulty of effective disentanglement. The fine-grained nature of retinal features makes disentanglement challenging, as it requires removing SA-related information while preserving fine details, which are necessary for accurate clinical interpretation. This trade-off between removing bias and maintaining diagnostic detail complicates the learning process and can hinder model performance. Furthermore, DR prediction may inadvertently rely on SA-related information, leading to decreased performance when this information is removed. In cases where confounding factors are present, latent dependencies between SA and DR-relevant features may persist even after standard disentanglement is applied [130], limiting the effectiveness of such techniques.

Given these results, we believe it is worthwhile to further refine disentanglement-based bias mitigation approaches in order to better suit the domain of fundus imaging.

Chapter 6

Conclusion and Future Work

It is of high importance to increase fairness concerns, particularly for AI healthcare, to prevent unequal treatment outcomes. Fairness studies ought to be conducted before clinical implementation. This dissertation addresses this critical issue by investigating fairness in diabetic retinopathy (DR) detection, as well as disentanglement as a strategy for mitigating unfairness.

The main goals of this work were successfully achieved. First, three image-based models were trained and evaluated regarding fairness using the mBRSET and BRSET datasets. The mBRSET dataset, collected using handheld cameras, demonstrated strong classification performance, although slightly inferior to BRSET, collected using table-top cameras. Moreover, it exhibited greater disparities between subgroups compared to the BRSET dataset. In particular, the ConvNeXt V2 model consistently achieved the highest predictive performance while maintaining the lowest disparity gaps across both datasets. Second, a disentanglement network was developed to mitigate the effects of confounding information, and a series of validation studies were conducted to confirm its effectiveness. Third, we investigated the impact of feature disentanglement in mitigating this bias. Results were only effective for the DINOv2 model, due to the complexity of performing disentanglement in fundus images, so further adaptations to the architecture are recommended.

6.1 Future Work

Overall, while disentanglement remains a promising direction, its effectiveness in fundus imaging remains limited by domain-specific challenges. Bias mitigation in this context is non-trivial, and more robust techniques may be necessary to improve fairness without compromising predictive performance.

In future work, several directions could be pursued. One possibility is to synthetically introduce bias into the datasets to study its impact on fairness and shortcut learning behavior, integrating tools like ShorT (Shortcut Testing), and assess whether disentanglement can mitigate such bias. Additionally, the proposed disentanglement framework could be applied to other medical imaging domains where disentanglement may be less challenging (e.g., standard X-rays or MRI scans). Exploring alternative

disentanglement architectures that support high-resolution image generation would also be of interest, for instance, the approach proposed by Du et al. [106]. Finally, it would be interesting to implement disentanglement with other state-of-the-art architectures for DR prediction, such as those proposed by Vasireddi et al. [64] and Almas et al. [65].

Bibliography

[1] Wei Wang and Amy C. Y. Lo. "Diabetic Retinopathy: Pathophysiology and Treatments". In: *International Journal of Molecular Sciences* 19.6 (June 2018), p. 1816. ISSN: 1422-0067. DOI: 10.3390/ijms19061816. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6032159/.

[2] *Diabetic Retinopathy | National Eye Institute*. URL: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy.

[3] Yoshihiro Yonekawa et al. "American Society of Retina Specialists Clinical Practice Guidelines: Management of Nonproliferative and Proliferative Diabetic Retinopathy Without Diabetic Macular Edema". In: *Journal of Vitreoretinal Diseases* 4.2 (Jan. 2020), pp. 125–135. ISSN: 2474-1264. DOI: 10.1177/2474126419893829. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8297841/.

[4] Bassirou A. M. Mounirou et al. "Diabetic Retinopathy: An Overview of Treatments". en. In: *Indian Journal of Endocrinology and Metabolism* 26.2 (Mar. 2022), pp. 111–118. ISSN: 2230-9500. DOI: 10.4103/ijem.ijem_480_21. URL: https://journals.lww.com/10.4103/ijem.ijem_480_21.

[5] Zhen Ling Teo et al. "Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis". In: *Ophthalmology* 128.11 (Nov. 2021), pp. 1580–1591. ISSN: 0161-6420. DOI: 10.1016/j.ophtha.2021.04.027. URL: https://www.sciencedirect.com/science/article/pii/S0161642021003213.

[6] Zhihui Wang et al. "Risk factors for diabetic retinopathy in young and middle-aged patients: a retrospective study". en. In: *BMC Ophthalmology* 24.1 (Dec. 2024), p. 544. ISSN: 1471-2415. DOI: 10.1186/s12886-024-03821-y. URL: https://bmcophthalmol.biomedcentral.com/articles/10.1186/s12886-024-03821-y.

[7] Jennifer Manne-Goehler et al. "Health system performance for people with diabetes in 28 low- and middle-income countries: A cross-sectional study of nationally representative surveys". In: *PLoS Medicine* 16.3 (Mar. 2019), e1002751. ISSN: 1549-1277. DOI: 10.1371/journal.pmed.1002751. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6396901/.

[8] Serge Resnikoff et al. “Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): will we meet the needs?” en. In: *British Journal of Ophthalmology* 104.4 (Apr. 2020), pp. 588–592. ISSN: 0007-1161, 1468-2079. DOI: 10.1136/bjophthalmol-2019-314336. URL: https://bjo.bmj.com/lookup/doi/10.1136/bjophthalmol-2019-314336.

[9] Konstantinos Mavrogiorgos et al. “Bias in Machine Learning: A Literature Review”. en. In: *Applied Sciences* 14.19 (Oct. 2024), p. 8860. ISSN: 2076-3417. DOI: 10.3390/app14198860. URL: https://www.mdpi.com/2076-3417/14/19/8860.

[10] Long Bai et al. “The Influence of Age and Gender Information on the Diagnosis of Diabetic Retinopathy: Based on Neural Networks”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. Nov. 2021, pp. 3514–3517. DOI: 10.1109/EMBC46164.2021.9629607. URL: https://ieeexplore.ieee.org/document/9629607.

[11] Caleb Robinson et al. “Deep learning models for COVID-19 chest x-ray classification: Preventing shortcut learning using feature disentanglement”. en. In: (Feb. 2021). DOI: 10.1101/2021.02.11.20196766. URL: http://medrxiv.org/lookup/doi/10.1101/2021.02.11.20196766.

[12] European Commission. *Sustainable Development Goals*. English. URL: https://commission.europa.eu/strategy-and-policy/sustainable-development-goals_en.

[13] Michael D. Abramoff, Mona K. Garvin, and Milan Sonka. “Retinal Imaging and Image Analysis”. In: *IEEE Reviews in Biomedical Engineering* 3 (2010), pp. 169–208. ISSN: 1937-3333, 1941-1189. DOI: 10.1109/RBME.2010.2084567. URL: http://ieeexplore.ieee.org/document/5660089/.

[14] Maurice Ptito, Maxime Bleau, and Joseph Bouskila. “The Retina: A Window into the Brain”. In: *Cells* 10.12 (Nov. 2021), p. 3269. ISSN: 2073-4409. DOI: 10.3390/cells10123269. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8699497/.

[15] Barry W. Fitzgerald. “Using Hawkeye from the Avengers to communicate on the eye”. en. In: *Advances in Physiology Education* 42.1 (Mar. 2018), pp. 90–98. ISSN: 1043-4046, 1522-1229. DOI: 10.1152/advan.00161.2017. URL: https://www.physiology.org/doi/10.1152/advan.00161.2017.

[16] Silke Aumann et al. “Optical Coherence Tomography (OCT): Principle and Technical Realization”. eng. In: *High Resolution Imaging in Microscopy and Ophthalmology: New Frontiers in Biomedical Optics*. Ed. by Josef F. Bille. Cham (CH): Springer, 2019. ISBN: 9783030166373. URL: http://www.ncbi.nlm.nih.gov/books/NBK554044/.

[17] David A. Salz and Andre J. Witkin. “Imaging in Diabetic Retinopathy”. In: *Middle East African Journal of Ophthalmology* 22.2 (2015), pp. 145–150. ISSN: 0974-9233. DOI: 10.4103/0974-9233.151887. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4411609/.

[18] James Kang Hao Goh et al. “Retinal Imaging Techniques for Diabetic Retinopathy Screening”. In: *Journal of Diabetes Science and Technology* 10.2 (Feb. 2016), pp. 282–294. ISSN: 1932-2968. DOI: 10.1177/1932296816629491. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4773981/.

[19] Surabhi Ruia and Koushik Tripathy. “Fluorescein Angiography”. eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2025. URL: http://www.ncbi.nlm.nih.gov/books/NBK576378/.

[20] Yasir J. Sepah et al. “Fundus autofluorescence imaging: Fundamentals and clinical relevance”. In: *Saudi Journal of Ophthalmology* 28.2 (Apr. 2014), pp. 111–116. ISSN: 1319-4534. DOI: 10.1016/j.sjopt.2014.03.008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4023118/.

[21] Vinod Kumar et al. “Ultra-wide field retinal imaging: A wider clinical perspective”. In: *Indian Journal of Ophthalmology* 69.4 (Apr. 2021), pp. 824–835. ISSN: 0301-4738. DOI: 10.4103/ijo.IJO_1403_20. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8012972/.

[22] Eric J. Kim et al. “Treatment of Diabetic Macular Edema”. en. In: *Current Diabetes Reports* 19.9 (July 2019), p. 68. ISSN: 1539-0829. DOI: 10.1007/s11892-019-1188-4. URL: https://doi.org/10.1007/s11892-019-1188-4.

[23] Thomas C. Michels and Oana Ivan. “Glaucoma: Diagnosis and Management”. en-US. In: *American Family Physician* 107.3 (Mar. 2023), pp. 253–262. ISSN: 1532-0650, 0002-838X. URL: https://www.aafp.org/pubs/afp/issues/2023/0300/glaucoma.html.

[24] Hasan Kiziltoprak et al. “Cataract in diabetes mellitus”. In: *World Journal of Diabetes* 10.3 (Mar. 2019), pp. 140–153. ISSN: 1948-9358. DOI: 10.4239/wjd.v10.i3.140. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6422859/.

