T.C.
IŞIK UNIVERSITY
SCHOOL OF GRADUATE STUDIES

DOCTORAL THESIS
DEPARTMENT OF COMPUTER ENGINEERING
COMPUTER ENGINEERING PROGRAM

Yasemin TURKAN

DEEP LEARNING-BASED ANALYSIS OF RETINAL OCT SCANS FOR DETECTION OF ALZHEIMER'S DISEASE

SUPERVISOR
Assoc. Prof. Faik Boray TEK

İSTANBUL, January 2026

T.C.
IŞIK UNIVERSITY
SCHOOL OF GRADUATE STUDIES

DOCTORAL THESIS
DEPARTMENT OF COMPUTER ENGINEERING
COMPUTER ENGINEERING PROGRAM

Yasemin TURKAN
(219DCS8081)

DEEP LEARNING-BASED ANALYSIS OF RETINAL OCT SCANS FOR DETECTION OF ALZHEIMER'S DISEASE

SUPERVISOR
Assoc. Prof. Faik Boray TEK

İSTANBUL, January 2026

T.C.
IŞIK UNIVERSITY
SCHOOL OF GRADUATE STUDIES

DOCTORAL THESIS
DEPARTMENT OF COMPUTER ENGINEERING
PROGRAM

Yasemin TURKAN

(219DCS8081)

DEEP LEARNING-BASED ANALYSIS OF RETINAL OCT SCANS FOR DETECTION OF ALZHEIMER'S DISEASE

Date: 23.01.2026

Thesis Supervisor: Assoc. Prof. Faik Boray TEK / Istanbul Technical University

Jury Members: Prof. Dr. Selim AKSOY / Bilkent University

Assoc. Prof. İlkay ÖKSÜZ / Istanbul Technical University

Asst. Prof. Emine EKİN / Işık University

Asst. Prof. Tuğba ERKOÇ / Işık University

İSTANBUL, January 2026

ÖZET

ALZHEİMER HASTALIĞI TEŞHİSİNDE RETİNAL OKT GÖRÜNTÜLERİNİN DERİN ÖĞRENME TEMELLİ ANALİZİ

Ağ tabakası kalınlığındaki değişimler, Alzheimer hastalığı (AH) gibi nörodejen-eratif hastalıklarla ilis¸kilendirilmis¸tir. Bu yapısal değişiklikler, Optik Koherens Tomografi (OCT) adı verilen giris¸imsel olmayan bir görüntüleme teknolojisi kul-lanılarak ölçülebilmektedir. Önceki aras¸tırmalar çoğunlukla, OCT veya OCTA aygıtlarından elde edilen bölütlenmis¸ ağ tabakası kalınlığı ile AH arasındaki is-tatistiksel ilis¸kilere odaklanmıs¸tır. Geleneksel tıbbi görüntü sınıflandırma görev-lerinin aksine, görüntülemenin klinik tanıdan birkaç yıl önce gelmesi nedeniyle erken kestirim (tespit), tanı koymaktan daha zorludur. Derin öğrenme (DÖ), özellikle evris¸imli sinir ağları ve aktarımlı öğrenme aracılığıyla, görüntü tabanlı hastalık tespiti görevlerinde güçlü bir bas¸arım sergilemis¸tir. Ancak, erken AH tespiti için DÖ'nün doğrudan bölütlenmemis¸ ham OKT B-taraması görüntüleri üzerinde uygulanması henüz yeterince aras¸tırılmamıs¸tır. Bu nedenle, bu tezde, erken Alzheimer hastalığı tespiti için ham OCT görüntülerini kullanan derin öğrenme tabanlı bir yaklas¸ım önererek bu aras¸tırma boşluğunu ele alıyoruz.

Literatürdeki ilgili tüm çalıs¸malar, büyük ölçüde birlikte çalıs¸abilirlikten yoksun olan özel ve kurumsal kohortlara dayanmaktadır. Buna kars¸ılık, UK Biobank, 2022, ağ tabakası yapısı ile sistemik sağlık arasındaki ilis¸kileri aras¸tırmak için benzersiz bir kaynak sunmakta olup bilis¸sel ve sağlıkla ilgili verilerle bağlantılı 85.000'den fazla OCT taramasını içermektedir. İlk tarama (2010–2015) ile Temmuz 2023 arasında, veri kümesindeki 539 katılımcıya AH tanısı konmuştur.