[25] Chitaranjan Mishra and Koushik Tripathy. “Retinal Traction Detachment”. eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2025. URL: http://www.ncbi.nlm.nih.gov/books/NBK558952/.

[26] URL: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/age-related-macular-degeneration.

[27] Daisuke Yamazaki, Hirofumi Hitomi, and Akira Nishiyama. “Hypertension with diabetes mellitus complications”. en. In: *Hypertension Research* 41.3 (Mar. 2018), pp. 147–156. ISSN: 1348-4214. DOI: 10.1038/s41440-017-0008-y. URL: https://www.nature.com/articles/s41440-017-0008-y.

[28] Frederike C. Oertel et al. “Novel uses of retinal imaging with optical coherence tomography in multiple sclerosis”. en. In: *Expert Review of Neurotherapeutics* 19.1 (Jan. 2019), pp. 31–43. ISSN: 1473-7175, 1744-8360. DOI: 10.1080/14737175.2019.1559051. URL: https://www.tandfonline.com/doi/full/10.1080/14737175.2019.1559051.

[29] Dong-Rong Yang et al. “Endothelial dysfunction in vascular complications of diabetes: a comprehensive review of mechanisms and implications”. English. In: 15 (Apr. 2024). ISSN: 1664-2392. DOI: 10.3389/fendo.2024.1359255. URL: https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2024.1359255/full.

[30] Jimena Fernández-Carneado et al. “Quantification of microvascular lesions in the central retinal field: could it predict the severity of diabetic retinopathy?” en. In: (Mar. 2023). DOI: 10.1101/2023.03.21.23286574. URL: https://www.medrxiv.org/content/10.1101/2023.03.21.23286574v1.

[31] Elizabeth Pearce et al. “Intraretinal Microvascular Abnormalities and Venous Beading Have Different Genetic Profiles in Caucasian Patients with Non-Proliferative Diabetic Retinopathy”. In: *Vision* 7.1 (Mar. 2023), p. 18. ISSN: 2411-5150. DOI: 10.3390/vision7010018. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10051057/.

[32] T. M. Curtis, T. A. Gardiner, and A. W. Stitt. “Microvascular lesions of diabetic retinopathy: clues towards understanding pathogenesis?” en. In: *Eye* 23.7 (July 2009), pp. 1496–1508. ISSN: 1476-5454. DOI: 10.1038/eye.2009.108. URL: https://www.nature.com/articles/eye2009108.

[33] Sayon Roy, Shruti Amin, and Sumon Roy. “Retinal Fibrosis in Diabetic Retinopathy”. In: *Experimental eye research* 142 (Jan. 2016), pp. 71–75. ISSN: 0014-4835. DOI: 10.1016/j.exer.2015.04.004. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4683353/.

[34] In: *Ophthalmology* 98.5, Supplement (May 1991), pp. 741–756. ISSN: 0161-6420. DOI: 10.1016/S0161-6420(13)38009-9. URL: https://www.sciencedirect.com/science/article/pii/S0161642013380099.

[35] Zhengwei Yang et al. “Classification of diabetic retinopathy: Past, present and future”. In: *Frontiers in Endocrinology* 13 (Dec. 2022), p. 1079217. ISSN: 1664-2392. DOI: 10.3389/fendo.2022.1079217. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9800497/.

[36] C. P Wilkinson et al. “Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales”. In: *Ophthalmology* 110.9 (Sept. 2003), pp. 1677–1682. ISSN: 0161-6420. DOI: 10.1016/S0161-6420(03)00475-5. URL: https://www.sciencedirect.com/science/article/pii/S0161642003004755.

[37] en. May 2012. URL: https://www.aao.org/education/clinical-statement/international-clinical-classification-system-diabe.

[38] Sonia Zachariah, William Wykes, and David Yorston. "Grading diabetic retinopathy (DR) using the Scottish grading protocol". In: *Community Eye Health* 28.92 (2015), pp. 72–73. ISSN: 0953-6833. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4944099/.

[39] *How To Manage and Live With Diabetes-Related Retinopathy*. en. URL: https://my.clevelandclinic.org/health/diseases/8591-diabetic-retinopathy.

[40] Tyler A. Bahr and Sophie J. Bakri. "Update on the Management of Diabetic Retinopathy: Anti-VEGF Agents for the Prevention of Complications and Progression of Nonproliferative and Proliferative Retinopathy". In: *Life* 13.5 (Apr. 2023), p. 1098. ISSN: 2075-1729. DOI: 10.3390/life13051098. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10222751/.

[41] Spencer M. Moore and Daniel L. Chao. "Application of subthreshold laser therapy in retinal diseases: a review". In: *Expert review of ophthalmology* 13.6 (2018), pp. 311–320. ISSN: 1746-9899. DOI: 10.1080/17469899.2018.1555035. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6831099/.

[42] San-Ni Chen et al. "Refining vitrectomy for proliferative diabetic retinopathy". In: *Graefe's Archive for Clinical and Experimental Ophthalmology* 261.12 (2023), pp. 3659–3670. ISSN: 0721-832X. DOI: 10.1007/s00417-023-06134-w. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10667443/.

[43] Jorge Chawan-Saad et al. "Corticosteroids for Diabetic Macular Edema". In: *Taiwan Journal of Ophthalmology* 9.4 (Dec. 2019), pp. 233–242. ISSN: 2211-5056. DOI: 10.4103/tjo.tjo_68_19. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6947754/.

[44] Pouya Saeedi et al. "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition". en. In: *Diabetes Research and Clinical Practice* 157 (Nov. 2019), p. 107843. ISSN: 01688227. DOI: 10.1016/j.diabres.2019.107843. URL: https://linkinghub.elsevier.com/retrieve/pii/S0168822719312306.

[45] Kustaa Hletala et al. "Age at Onset and the Risk of Proliferative Retinopathy in Type 1 Diabetes". en. In: *Diabetes Care* 33.6 (June 2010), pp. 1315–1319. ISSN: 0149-5992, 1935-5548. DOI: 10.2337/dc09-2278. URL: https://diabetesjournals.org/care/article/33/6/1315/27360/Age-at-Onset-and-the-Risk-of-Proliferative.

[46] Martina Kropp et al. "Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications-risks and mitigation". eng. In: *The EPMA journal* 14.1 (Mar. 2023), pp. 21–42. ISSN: 1878-5077. DOI: 10.1007/s13167-023-00314-8.

[47] Jaimie D Steinmetz et al. “Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study”. en. In: *The Lancet Global Health* 9.2 (Feb. 2021), e144–e160. ISSN: 2214109X. DOI: 10.1016/S2214-109X(20)30489-7. URL: https://linkinghub.elsevier.com/retrieve/pii/S2214109X20304897.

[48] Taraprasad Das et al. “Recently updated global diabetic retinopathy screening guidelines: commonalities, differences, and future possibilities”. In: *Eye* 35.10 (Oct. 2021), pp. 2685–2698. ISSN: 0950-222X. DOI: 10.1038/s41433-021-01572-4. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8452707/.

[49] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. en. In: *Nature* 521.7553 (May 2015), pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539. URL: https://www.nature.com/articles/nature14539.

[50] Elysse Tom et al. “Protecting Data Privacy in the Age of AI-Enabled Ophthalmology”. In: *Translational Vision Science Technology* 9.2 (July 2020), p. 36. ISSN: 2164-2591. DOI: 10.1167/tvst.9.2.36. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7424948/.

[51] B. Foot and C. MacEwen. “Surveillance of sight loss due to delay in ophthalmic treatment or review: frequency, cause and outcome”. eng. In: *Eye (London, England)* 31.5 (May 2017), pp. 771–775. ISSN: 1476-5454. DOI: 10.1038/eye.2017.1.

[52] Daniel Avidor et al. “Cost-effectiveness of diabetic retinopathy screening programs using telemedicine: a systematic review”. en. In: *Cost Effectiveness and Resource Allocation* 18.1 (Dec. 2020), p. 16. ISSN: 1478-7547. DOI: 10.1186/s12962-020-00211-1. URL: https://resource-allocation.biomedcentral.com/articles/10.1186/s12962-020-00211-1.

[53] Georgios Papadakis et al. “Deep learning opens new horizons in personalized medicine (Review)”. In: *Biomedical Reports* (Mar. 2019). ISSN: 2049-9434, 2049-9442. DOI: 10.3892/br.2019.1199. URL: http://www.spandidos-publications.com/10.3892/br.2019.1199.

[54] Zhen Ling Teo et al. “Do we have enough ophthalmologists to manage vision-threatening diabetic retinopathy? A global perspective”. en. In: *Eye* 34.7 (July 2020), pp. 1255–1261. ISSN: 1476-5454. DOI: 10.1038/s41433-020-0776-5. URL: https://www.nature.com/articles/s41433-020-0776-5.

[55] en. In: *JMA Journal* 8.1 (2025), pp. 66–75. ISSN: 2433-3298, 2433-328X. DOI: 10.31662/jmaj.2024-0139. URL: https://www.jmaj.jp/detail.php?id=10.31662/jmaj.2024-0139.

[56] Peter Heydon et al. “Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients”. en. In: *British Journal of Ophthalmology* 105.5 (May 2021), pp. 723–728. ISSN: 0007-1161, 1468-2079. DOI: 10.1136/bjophthalmol-2020-316594. URL: https://bjo.bmj.com/lookup/doi/10.1136/bjophthalmol-2020-316594.

[57] Ryan R Ramoutar. “An Economic Analysis for the Use of Artificial Intelligence in Screening for Diabetic Retinopathy in Trinidad and Tobago”. In: *Cureus* 16.3 (), e55745. ISSN: 2168-8184. DOI: 10.7759/cureus.55745. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10999161/.

[58] Rachel Marjorie Wei Wen Tseng et al. “Considerations for Artificial Intelligence Real-World Implementation in Ophthalmology: Providers’ and Patients’ Perspectives”. en. In: *Asia-Pacific Journal of Ophthalmology* 10.3 (May 2021), pp. 299–306. ISSN: 21620989. DOI: 10.1097/APO.0000000000000400. URL: https://linkinghub.elsevier.com/retrieve/pii/S2162098923001032.