Bu nedenle, UK Biobank OKTA taramalarının eksikliği nedeniyle bir mik-tar sınırlı olsa da, bu tezde OCT taramalarını kullanarak erken AH tespiti

yap-mak için bu veri kümesini kullandık. Titiz bir veri dıs¸lama sürecinin ardından bu çalıs¸ma, temel değerlendirmelerinden sonraki 4 yıl içinde AH tanısı konan katılımcıları seçerek hedeflenmis¸ 4 yıllık bir pencere kullanmıs¸tır. AH grubu; yas¸, cinsiyet, göz ve örneğe göre rastgele seçilmis¸ dengeli bir Sağlıklı Kontrol grubu (N = 30) ile es¸les¸tirilmis¸tir.

Ilk olarak, önceden eğitilmiş derin öğrenme mimarilerini kullanarak yalıtılmıs¸ 2B B-taramalarının kestirimsel değerini değerlendirdik. Bu testlerde, ResNet-34 modeli 0.624 ± 0.060 değerinde bir Ortalama AUC elde etmis¸tir. Bu B-taramalarının belirginlik haritası analizi, merkezi maküler bölgenin kritik önem-ini vurgularken, çevresel alanların modelin kararına göz ardı edilebilir bir katkı sağladığını göstermis¸tir. Yalıtılmıs¸ B-taramalarının sınırlamalarını as¸mak ve 3B bilgiden yararlanmak için, OCT B-taramalarından 3B tabanlı bir yüzeysel (en-face) kalınlık izdüs¸üm haritası olus¸turduk. Bu is¸ hattı, çevresel gürültüyü etkili bir s¸ekilde filtreleyen ve tanısal açıdan ilgili olan 3 mm'lik iç maküler bölg-eye odaklanacak s¸ekilde eniyilenmis¸tir. Kalınlık haritaları üzerine yaptığımız çalıs¸ma, Gangliyon Hücre Tabakasını (GHT), klinik öncesi AH'nin en önemli göstergesi olarak belirlemis¸tir. Yıl ağırlıklı bir yitim fonksiyonu ile GHT kalınlık haritaları üzerinde eğitilen VGG-19 modeli, 0.750 ± 0.037 ile en yüksek Orta-lama AUC değerine ulas¸mıs¸tır. Dikkat çekici bir s¸ekilde, geleneksel klinik ölçüt olan Ağ Tabakası Sinir Lifi Tabakası (RNFL), bu belirti öncesi kohortta göz ardı edilebilir bir kestirimsel değer sergilemis¸tir.

Kestirim doğruluğunu daha da artırmak ve klinik karar verme sürecini taklit etmek için Çok Kipli Yumus¸ak Oylamalı Topluluk modeli gelis¸tirdik. Bu model, B-taramalarından ve GHT-˙IPT kalınlık haritalarından elde edilen yapısal içgörüleri klinik ve demografik verilerle bütünles¸tirmektedir. Bu topluluk yaklas¸ımı, 0.85 ile en yüksek ortalama AUC değerine ulas¸mıs¸ ve bireysel kiplikleri anlamlı ölçüde geride bırakmıs¸tır. Ayrıca, yalnızca görüntü kipliklerini (B-taramaları ve kalınlık haritaları) kullanan bir eksiltme çalıs¸ması 0.84'lük bir AUC sağlamıştır. Bu sonuç, birles¸tirilmis¸ yapısal verilerin güçlü tamamlayıcı değerini vurgulamaktadır.

Boylamsal duyarlılık analizi ayrıca ağ tabakasına ait biyobelirteçler için bir "tanı ufku" belirlemis¸tir. Kestirim doğruluğunun, klinik tanıdan 4 ila 8 yıl öncesinde en yüksek seviyede olduğunu gözlemledik. Ancak bu sinyaller, 12. yıla gelindiğinde kademeli olarak taban çizgisine yakınsamaktadır. Mevcut literatürle kıyaslandığında, çerçevemiz semptomatik Hafif Bilis¸sel Bozukluk tanısı için mevcut temel çizgilerden daha iyi bir bas¸arım göstermis¸tir. Bu durum, klinik öncesi kestirim gibi çok daha zorlu bir görevde modelin gürbüzlüğünü kanıtlamaktadır. Sonuç olarak bu çalıs¸ma, ağ tabakası görüntülemesinin Alzheimer Hastalığı için erken tanı is¸ hattına tümles¸tirilmesi adına uygulanabilir bir yol olus¸turmaktadır.

Anahtar Kelimeler: Alzheimer Hastalığı, Retinal OKT, Derin Öğrenme, Erken Tahmin, UK Biobank.

ABSTRACT

DEEP LEARNING-BASED ANALYSIS OF RETINAL OCT SCANS FOR DETECTION OF ALZHEIMER'S DISEASE

Alterations in retinal layer thickness have been associated with neurodegenerative diseases such as Alzheimer's disease (AD). These structural changes can be measured using a noninvasive imaging technology called Optical Coherence Tomography (OCT). Previous research has mostly focused on the statistical as-sociations between segmented retinal layer thickness and AD derived from OCT or OCTA devices. Unlike conventional medical image classification tasks, early detection is more challenging than diagnosis because imaging precedes clinical diagnosis by several years. Deep learning (DL), particularly through convolu-tional neural networks (CNNs) and transfer learning, has demonstrated strong performance in image-based disease detection tasks. However, the application of DL directly on unsegmented raw OCT B-scan images for early AD detection remains underexplored. Therefore, in this thesis, we address this research gap by proposing a deep learning-based approach that uses raw OCT images for early Alzheimer's disease detection.

All related studies in the literature have heavily relied on private and in-situational cohorts that lack interoperability. In contrast, the UK Biobank, 2022 offers a unique resource for investigating the associations between retinal struc-ture and systemic health, comprising over 85,000 OCT scans linked to cognitive and health-related data. Between the initial scan (2010–2015) and July 2023, 539 participants in the dataset were diagnosed with AD. Therefore, although the UK Biobank is somewhat limited by the absence of OCTA scans, we used this dataset in our thesis to detect early AD using OCT scans. After a rigorous data-exclusion process, this study used a targeted 4-year window, selecting par-ticipants diagnosed with AD within 4 years of their baseline assessments. The

AD group was matched by age, sex, eye, and instance with a randomly selected balanced Healthy Control group (N = 30).

We first evaluated the predictive value of isolated 2D B-scans using pre-trained deep learning architectures. In these tests, the ResNet-34 model achieved Mean AUC of 0.624 ± 0.060. Saliency map analysis of these B-scans highlighted the critical importance of the central macular region, whereas peripheral areas showed a negligible contribution to the model's decision. To overcome the limitations of isolated B-scans and leverage 3D information, we generated a 3D-informed en-face thickness projection map from the OCT B-scans. This pipeline was optimized to focus on the diagnostically relevant 3mm inner macular region, which effectively filtered out peripheral noise. Our study of thickness maps identified the Ganglion Cell Layer (GCL) as the most significant indicator of preclinical AD. The VGG-19 model, trained on GCL thickness maps with a year-weighted loss function, achieved a peak Mean AUC of 0.750 ± 0.037. Notably, the traditional clinical benchmark, the Retinal Nerve Fiber Layer (RNFL), exhibited negligible predictive value in this pre-symptomatic cohort.

We also developed a Multi-Modal Soft-Voting Ensemble model to further increase the predictive accuracy and emulate clinical decision-making. This model integrates structural insights from B-scans and GCIPL thickness maps with clinical and demographic data. This ensemble approach achieved the highest Mean AUC of 0.85 and significantly outperformed the individual modalities. Furthermore, an ablation study using only image modalities (B-scans and thickness maps) yielded an AUC of 0.84. This result highlights the strong complementary value of combined structural data.

Longitudinal sensitivity analysis also established a "diagnostic horizon" for retinal biomarkers. We observed that predictive accuracy is highest between 4 and 8 years prior to clinical diagnosis. However, these signals progressively converge toward baseline by the 12-year mark. When benchmarked against the current literature, our framework outperformed existing baselines for the diagnosis of symptomatic Mild Cognitive Impairment (MCI). This demonstrates its

robustness in the much more challenging task of preclinical prediction. Consequently, it establishes a viable pathway for integrating retinal imaging into the early diagnostic pipeline for Alzheimer's Disease.
Keywords: Alzheimer's Disease, Retinal OCT, Deep Learning, Early Prediction, UK Biobank.

ACKNOWLEDGEMENT

Firstly, I would like to express my gratitude to my advisor, Assoc. Prof. Dr. Faik Boray Tek for his invaluable guidance and support throughout this thesis.

I would also like to thank my family, friends, and colleagues for their support, encouragement, and guidance throughout this journey.

Finally, I would like to thank TUBITAK, the UK Biobank, and the Küçükaslan family for their financial support.