[59] Fernando Korn Malerbi et al. “Diabetic Retinopathy Screening Using Artificial Intelligence and Handheld Smartphone-Based Retinal Camera”. en. In: *Journal of Diabetes Science and Technology* 16.3 (May 2022), pp. 716–723. ISSN: 1932-2968, 1932-2968. DOI: 10.1177/1932296820985567. URL: https://journals.sagepub.com/doi/10.1177/1932296820985567.

[60] Molla Imaduddin Ahmed et al. “A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare”. In: *Cureus* 15.10 (), e46454. ISSN: 2168-8184. DOI: 10.7759/cureus.46454. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10623210/.

[61] M. Usman Akram et al. “RIDB: A Dataset of fundus images for retina based person identification”. In: *Data in Brief* 33 (Dec. 2020), p. 106433. ISSN: 2352-3409. DOI: 10.1016/j.dib.2020.106433. URL: https://www.sciencedirect.com/science/article/pii/S2352340920313159.

[62] Giona Kleinberg et al. “Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare”. en. In: *Journal of Biomed Research* Volume 3.Issue 1 (Oct. 2022), pp. 42–47. ISSN: 2693-5910. DOI: 10.46439/biomedres.3.025. URL: https://www.probiologists.com/article/racial-underrepresentation-in-dermatological-datasets-leads-to-biased-machine-learning-models-and-inequitable-healthcare.

[63] Tyler Hyungtaek Rim et al. “Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms”. en. In: *The Lancet Digital Health* 2.10 (Oct. 2020), e526–e536. ISSN: 25897500. DOI: 10.1016/S2589-7500(20)30216-8. URL: https://linkinghub.elsevier.com/retrieve/pii/S2589750020302168.

[64] Hemanth Kumar Vasireddi, K. Suganya Devi, and G. N. V. Raja Reddy. "DR-XAI: Explainable Deep Learning Model for Accurate Diabetic Retinopathy Severity Assessment". en. In: *Arabian Journal for Science and Engineering* 49.9 (Sept. 2024), pp. 12899–12917. ISSN: 2191-4281. DOI: 10.1007/s13369-024-08836-7. URL: https://doi.org/10.1007/s13369-024-08836-7.

[65] Shagufta Almas et al. "Visual impairment prevention by early detection of diabetic retinopathy based on stacked auto-encoder". en. In: *Scientific Reports* 15.1 (Jan. 2025), p. 2554. ISSN: 2045-2322. DOI: 10.1038/s41598-025-85752-2. URL: https://www.nature.com/articles/s41598-025-85752-2.

[66] Varun Gulshan et al. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs". en. In: *JAMA* 316.22 (Dec. 2016), p. 2402. ISSN: 0098-7484. DOI: 10.1001/jama.2016.17216. URL: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2016.17216.

[67] Etienne Decencière et al. "FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE: THE MESSIDOR DATABASE". In: *Image Analysis Stereology* 33.3 (Aug. 2014), p. 231. ISSN: 1854-5165, 1580-3139. DOI: 10.5566/ias.1155. URL: https://www.ias-iss.org/ojs/IAS/article/view/1155.

[68] Daniel Shu Wei Ting et al. "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes". en. In: *JAMA* 318.22 (Dec. 2017), p. 2211. ISSN: 0098-7484. DOI: 10.1001/jama.2017.18152. URL: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.18152.

[69] Hongyang Jiang et al. "An Interpretable Ensemble Deep Learning Model for Diabetic Retinopathy Disease Classification". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Berlin, Germany: IEEE, July 2019, pp. 2045–2048. ISBN: 9781538613115. DOI: 10.1109/EMBC.2019.8857160. URL: https://ieeexplore.ieee.org/document/8857160/.

[70] Ashish Bora et al. "Predicting the risk of developing diabetic retinopathy using deep learning". en. In: *The Lancet Digital Health* 3.1 (Jan. 2021), e10–e19. ISSN: 25897500. DOI: 10.1016/S2589-7500(20)30250-8. URL: https://linkinghub.elsevier.com/retrieve/pii/S2589750020302508.

[71] Paisan Ruamviboonsuk et al. "Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program". en. In: *npj Digital Medicine* 2.1 (Apr. 2019), p. 25. ISSN: 2398-6352. DOI: 10.1038/s41746-019-0099-8. URL: https://www.nature.com/articles/s41746-019-0099-8.

[72] Sambit S. Mondal et al. "EDLDR: An Ensemble Deep Learning Technique for Detection and Classification of Diabetic Retinopathy". In: *Diagnostics* 13.1 (Dec. 2022), p. 124. ISSN: 2075-4418. DOI: 10.3390/diagnostics13010124. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9818466/.

[73] en. URL: https://kaggle.com/aptos2019-blindness-detection.

[74] T Kauppi et al. *The DIARETDB1 diabetic retinopathy database and evaluation protocol. British Machine Vision Conference*. 2007.

[75] Haoxuan Che, Haibo Jin, and Hao Chen. "Learning Robust Representation for Joint Grading of Ophthalmic Diseases via Adaptive Curriculum and Feature Disentanglement". en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by Linwei Wang et al. Vol. 13433. Cham: Springer Nature Switzerland, 2022, pp. 523–533. ISBN: 9783031164361. DOI: 10.1007/978-3-031-16437-8_50. URL: https://link.springer.com/10.1007/978-3-031-16437-8_50.

[76] Xianjing Liu et al. *DeepDRiD: The DeepDR diabetic retinopathy image dataset (DeepDRiD) website*. https://isbi.deepdr.org. 2025.

[77] Prasanna Porwal et al. "Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research". en. In: *Data* 3.3 (July 2018), p. 25. ISSN: 2306-5729. DOI: 10.3390/data3030025. URL: https://www.mdpi.com/2306-5729/3/3/25.

[78] Quang Toan Dao, Hoang Quan Trinh, and Viet Anh Nguyen. "An effective and comprehensible method to detect and evaluate retinal damage due to diabetes complications". en. In: *PeerJ Computer Science* 9 (Sept. 2023), e1585. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.1585. URL: https://peerj.com/articles/cs-1585.

[79] Elmeeh Hasan Shipra and Md. Sazzadur Rahman. "An Explainable Artificial Intelligence Strategy for Transparent Deep Learning in the Classification of Eye Diseases". In: *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*. Cox's Bazar, Bangladesh: IEEE, Sept. 2024, pp. 1–6. ISBN: 9798331529765. DOI: 10.1109/COMPAS60761.2024.10797058. URL: https://ieeexplore.ieee.org/document/10797058/.

[80] Guna Venkat Doddi. *Eye_diseases_classification*. 2022.

[81] Gary S Collins et al. "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods". en. In: *BMJ* (Apr. 2024), e078378. ISSN: 1756-1833. DOI: 10.1136/bmj-2023-078378. URL: https://www.bmj.com/lookup/doi/10.1136/bmj-2023-078378.

[82] Alessandro Castelnovo et al. "A clarification of the nuances in the fairness metrics landscape". en. In: *Scientific Reports* 12.1 (Mar. 2022), p. 4209. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07939-1. URL: https://www.nature.com/articles/s41598-022-07939-1.

[83] Emma A.M. Stanley et al. “Where, why, and how is bias learned in medical image analysis models? A study of bias encoding within convolutional networks using synthetic data”. en. In: *eBioMedicine* 111 (Jan. 2025), p. 105501. ISSN: 23523964. DOI: 10.1016/j.ebiom.2024.105501. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352396424005371.

[84] Will Fleisher. “What’s Fair About Individual Fairness?” en. In: *SSRN Electronic Journal* (2021). ISSN: 1556-5068. DOI: 10.2139/ssrn.3819799. URL: https://www.ssrn.com/abstract=3819799.

[85] Reuben Binns. “On the apparent conflict between individual and group fairness”. en. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona Spain: ACM, Jan. 2020, pp. 514–524. ISBN: 9781450369367. DOI: 10.1145/3351095.3372864. URL: https://dl.acm.org/doi/10.1145/3351095.3372864.

[86] Cong Su et al. “A review of causality-based fairness machine learning”. In: *Intelligence Robotics* 2.3 (2022), pp. 244–274. ISSN: 27703541. DOI: 10.20517/ir.2022.17. URL: https://www.oaepublish.com/articles/ir.2022.17.

[87] Yan Luo et al. “FairVision: Equitable Deep Learning for Eye Disease Screening via Fair Identity Scaling”. In: (2023). DOI: 10.48550/ARXIV.2310.02492. URL: https://arxiv.org/abs/2310.02492.

[88] Alexander Brown et al. “Detecting shortcut learning for fair medical AI using shortcut testing”. en. In: *Nature Communications* 14.1 (July 2023), p. 4314. ISSN: 2041-1723. DOI: 10.1038/s41467-023-39902-7. URL: https://www.nature.com/articles/s41467-023-39902-7.

[89] Laleh Seyyed-Kalantari et al. “CheXclusion: Fairness gaps in deep chest X-ray classifiers”. In: (2020). DOI: 10.48550/ARXIV.2003.00827. URL: https://arxiv.org/abs/2003.00827.

[90] Min Shi et al. “Equitable deep learning for diabetic retinopathy detection using multi-dimensional retinal imaging with fair adaptive scaling: a retrospective study”. en. In: (Apr. 2024). DOI: 10.1101/2024.04.13.24305759. URL: http://medrxiv.org/lookup/doi/10.1101/2024.04.13.24305759.

[91] Agostina J. Larrazabal et al. “Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis”. en. In: *Proceedings of the National Academy of Sciences* 117.23 (June 2020), pp. 12592–12594. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1919012117. URL: https://pnas.org/doi/full/10.1073/pnas.1919012117.

[92] Nina Weng et al. “Are Sex-based Physiological Differences the Cause of Gender Bias for Chest X-ray Diagnosis?” In: (2023). DOI: 10.48550/ARXIV.2308.05129. URL: https://arxiv.org/abs/2308.05129.

[93] Jose Benitez-Aurioles et al. “Understanding algorithmic fairness for clinical prediction in terms of subgroup net benefit and health equity”. In: (2024). DOI: 10.48550/ARXIV.2412.07879. URL: https://arxiv.org/abs/2412.07879.