Yasemin TURKAN

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS LIST

AD Alzheimer's Disease

AMD Age Related Macula Degeneration

CNN Convolutional Neurol Networks

DL Deep Learning

FAZ Foveal Avascular Zone

Grad-CAM Gradient-Weighted Class Activation Map

OCT Optical Coherence Tomography

OCTA Optical Coherence Tomography Angiography

ILM Inner Limiting Membrane

RNFL Retinal Nerve Fiber Layer

GCL Ganglion Cell Layer

IPL Inner Plexiform Layer

INL Inner Nuclear Layer

OPL Outer Plexiform Layer

HFL Henle Fiber Layer

BMEIS Boundary between Myoid and Ellipsoid Zone

IS/OSJ Inner Segment/Outer Segment Junction

OPR Outer Photoreceptor Layer

RPE Retinal Pigment Epithelium

SVP Superficial Vascular Plexus

DVP Deep Vascular Plexus

CHAPTER 1

1. INTRODUCTION

Dementia is a major global health concern, especially affecting older adults. It is the seventh leading cause of death as reported by World Health Organization (WHO, 2022). It is important to note that dementia is not a single disease but rather a broad term, similar to heart disease, that includes many different medical conditions (Dening & Sandilyan, 2015) mainly trouble with memory, language and problem-solving; difficulty concentrating; and struggling to understand and express thoughts.

Alzheimer's disease (AD) is the most common type in dementia with a rate of 60%-80%. Currently, there is no cure for AD. The disease is marked by progressive brain degeneration, caused by abnormal buildups of proteins called amyloid-beta and tau inside brain cells (WHO, 2022). The disease starts almost 20 years before clinical symptoms appear and develops into a condition called Mild Cognitive Impairment (MCI) (Gaugler et al., 2022). Studies on human donors have shown that the same harmful proteins seen in the AD brain also accumulate in the retina (London et al., 2013). For this reason, researchers are now interested in using various eye imaging methods to detect early changes in the retina of AD patients. Optical coherence tomography (OCT) and optical coherence tomography angiography (OCTA) are used to detect structural and vascular changes. Compared to traditional diagnostic tools such as magnetic resonance imaging (MRI), cerebrospinal fluid (CSF) analysis, or genetic testing, OCT is non-invasive, quick, affordable, and already widely used in eye clinics.

Detecting AD early gives patients the chance to plan and start treatment or lifestyle changes that could slow down its progression. Besides, patients and caregivers can plan, and organize support while they are still cognitively stable. Therefore, early detection may lower healthcare costs by delaying the need for intensive care.

However, detecting early signs of AD from retinal imaging is challenging as changes are subtle, temporally misaligned with diagnosis, and confounded by age and ocular conditions. Conventional diagnostic methods from medical images greatly depend on physicians' professional experience and knowledge. Artificial intelligence (AI) has improved the performance of many challenging tasks when working with high-resolution, complex imaging data. Artificial neural networks are a subset of AI inspired by a simplification of neurons and their connections in the brain. Deep learning (DL) is a multi-layer structure of neural networks that mimics human learning by analyzing data with a given logical structure.

1.1 STATEMENT OF THE PROBLEM

Optical Coherence Tomography (OCT) is a non-invasive imaging technology widely used in ophthalmology to capture high-resolution cross-sectional images of the retina. Since the retina and brain share the same embryological origin, retinal changes—particularly in structural layers—are believed to reflect neurodegenerative processes occurring in the brain. Therefore, OCT has recently emerged as a promising tool for AD detection.

Previous research has mostly focused on statistical associations between segmented retinal layer thicknesses and AD derived from OCT or OCTA devices. However, these approaches often depend on pre-processed measurements and lack the flexibility to detect subtle, complex patterns in raw images. Deep learning (DL), particularly through convolutional neural networks (CNNs) and transfer learning, has shown strong performance in image-based disease detection tasks. Yet, the application of DL directly on unsegmented raw OCT B-scan images for early AD prediction remains underexplored.

In this thesis, we address this research gap by proposing a deep learning-based approach that uses raw OCT images for early-stage Alzheimer's detection. Our method includes preprocessing pipelines tailored for retinal anatomy and explores the integration of OCT-derived maps to enhance the learning process. By

leveraging transfer learning on raw and multichannel OCT data, this work aims to establish an effective and scalable framework for preclinical AD prediction using non-invasive retinal imaging.

1.2 SUMMARY OF CONTRIBUTIONS

This study builds on two conference papers and two journal papers published (or under review/preparation) during my doctoral journey. Here, we list the papers and summarize the main contributions of each paper. A detailed list of contributions is provided in the corresponding chapters on related research.

1.Systematic Review – Contributions (Turkan et al., 2024): Turkan, Y., Tek, F. B., Arpacı, F., Arslan, O., Toslak, D., Bulut, M., & Yaman, A. (2024). Automated diagnosis of Alzheimer's Disease using OCT and OCTA: A systematic review.
IEEE Access. https://doi.org/10.1109/access.2024.3434670
We conducted a PRISMA-based systematic review of the literature regard-ing the automated diagnosis of Alzheimer's Disease using both structural OCT and OCT-Angiography (OCTA).
We identified and categorized the most significant retinal biomarkers, highlighting the shift from total macular thickness to specific inner retinal lay-ers and vascular density metrics.
We examined the diagnostic performance (AUC, Accuracy) of various ma-chine learning and deep learning methodologies across diverse private and public datasets.

- We discussed the current challenges in the field, such as dataset heterogeneity and the lack of longitudinal studies, providing a roadmap for future research.

2.B-Scan Deep Learning Study – Contributions (Turkan et al., In Publication): Turkan, Y., Tek, F. B., Nazlı, M. S., & Eren, Ö. (2025). Early Alzheimer's Disease Detection from Retinal OCT Images: A UK Biobank Study. (IWW).

- We introduced the first application of deep learning to raw OCT B-scans for the early prediction of AD (up to 4 years before diagnosis) using the UK Biobank dataset.
- We evaluated multiple state-of-the-art architectures, including ResNet-34 and the OCT-specific foundation model RETFound, establishing a repro-ducible baseline for raw image classification.
- We utilized Grad-CAM explainability analysis to confirm that the model's predictive signal was localized to anatomically relevant layers, specifically the BMEIS and IS/OSJ.

3.**Interpretable ML on Derived Data from OCT Images – Contribu-tions** (Turkan et al., In Prep): Turkan, Y., Kırbıyık, E., & Tek, F. B. (2024). Retinal Biomarkers of Alzheimer's Disease with Interpretable Machine Learn-ing on UK Biobank OCT Data. ACM Transactions on Computing for Healthcare (Under Review).

- We analyzed the predictive power of segmented retinal layer volumes and clinical features from the UK Biobank using an interpretable XGBoost framework.
- We investigated the longitudinal hazard of AD development through Cox proportional hazards models, identifying specific retinal metrics associated with increased risk.
- We employed SHAP (SHapley Additive exPlanations) to rank the impor-tance of retinal features, providing a transparent look at how structural thinning correlates with future cognitive decline.

4.En-face Ganglion Cell Layer (GCL) Thickness Analysis – Contri-butions (Turkan & Tek, In Prep):Turkan, Y., & Tek, F. B. Deep Learning-Based Early AD Prediction Using 3D-Informed En-face GCL Thickness Maps. (In Preparation / In Progress).

- We proposed a novel 3D-informed en-face thickness mapping pipeline that projects volumetric OCT data into 2D maps, significantly improving

com-putational efficiency while preserving spatial context.

- We identified the Ganglion Cell Layer (GCL) as the superior longitudinal biomarker for AD prediction, outperforming the traditionally utilized RNFL over a 12-year diagnostic window.
- We achieved a state-of-the-art AUC of 0.750 using a VGG-19 architecture, demonstrating the effectiveness of CNN-based inductive biases for specialized thickness maps compared to Vision Transformers.

1.3. ORGANIZATION

This thesis is organized into nine chapters, providing a logical progression from the theoretical background to data analysis, model development, and final validation. The content of each chapter is summarized as follows:

Chapter 1: Introduction presents the motivation for the study, defines the research problem regarding the early detection of Alzheimer's Disease (AD) using retinal imaging, and summarizes the main contributions of this thesis.

Chapter 2: Background provides the necessary medical and technical con-text. It covers the principles of Optical Coherence Tomography (OCT) and OCT-Angiography (OCTA), followed by a comprehensive literature survey on deep learning applications for AD detection. This chapter also categorizes existing studies based on dataset curation, model training, and validation strategies.

Chapter 3: UK Biobank Dataset details the source of the data used in this research. It describes the rigorous exclusion criteria, quality control measures, and the statistical analysis of the study population, defining the specific cohorts used for training and testing.

Chapter 4: Early AD Prediction in UK Biobank Dataset focuses on the analysis of tabular data. It evaluates the predictive power of retinal and sys-temic risk factors using gradient-boosted decision trees (XGBoost) and survival analysis (Cox Proportional Hazards). This chapter also utilizes SHAP (SHapley Additive exPlanations) to identify key biomarkers and interprets their impact on

AD risk.