[94] Zikang Xu et al. “Fairness in Medical Image Analysis and Healthcare: A Literature Survey”. In: (Oct. 2023). DOI: 10.36227/techrxiv.24324979. URL: https://www.techrxiv.org/articles/preprint/Fairness_in_Medical_Image_Analysis_and_Healthcare_A_Literature_Survey/24324979.

[95] Valeriia Cherepanova et al. “Technical Challenges for Training Fair Neural Networks”. In: (2021). DOI: 10.48550/ARXIV.2102.06764. URL: https://arxiv.org/abs/2102.06764.

[96] Siyi Du et al. “FairDisCo: Fairer AI in Dermatology via Disentanglement Contrastive Learning”. en. In: *Computer Vision – ECCV 2022 Workshops*. Ed. by Leonid Karlinsky, Tomer Michaeli, and Ko Nishino. Vol. 13804. Cham: Springer Nature Switzerland, 2023, pp. 185–202. ISBN: 9783031250682. DOI: 10.1007/978-3-031-25069-9_13. URL: https://link.springer.com/10.1007/978-3-031-25069-9_13.

[97] Barath Narayanan Narayanan et al. “Hybrid machine learning architecture for automated detection and grading of retinal images for diabetic retinopathy”. In: *Journal of Medical Imaging* 7.3 (May 2020), p. 034501. ISSN: 2329-4302. DOI: 10.1117/1.JMI.7.3.034501. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7309178/.

[98] José Martins, Jaime S. Cardoso, and Filipe Soares. “Offline computer-aided diagnosis for Glaucoma detection using fundus images targeted at mobile devices”. In: *Computer Methods and Programs in Biomedicine* 192 (Aug. 2020), p. 105341. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2020.105341. URL: https://www.sciencedirect.com/science/article/pii/S0169260719312015.

[99] Qier Meng, Yohei Hashimoto, and Shin’ichi Satoh. “How to Extract More Information with Less Burden: Fundus Image Classification and Retinal Disease Localization with Ophthalmologist Intervention”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. Iowa City, IA, USA: IEEE, Apr. 2020, pp. 1373–1377. ISBN: 9781538693308. DOI: 10.1109/ISBI45749.2020.9098600. URL: https://ieeexplore.ieee.org/document/9098600/.

[100] Been Kim et al. “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)”. In: (2017). DOI: 10.48550/ARXIV.1711.11279. URL: https://arxiv.org/abs/1711.11279.

[101] Mhd Hasan Sarhan et al. “Learning Interpretable Disentangled Representations using Adversarial VAEs”. In: (2019). DOI: 10.48550/ARXIV.1904.08491. URL: https://arxiv.org/abs/1904.08491.

[102] Wilson Silva et al. "Interpretability-Guided Content-Based Medical Image Retrieval". en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by Anne L. Martel et al. Vol. 12261. Cham: Springer International Publishing, 2020, pp. 305–314. ISBN: 9783030597092. DOI: 10.1007/978-3-030-59710-8_30. URL: https://link.springer.com/10.1007/978-3-030-59710-8_30.

[103] Mengze Li et al. "IB-M: A Flexible Framework to Align an Interpretable Model and a Black-box Model". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Dec. 2020, pp. 643–649. DOI: 10.1109/BIBM49941.2020.9313119. URL: https://ieeexplore.ieee.org/abstract/document/9313119.

[104] Edward Korot et al. "Predicting sex from retinal fundus photographs using automated deep learning". en. In: *Scientific Reports* 11.1 (May 2021), p. 10286. ISSN: 2045-2322. DOI: 10.1038/s41598-021-89743-x. URL: https://www.nature.com/articles/s41598-021-89743-x.

[105] Ryan Poplin et al. "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning". en. In: *Nature Biomedical Engineering* 2.3 (Mar. 2018), pp. 158–164. ISSN: 2157-846X. DOI: 10.1038/s41551-018-0195-0. URL: https://www.nature.com/articles/s41551-018-0195-0.

[106] Sarah Müller et al. "Disentangling representations of retinal images with generative models". In: (2024). DOI: 10.48550/ARXIV.2402.19186. URL: https://arxiv.org/abs/2402.19186.

[107] Cathy Ong Ly et al. "Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data". en. In: *npj Digital Medicine* 7.1 (May 2024), pp. 1–10. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01118-4. URL: https://www.nature.com/articles/s41746-024-01118-4.

[108] Qingyu Zhao, Ehsan Adeli, and Kilian M. Pohl. "Training confounder-free deep learning models for medical applications". en. In: *Nature Communications* 11.1 (Nov. 2020), p. 6010. ISSN: 2041-1723. DOI: 10.1038/s41467-020-19784-9. URL: https://www.nature.com/articles/s41467-020-19784-9.

[109] Yuzhe Yang et al. "The limits of fair medical imaging AI in real-world generalization". en. In: *Nature Medicine* 30.10 (Oct. 2024), pp. 2838–2848. ISSN: 1546-170X. DOI: 10.1038/s41591-024-03113-4. URL: https://www.nature.com/articles/s41591-024-03113-4.

[110] Jiahong Ouyang et al. "Representation Disentanglement for Multi-modal Brain MRI Analysis". en. In: *Information Processing in Medical Imaging*. Ed. by Aasa Feragen et al. Vol. 12729. Cham: Springer International Publishing, 2021, pp. 321–333. ISBN: 9783030781903. DOI: 10.

1007/978-3-030-78191-0_25. URL: https://link.springer.com/10.1007/978-3-030-78191-0_25.

[111] Agisilaos Chartsias et al. "Disentangled representation learning in cardiac image analysis". In: *Medical image analysis* 58 (Dec. 2019), p. 101535. ISSN: 1361-8415. DOI: 10.1016/j.media.2019.101535. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6815716/.

[112] Francesco Locatello et al. "On the Fairness of Disentangled Representations". In: (2019). DOI: 10.48550/ARXIV.1905.13662. URL: https://arxiv.org/abs/1905.13662.

[113] Helena Montenegro and Jaime S. Cardoso. "Anonymizing medical case-based explanations through disentanglement". en. In: *Medical Image Analysis* 95 (July 2024), p. 103209. ISSN: 13618415. DOI: 10.1016/j.media.2024.103209. URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841524001348.

[114] Zhihao Zhao et al. "Unobtrusive biometric data de-identification of fundus images using latent space disentanglement". en. In: *Biomedical Optics Express* 14.10 (Oct. 2023), p. 5466. ISSN: 2156-7085, 2156-7085. DOI: 10.1364/BOE.495438. URL: https://opg.optica.org/abstract.cfm?URI=boe-14-10-5466.

[115] Peng Xia et al. "Generalizing to Unseen Domains in Diabetic Retinopathy with Disentangled Representations". en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Ed. by Marius George Linguraru et al. Cham: Springer Nature Switzerland, 2024, pp. 427–437. ISBN: 9783031721175. DOI: 10.1007/978-3-031-72117-5_40.

[116] Sharon Chokuwa and Muhammad H. Khan. "Generalizing Across Domains in Diabetic Retinopathy via Variational Autoencoders". In: (2023). DOI: 10.48550/ARXIV.2309.11301. URL: https://arxiv.org/abs/2309.11301.

[117] Luis Filipe Nakayama et al. *A Brazilian Multilabel Ophthalmological Dataset (BRSET)*. DOI: 10.13026/1PHT-2B69. URL: https://physionet.org/content/brazilian-ophthalmological/1.0.1/.

[118] Chenwei Wu et al. "mBRSET: A Portable Retina Fundus Photos Benchmark Dataset for Clinical and Demographic Prediction". en. In: (July 2024). DOI: 10.1101/2024.07.11.24310293. URL: http://medrxiv.org/lookup/doi/10.1101/2024.07.11.24310293.

[119] en. Nov. 2021. URL: https://365datascience.com/tutorials/machine-learning-tutorials/decision-trees/.

[120] Cristian Leo. *The Math Behind XGBoost*. en. Jan. 2024. URL: https://medium.com/@cristianleo120/the-math-behind-xgboost-3068c78aad9d.

[121] Sanghyun Woo et al. "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders". In: (2023). DOI: 10.48550/ARXIV.2301.00808. URL: https://arxiv.org/abs/2301.00808.

[122] Maxime Oquab et al. “DINOv2: Learning Robust Visual Features without Supervision”. In: (2023). DOI: 10.48550/ARXIV.2304.07193. URL: https://arxiv.org/abs/2304.07193.

[123] Ze Liu et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. In: (2021). DOI: 10.48550/ARXIV.2103.14030. URL: https://arxiv.org/abs/2103.14030.

[124] Tsung-Yi Lin et al. “Focal Loss for Dense Object Detection”. In: (2017). DOI: 10.48550/ARXIV.1708.02002. URL: https://arxiv.org/abs/1708.02002.

[125] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: (2014). DOI: 10.48550/ARXIV.1412.6980. URL: https://arxiv.org/abs/1412.6980.

[126] Andrew J. Vickers, Ben Van Calster, and Ewout W. Steyerberg. “A simple, step-by-step guide to interpreting decision curve analysis”. en. In: *Diagnostic and Prognostic Research* 3.1 (Dec. 2019), p. 18. ISSN: 2397-7523. DOI: 10.1186/s41512-019-0064-7. URL: https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-019-0064-7.

[127] Ben Van Calster et al. “Performance evaluation of predictive AI models to support medical decisions: Overview and guidance”. In: (2024). DOI: 10.48550/ARXIV.2412.10288. URL: https://arxiv.org/abs/2412.10288.

[128] Chaya Langerman, Angus Forbes, and Glenn Robert. “The experiences of insulin use among older people with Type 2 diabetes mellitus: A thematic synthesis”. en. In: *Primary Care Diabetes* 16.5 (Oct. 2022), pp. 614–626. ISSN: 17519918. DOI: 10.1016/j.pcd.2022.08.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S175199182200136X.

[129] Glenn Linde et al. “A comparative evaluation of deep learning approaches for ophthalmology”. en. In: *Scientific Reports* 14.1 (Sept. 2024), p. 21829. ISSN: 2045-2322. DOI: 10.1038/s41598-024-72752-x. URL: https://www.nature.com/articles/s41598-024-72752-x.