Chapter 5: Early AD Detection from Retinal OCT B-Scans introduces the first deep learning approach utilizing raw OCT B-scans. It details the preprocessing pipeline, the matching strategy for the study cohort, and the training of Convolutional Neural Networks (e.g., ResNet-34) to detect early AD signs. It also presents interpretability results using saliency maps to localize diagnostically relevant retinal regions.

Chapter 6: Early AD Detection from Retinal OCT C-Scans advances the analysis from 2D slices to 3D-informed representations. It describes the generation of en-face thickness projection maps and evaluates the performance of VGG-19 and Swin Transformer architectures. This chapter highlights the superior predictive value of the Ganglion Cell Layer (GCL) compared to the Retinal Nerve Fiber Layer (RNFL).

Chapter 7: Multi-Modal Ensemble via Soft Voting presents a unified frame-work that combines the strengths of the previously developed models. It details the methodology for a Soft-Voting Ensemble that integrates structural insights from B-scans, morphological data from thickness maps, and clinical risk factors to achieve the highest predictive accuracy.

Chapter 8: Discussion synthesizes the findings from all chapters and com-pares the performance of the proposed models against the current state of the art. It discusses the clinical implications of the "diagnostic horizon" and the potential of retinal biomarkers for screening.

Chapter 9: Conclusion summarizes the thesis’s key outcomes and suggests directions for future research.

CHAPTER 2

2. BACKGROUND

2.1 MEDICAL BACKGROUND

The retina and optic nerve grow from the same neural tube during embryonic development (Blazes & Lee, 2021). Figure 2.1. shows the details of the retinal layers. Only photoreceptors are sensitive to light. When light reaches the retina, photoreceptors are triggered, and the signal is transmitted through bipolar cells to ganglion cells. These ganglion cells generate action potentials in response to incoming signals. Other retinal neurons are indirectly affected by light via various synaptic connections. Ganglion cells transmit visual information to the brain via the optic nerve (Bear et al., 2016).

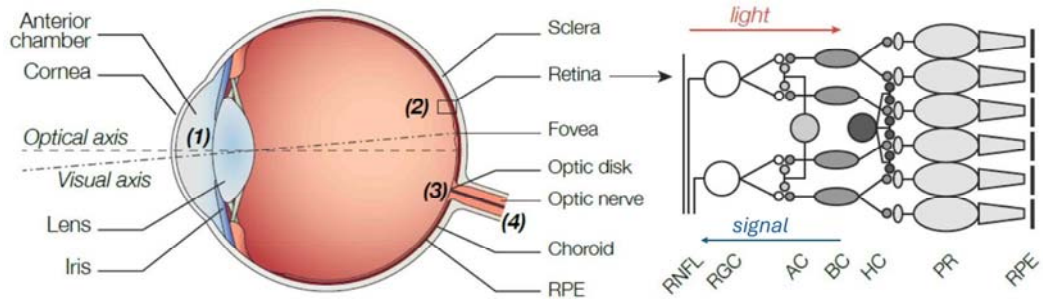


Figure 2.1: Diagram of the eye, the retina, and location of the various retinal implants. Retinal layers, from bottom to top: retinal pigment epithelium (RPE), photoreceptors (PR), horizontal cells (HC), bipolar cells (BC), amacrine cells (AC), ganglion cells (RGC), nerve fiber layer (RNFL). (Wikipedia, 2025)

The cardiovascular system and inner blood-retinal barrier can be directly observed in the eye (London et al., 2013). Therefore, retinal imaging is not only useful for examining eye conditions but is also employed in the analysis of systemic diseases such as cardiovascular disorders (Wagner et al., 2020). Compared

to the brain, the eye is more easily accessible, making imaging techniques more practical, generally less invasive, and more affordable than methods used for brain imaging, such as MRI, PET, or CT scans (Cunha et al., 2022; Song et al., 2021).

Neurodegeneration in Alzheimer's disease (AD) affects not only brain neurons but also retinal neurons. Both animal studies (Carelli et al., 2017; Gardner et al., 2020; Hadoux et al., 2019) and human research (den Haan et al., 2018) have clearly demonstrated the accumulation of retinal Aβ and neurofibrillary tangles (NFTs) in the retina of the eye. Therefore, ophthalmic imaging techniques have become popular for investigating Alzheimer's disease and related disorders. Table 2.1 summarizes the various imaging methods used for these diagnoses.