[130] Xianjing Liu et al. “AI-based association analysis for medical imaging using latent-space geometric confounder correction”. eng. In: *Medical Image Analysis* 102 (May 2025), p. 103529. ISSN: 1361-8423. DOI: 10.1016/j.media.2025.103529.

Appendix A

Complete results

In Section A.1, we present the complete results of the decision curve analysis and risk distribution plots for each sensitive attribute subgroup in both mBRSET and BRSET. We evaluate the three models, ConvNeXt V2, DINOv2, and Swin V2, across these attributes. For mBRSET, the attributes include sex, age, insurance, educational level, and obesity; for BRSET, the attributes include sex and age. Each attribute defines a division of the dataset into corresponding subgroups.

In Section A.2, we present the results of all disentanglement experiments. These includes disentangling age and sex information from ConvNeXt V2, as well as sex and insurance information from both DINOv2 and Swin V2.

A.1 Complete results of fairness analysis for baseline models

A.1.0.1 mBRSET - Sex

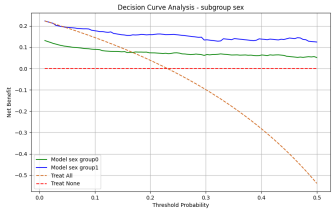


Figure A.1: Decision Curve Analysis - ConvNeXt V2 mBRSET.

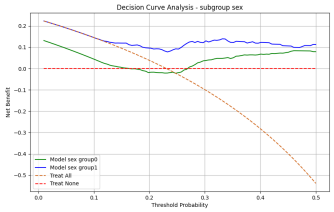


Figure A.2: Decision Curve Analysis - DINOv2 mBRSET.

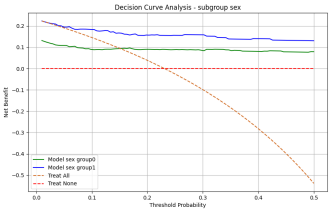


Figure A.3: Decision Curve Analysis - Swin V2 mBRSET.

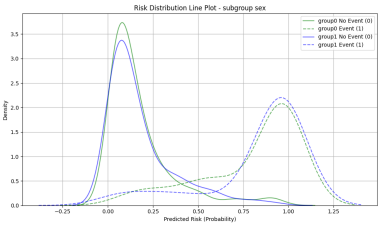


Figure A.4: Risk Distribution - ConvNeXt V2 mBRSET.

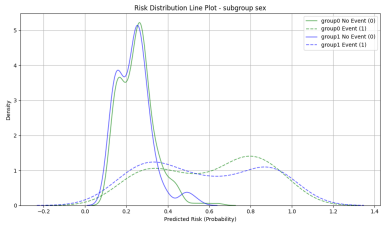


Figure A.5: Risk Distribution - DINOv2 mBRSET.

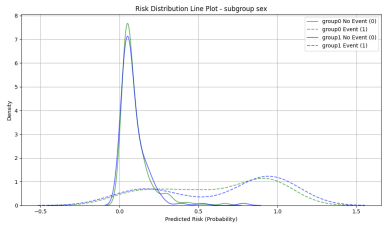


Figure A.6: Risk Distribution - Swin V2 mBRSET.

A.1.0.2 mBRSET - Age

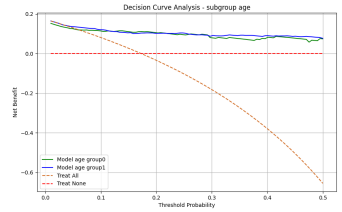


Figure A.7: Decision Curve Analysis - ConvNeXt V2 mBRSET.

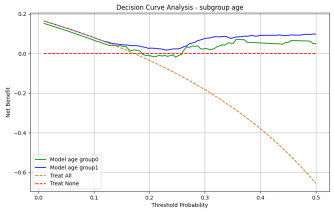


Figure A.8: Decision Curve Analysis - DINOv2 mBRSET.

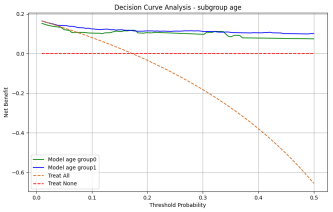


Figure A.9: Decision Curve Analysis - Swin V2 mBRSET.

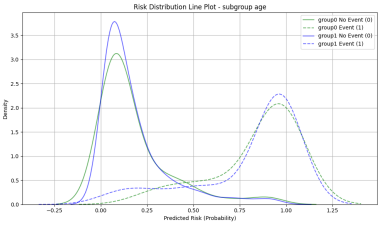


Figure A.10: Risk Distribution - ConvNeXt V2 mBRSET.

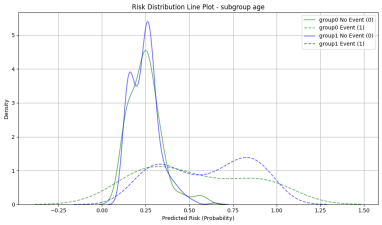


Figure A.11: Risk Distribution - DINOv2 mBRSET.

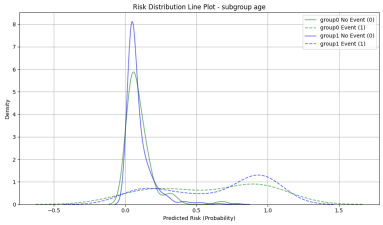


Figure A.12: Risk Distribution - Swin V2 mBRSET.

A.1.0.3 mBRSET - Insurance

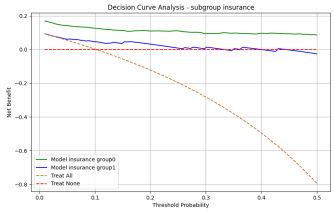


Figure A.13: Decision Curve Analysis - ConvNeXt V2 mBRSET.

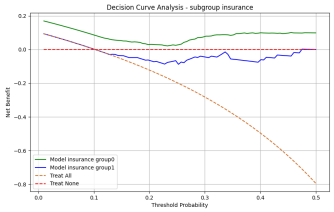


Figure A.14: Decision Curve Analysis - DINOv2 mBRSET.

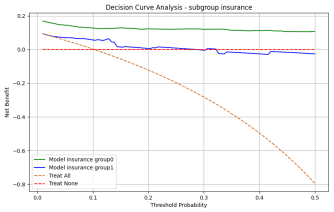


Figure A.15: Decision Curve Analysis - Swin V2 mBRSET.

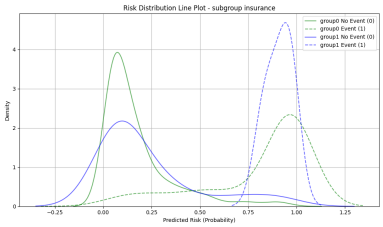


Figure A.16: Risk Distribution - ConvNeXt V2 mBRSET.

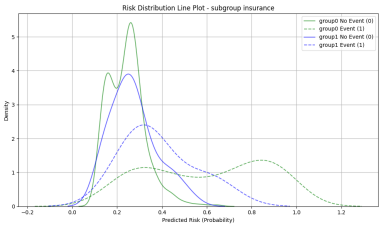


Figure A.17: Risk Distribution - DINOv2 mBRSET.

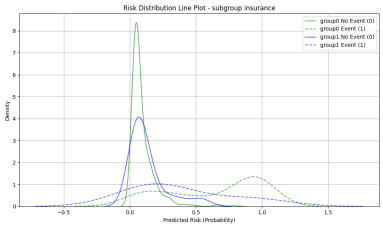


Figure A.18: Risk Distribution - Swin V2 mBRSET.

A.1.0.4 mBRSET - Educational level

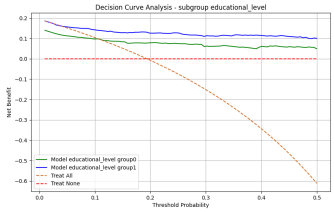


Figure A.19: Decision Curve Analysis - ConvNeXt V2 mBRSET.

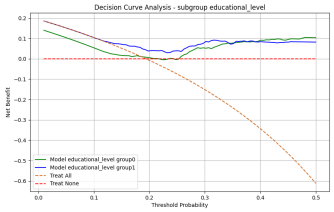


Figure A.20: Decision Curve Analysis - DINOv2 mBRSET.

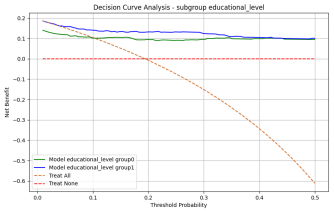


Figure A.21: Decision Curve Analysis - Swin V2 mBRSET.

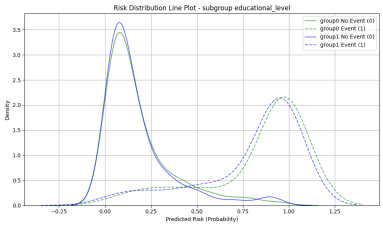


Figure A.22: Risk Distribution - ConvNeXt V2 mBRSET.

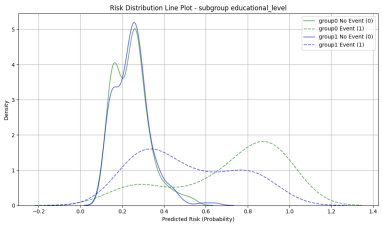


Figure A.23: Risk Distribution - DINOv2 mBRSET.

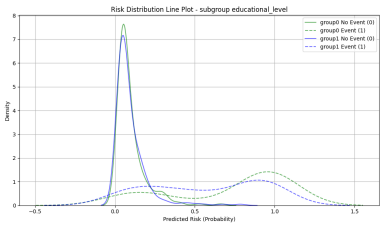


Figure A.24: Risk Distribution - Swin V2 mBRSET.

A.1.0.5 mBRSET - Obesity

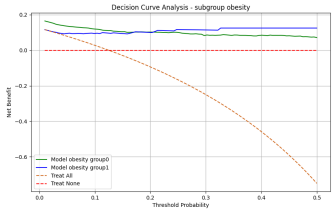


Figure A.25: Decision Curve Analysis - ConvNeXt V2 mBRSET.

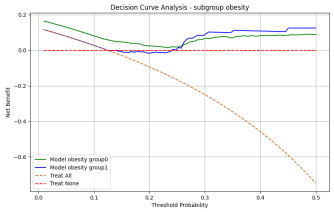


Figure A.26: Decision Curve Analysis - DINOv2 mBRSET.

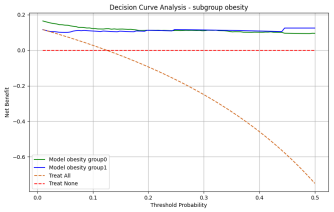


Figure A.27: Decision Curve Analysis - Swin V2 mBRSET.

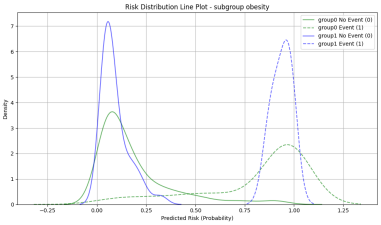


Figure A.28: Risk Distribution - ConvNeXt V2 mBRSET.

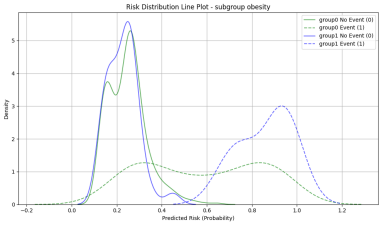


Figure A.29: Risk Distribution - DINOv2 mBRSET.

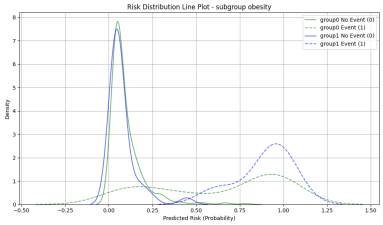


Figure A.30: Risk Distribution - Swin V2 mBRSET.

A.1.0.6 BRSET - Sex

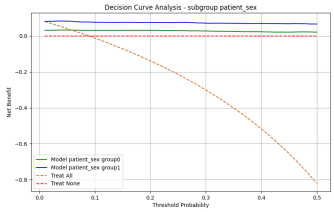


Figure A.31: Decision Curve Analysis - ConvNeXt V2 BRSET.

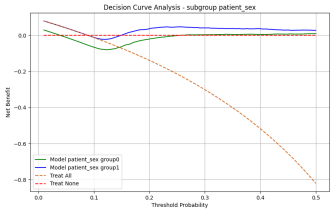


Figure A.32: Decision Curve Analysis - DINOv2 BRSET.

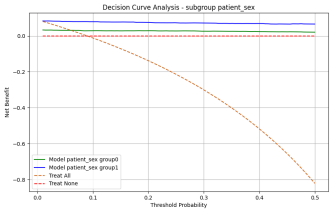


Figure A.33: Decision Curve Analysis - Swin V2 mBRSET.

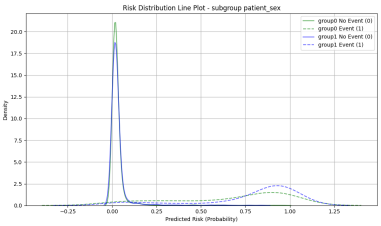


Figure A.34: Risk Distribution - ConvNeXt V2 BRSET.

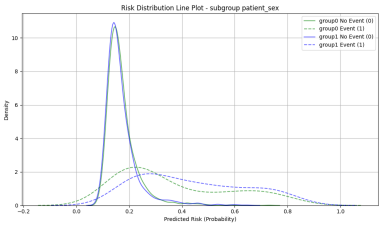


Figure A.35: Risk Distribution - DINOv2 BRSET.

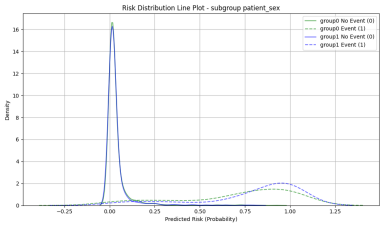


Figure A.36: Risk Distribution - Swin V2 mBRSET.

A.1.0.7 BRSET - Age

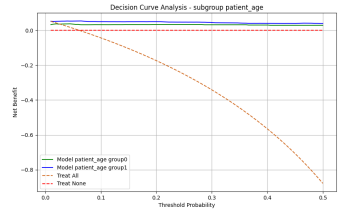


Figure A.37: Decision Curve Analysis - ConvNeXt V2 BRSET.

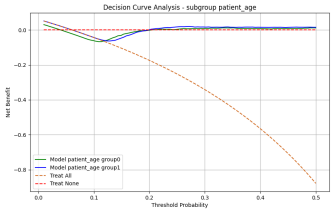


Figure A.38: Decision Curve Analysis - DINOv2 BRSET.

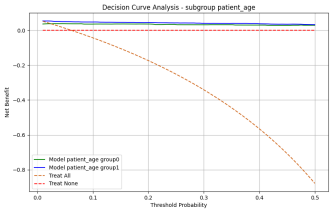


Figure A.39: Decision Curve Analysis - Swin V2 mBRSET.

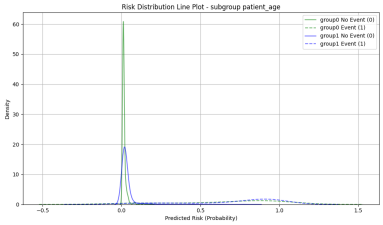


Figure A.40: Risk Distribution - ConvNeXt V2 BRSET.

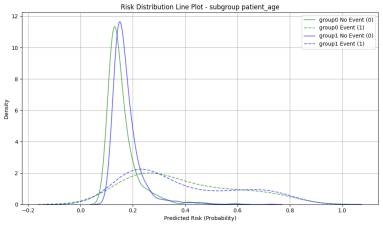


Figure A.41: Risk Distribution - DINOv2 BRSET.

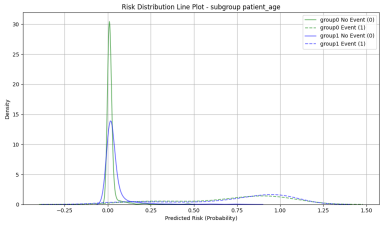
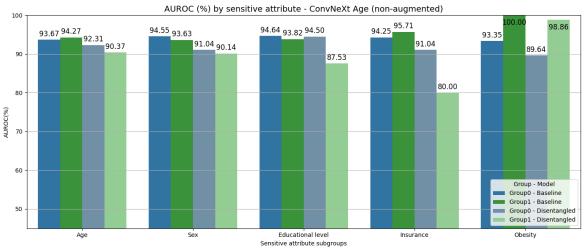


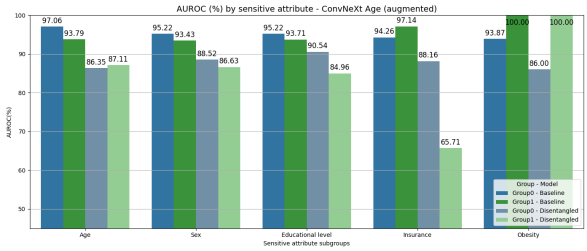
Figure A.42: Risk Distribution - Swin V2 mBRSET.

A.2 Complete Results of Disentanglement

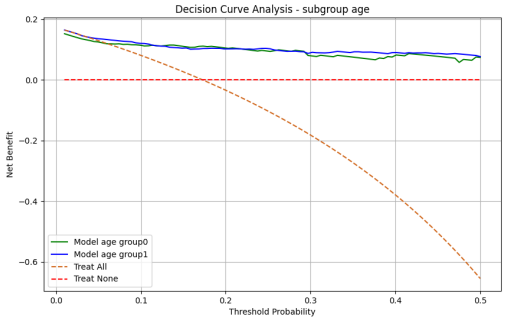
A.2.0.1 ConvNeXt V2 - Age



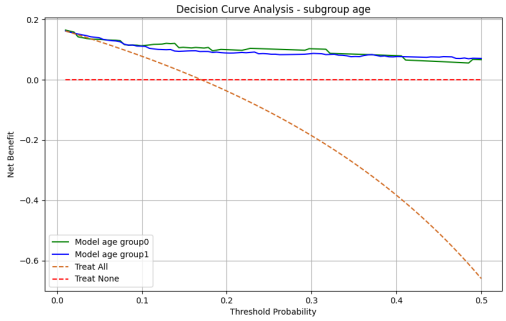
(a) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the non-augmented mBRSET dataset.



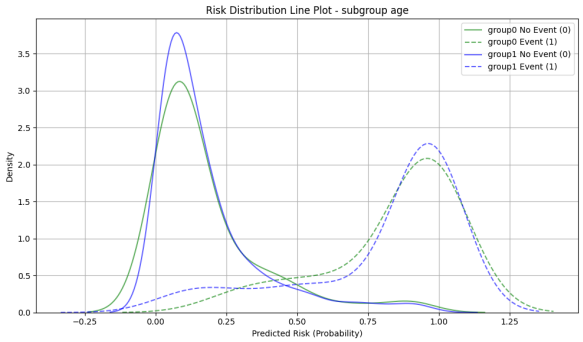
(b) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the augmented mBRSET dataset.



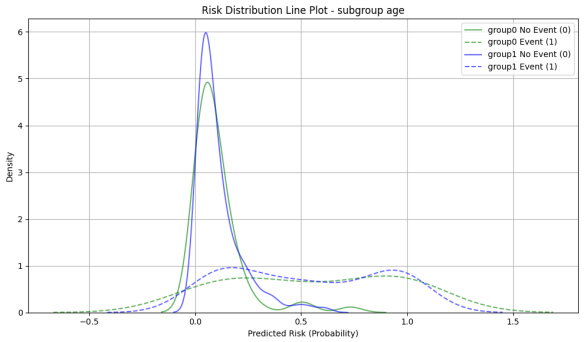
(c) Decision curve analysis of the baseline model on the augmented mBRSET dataset.



(d) Decision curve analysis of the disentangled model on the augmented mBRSET dataset.



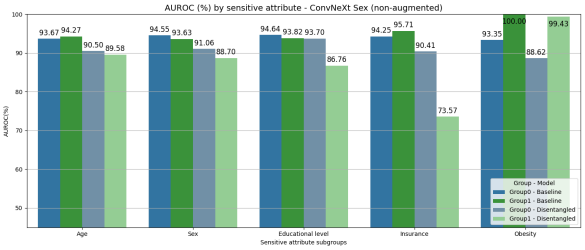
(e) Risk distribution plot of the baseline model on the augmented mBRSET dataset.