Table 2.1 Overview of Retinal Imaging Modalities Used in Alzheimer's Disease Research

Modality Type	**Examples / Techniques**
Structural	Fundus imaging, Optical Coherence Tomography (OCT), Autofluorescence imaging (Attiku et al., 2021), Widefield autofluorescence (Alber et al., 2020)
Vascular	Fundus fluorescence angiography, Optical Coherence Tomography Angiography (OCTA) (Attiku et al., 2021)
Functional	Electroretinogram (ERG) (Tsang & Sharma, 2018), Color and contrast sensitivity tests (London et al., 2013)

2.1.1 Optical coherence tomography (OCT) :

OCT is a non-invasive method used to obtain views of retinal structures in two-dimensional (2D), cross-sectional, and three-dimensional (3D) volumetric images (Snyder et al., 2020). It provides extensive information on retinal morphology and assists in diagnosing many diseases. It is often described as "optical ultrasound." While ultrasound uses sound waves to image tissue, OCT uses light. Because light travels much faster than sound, it is impossible to measure the echo time directly using standard electronics. Instead, OCT relies on

a technique called low-coherence interferometry (Aumann et al., 2019). Figure 2.2. shows how the OCT system works.

A low-coherence light source emits a beam of near-infrared light. Near infrared light is used because it penetrates biological tissue effectively while minimizing absorption. The light beam is directed into a beam splitter, which divides the light into two distinct paths: the Reference Beam (Directs to a mirror at a known distance) and the Sample Beam (Directs into the tissue or material being imaged). The light reflects off the reference mirror and keeps "echoing" back from different depths within the sample tissue. These two reflected beams meet back at the splitter. Interference (a readable signal) occurs only when the distance traveled by the light in the sample arm matches the distance in the reference arm to within a tiny fraction of a millimeter (the coherence length). By analyzing these interference patterns, the system can calculate the exact depth of the reflecting structure (Mokhtari et al., 2025).

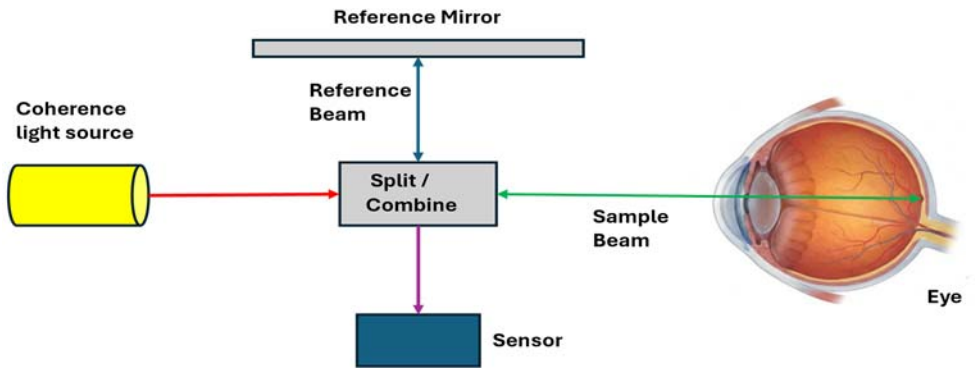


Figure 2.2: Schematic of an Optical Coherence Tomography (OCT) setup for retinal imaging.

A low-coherence light source emits a beam that is split into a reference arm and a sample arm. The reference beam reflects off a reference mirror, while the sample beam is directed into the eye to capture retinal structures. The reflected signals from both arms are recombined and detected by a sensor to generate depth-resolved images based on interferometric principles.

The raw data from the interferometer is processed to generate distinct types of outputs used for analysis (Mokhtari et al., 2025).

- A-Scan (Amplitude Scan): This is a one-dimensional graph representing the reflectivity of tissue at a single point in depth. It is effectively a "drill core" of data showing layers beneath a single pixel.
- B-Scan (Cross-Sectional Image): By combining a series of A-scans while moving the beam laterally across the sample, the system creates a two-dimensional cross-sectional image. This is the standard "slice" view most clinicians use, resembling a histology slide.
- C-Scan and 3D Volumetric Imaging: By acquiring multiple consecutive B-scans, the system can reconstruct a 3D volume of the tissue. This allows for "En-face" imaging, where the user can view the tissue from the front (like a photograph) at specific depths beneath the surface.