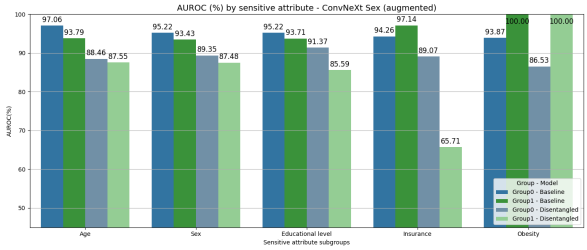
(f) Risk distribution plot of the disentangled model on the augmented mBRSET dataset.

Figure A.43: Comparison of baseline and disentangled model performance on the mBRSET dataset across AUROC, decision curve analysis, and risk distribution plots.

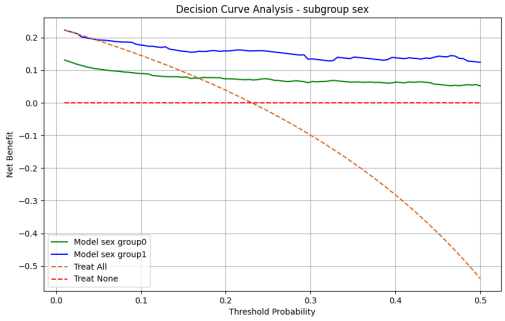
A.2.0.2 ConvNeXt V2 - Sex



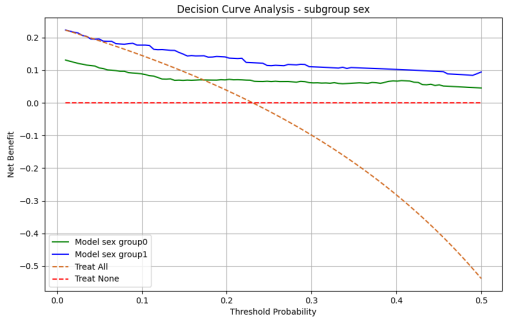
(a) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the non-augmented mBRSET dataset.



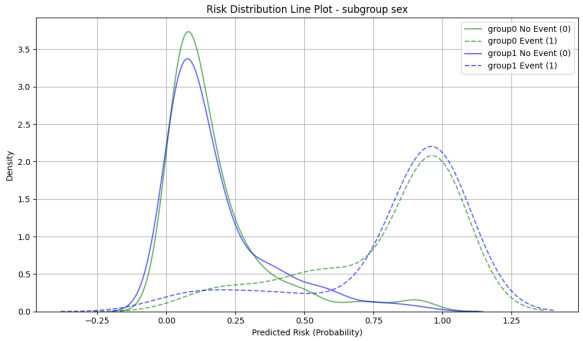
(b) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the augmented mBRSET dataset.



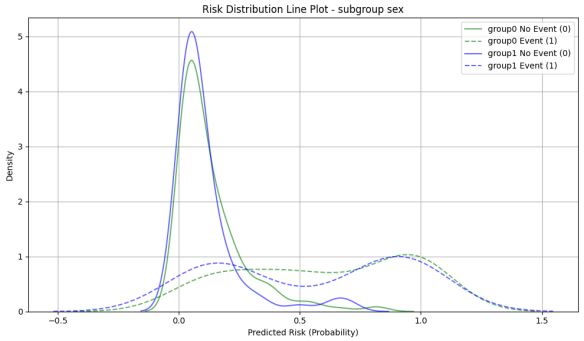
(c) Decision curve analysis of the baseline model on the augmented mBRSET dataset.



(d) Decision curve analysis of the disentangled model on the augmented mBRSET dataset.



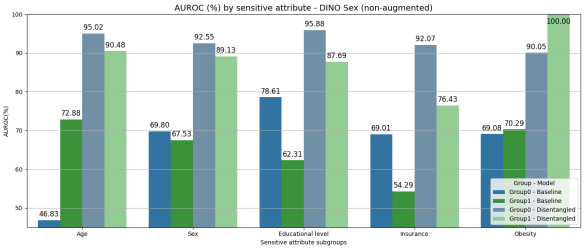
(e) Risk distribution plot of the baseline model on the augmented mBRSET dataset.



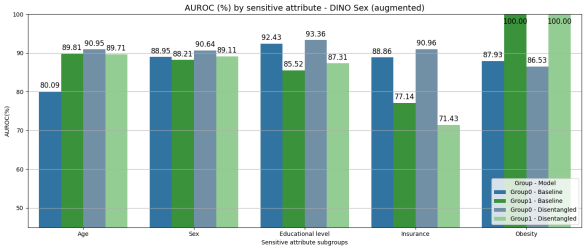
(f) Risk distribution plot of the disentangled model on the augmented mBRSET dataset.

Figure A.44: Comparison of baseline and disentangled model performance on the mBRSET dataset across AUROC, decision curve analysis, and risk distribution plots.

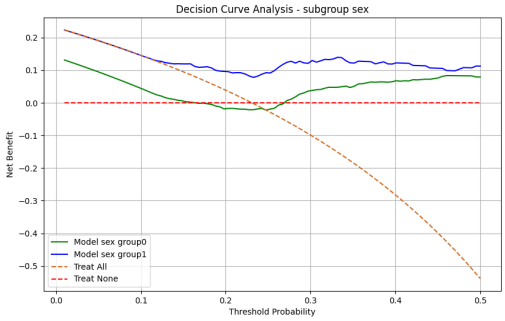
A.2.0.3 DINOv2 - Sex



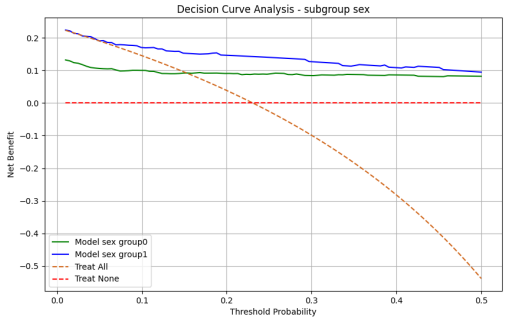
(a) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the non-augmented mBRSET dataset.



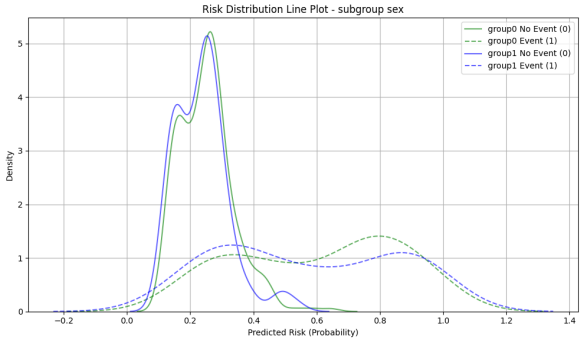
(b) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the augmented mBRSET dataset.



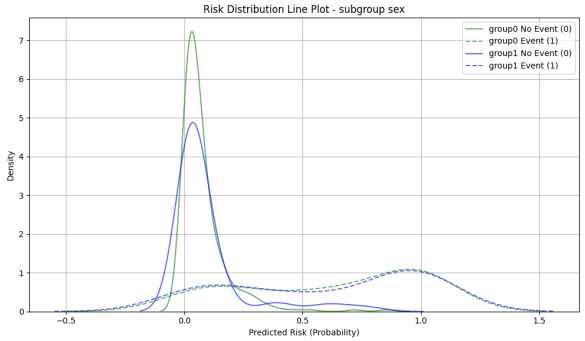
(c) Decision curve analysis of the baseline model on the augmented mBRSET dataset.



(d) Decision curve analysis of the disentangled model on the augmented mBRSET dataset.



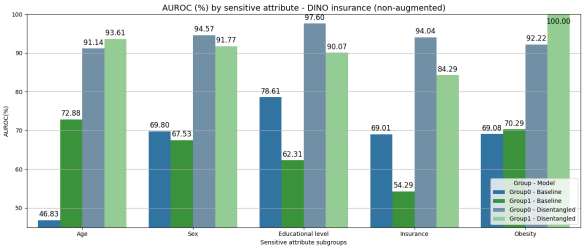
(e) Risk distribution plot of the baseline model on the augmented mBRSET dataset.



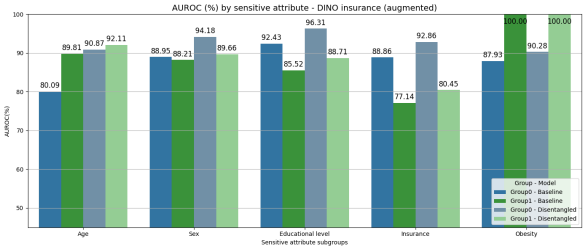
(f) Risk distribution plot of the disentangled model on the augmented mBRSET dataset.

Figure A.45: Comparison of baseline and disentangled model performance on the mBRSET dataset across AUROC, decision curve analysis, and risk distribution plots.

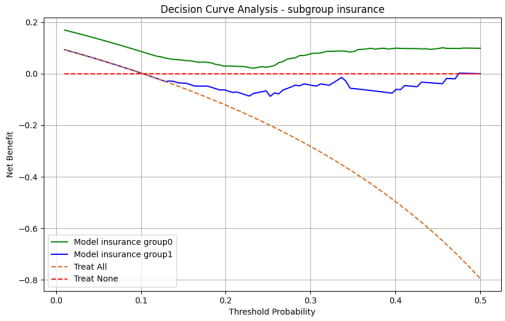
A.2.0.4 DINOv2 - Insurance



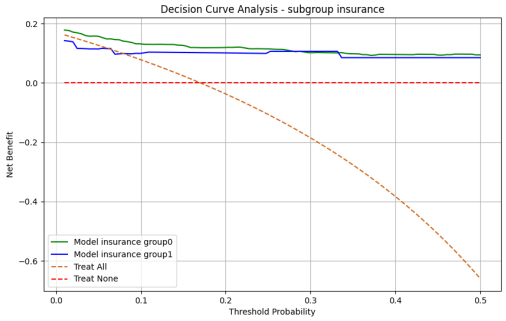
(a) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the non-augmented mBRSET dataset.



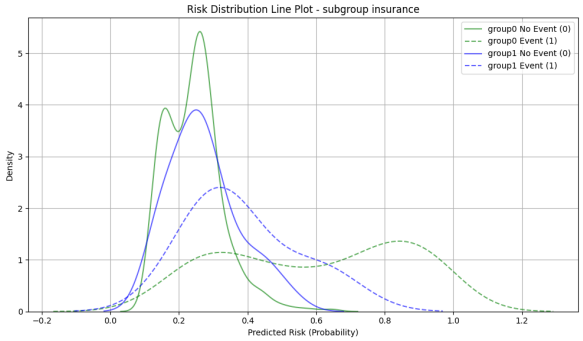
(b) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the augmented mBRSET dataset.



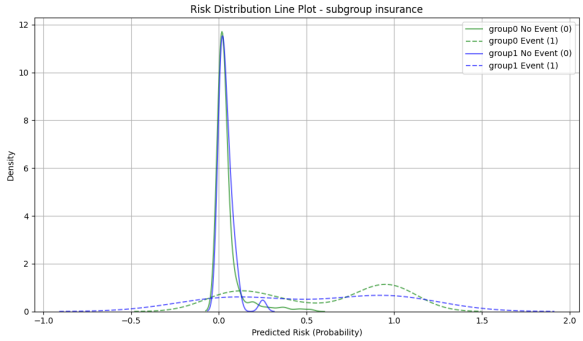
(c) Decision curve analysis of the baseline model on the augmented mBRSET dataset.



(d) Decision curve analysis of the disentangled model on the augmented mBRSET dataset.



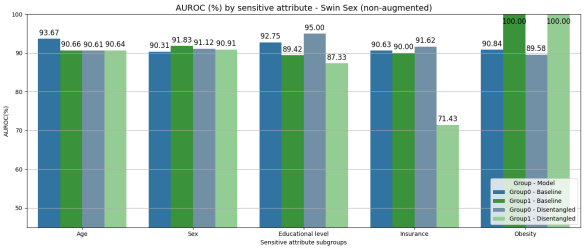
(e) Risk distribution plot of the baseline model on the augmented mBRSET dataset.



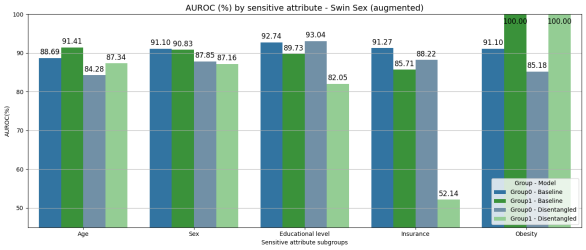
(f) Risk distribution plot of the disentangled model on the augmented mBRSET dataset.

Figure A.46: Comparison of baseline and disentangled model performance on the mBRSET dataset across AUROC, decision curve analysis, and risk distribution plots.

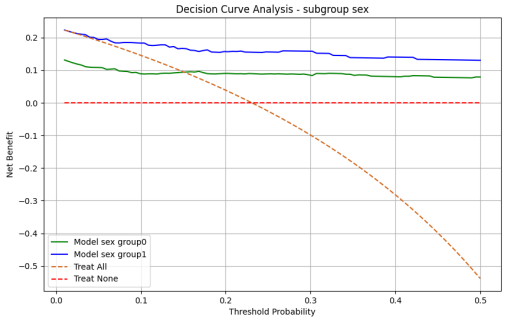
A.2.0.5 Swin V2 - Sex



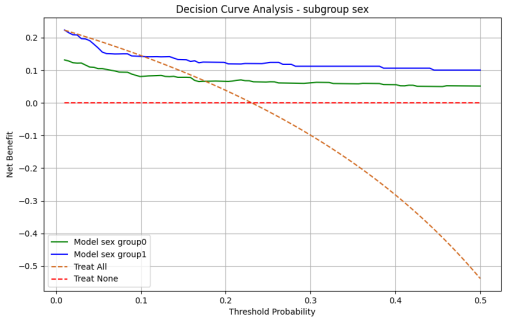
(a) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the non-augmented mBRSET dataset.



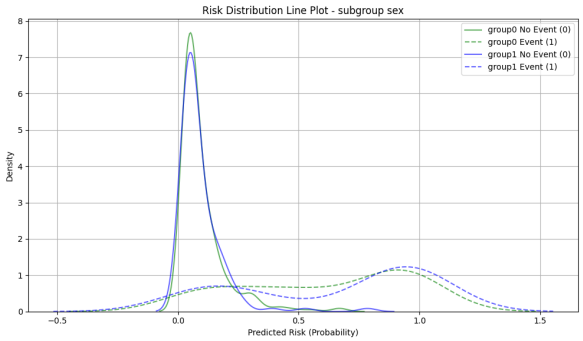
(b) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the augmented mBRSET dataset.



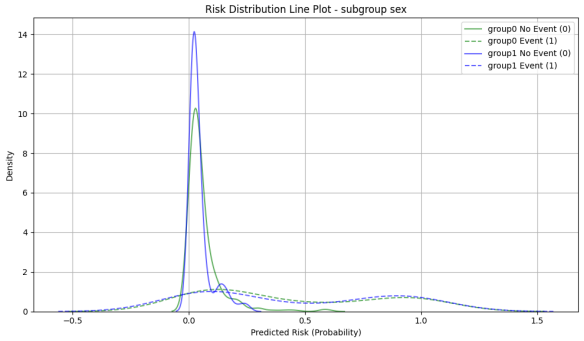
(c) Decision curve analysis of the baseline model on the augmented mBRSET dataset.



(d) Decision curve analysis of the disentangled model on the augmented mBRSET dataset.



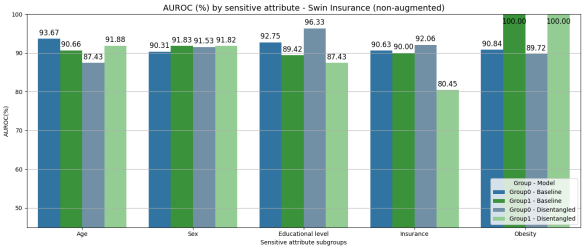
(e) Risk distribution plot of the baseline model on the augmented mBRSET dataset.



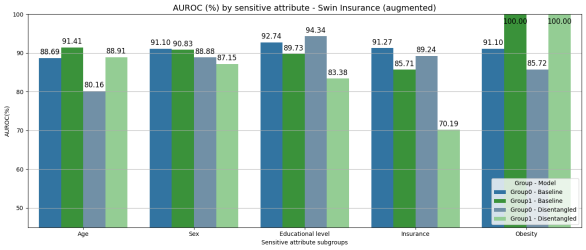
(f) Risk distribution plot of the disentangled model on the augmented mBRSET dataset.

Figure A.47: Comparison of baseline and disentangled model performance on the mBRSET dataset across AUROC, decision curve analysis, and risk distribution plots.

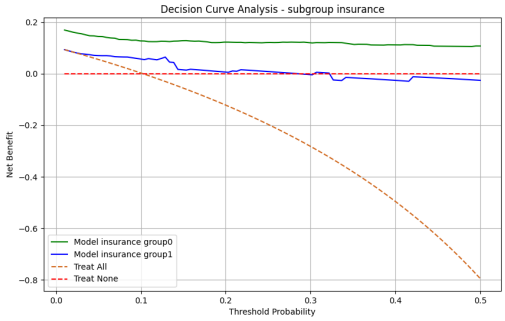
A.2.0.6 Swin V2 - Insurance



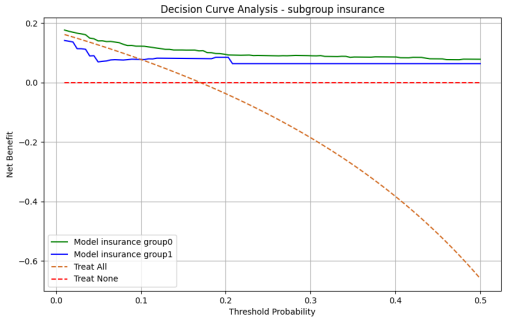
(a) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the non-augmented mBRSET dataset.



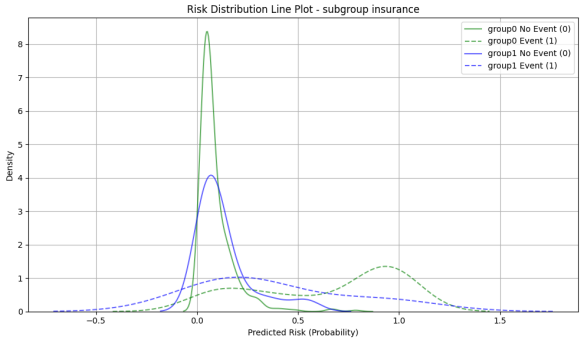
(b) Comparison of AUROC scores across SA groups for the baseline and disentangled models, using the augmented mBRSET dataset.



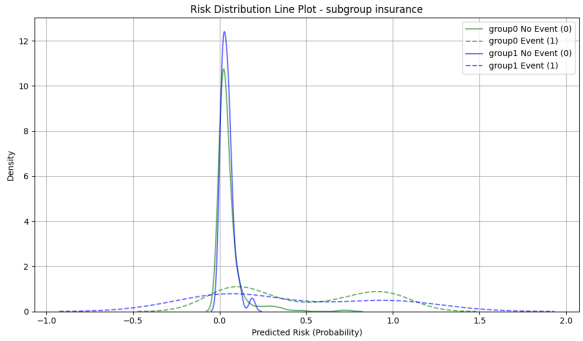
(c) Decision curve analysis of the baseline model on the augmented mBRSET dataset.



(d) Decision curve analysis of the disentangled model on the augmented mBRSET dataset.



(e) Risk distribution plot of the baseline model on the augmented mBRSET dataset.



(f) Risk distribution plot of the disentangled model on the augmented mBRSET dataset.

Figure A.48: Comparison of baseline and disentangled model performance on the mBRSET dataset across AUROC, decision curve analysis, and risk distribution plots.