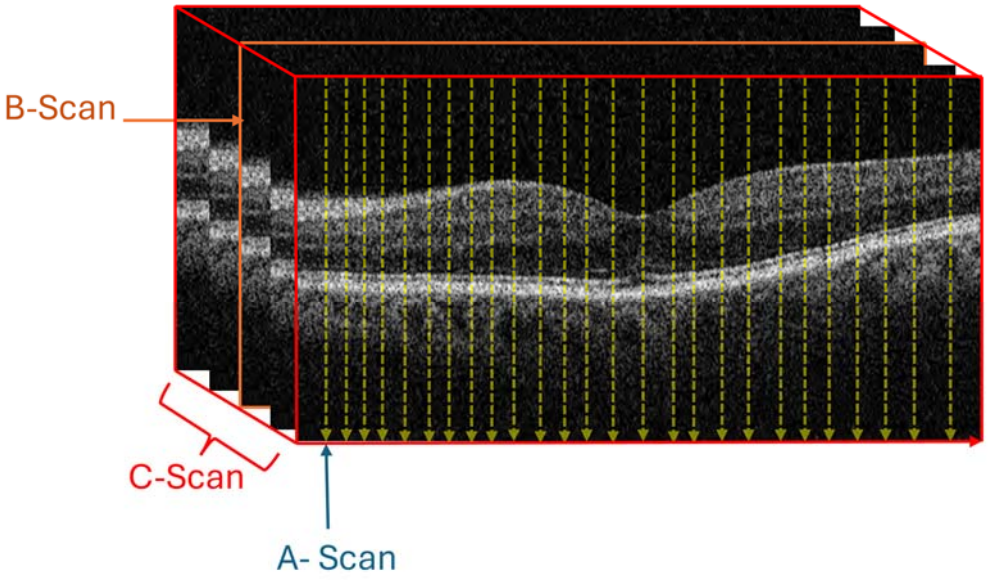


Figure 2.3 Visualization of OCT data acquisition hierarchy. The blue arrow indicates an A-scan, representing a single vertical depth scan of the retinal tissue. Multiple adjacent A-scans (yellow dashed arrows) differ laterally to form a B-scan (orange arrow), providing a cross-sectional view. A series of B-scans are compiled to generate a volumetric C-scan (red arrow), allowing for

3D reconstruction and en face analysis.

The processing power of modern OCT softwares move beyond simple visualization to advanced quantitative analysis. This relies on the automated identification of specific retinal boundaries, a process known as segmentation. The retinal layers are depicted in Figure 2.4 (c). Once the layers are segmented, the software calculates the distance between specific boundaries to determine thickness in microns (μm) such as RNFL layer thickness as shown in Figure 2.5. To standardize the analysis, OCT softwares map the thickness data onto the ETDRS (Early Treatment Diabetic Retinopathy Study) Grid (Early Treatment Diabetic Retinopathy Study Research Group, 1991). Besides the measured thickness value is overlaid on a normative database to compare them with population norms, and track changes over time.

In recent years, there has been an increase in research investigating the use of OCT to evaluate AD. These studies examine a range of novel parameters observed by OCT. Song et al. (2021) conducted an extensive literature review of diagnosing Alzheimer's Disease using OCT imaging modalities. Most commonly, researchers rely on the numerical measurements generated by OCT software—such as layer thickness and volume—rather than analyzing the raw A-, B-, or C-scans directly. The following are the significant findings in OCT in the literature:

- Thinning of the retina has been highlighted in the majority of the studies. Several neurodegenerative diseases such as AD, dementia, and Parkin-son's Disease, have a characteristic signature of Retinal Nerve Fiber Layer (RNFL) thinning. Figure 2.1 shows the layers of the retina. Retinal thin-ning often refers to the inner layers such as the RNFL (also called Stratum opticum), Ganglion Cell Layer, and Inner Plexiform Layer. These three layers are called ganglion cell complex. Ganglion cells are vulnerable to neurodegeneration due to mitochondrial dysfunction and their unique ar-chitecture of axons without myelin (Carelli et al., 2017). Figure 2.5 shows an OCT output for a patient,

showing the RNFL thickness values in both eyes. OD is the left eye, and OS is the right eye. The eye has a spherical shape. Therefore, 2D slice visualization requires a circular region to be mapped to a rectangular region.

- Reduced macular volume and thickness are the research area's second es-sential parameters. Macula provides sharp, clear, straight-ahead vision. It is responsible for central and color vision (Bear et al., 2016). Figure 2.5 shows the macula region in the eye. The middle of the retina images, along with their graphs, show the thickness of the macula.
- Some studies show that reduced choroidal volume and thickness are im-portant parameters, while others disagree with this finding (Song et al., 2021). The choroid layer is the supplier of nutrients to the retina, and it also maintains the temperature and volume of the eye (Bear et al., 2016).

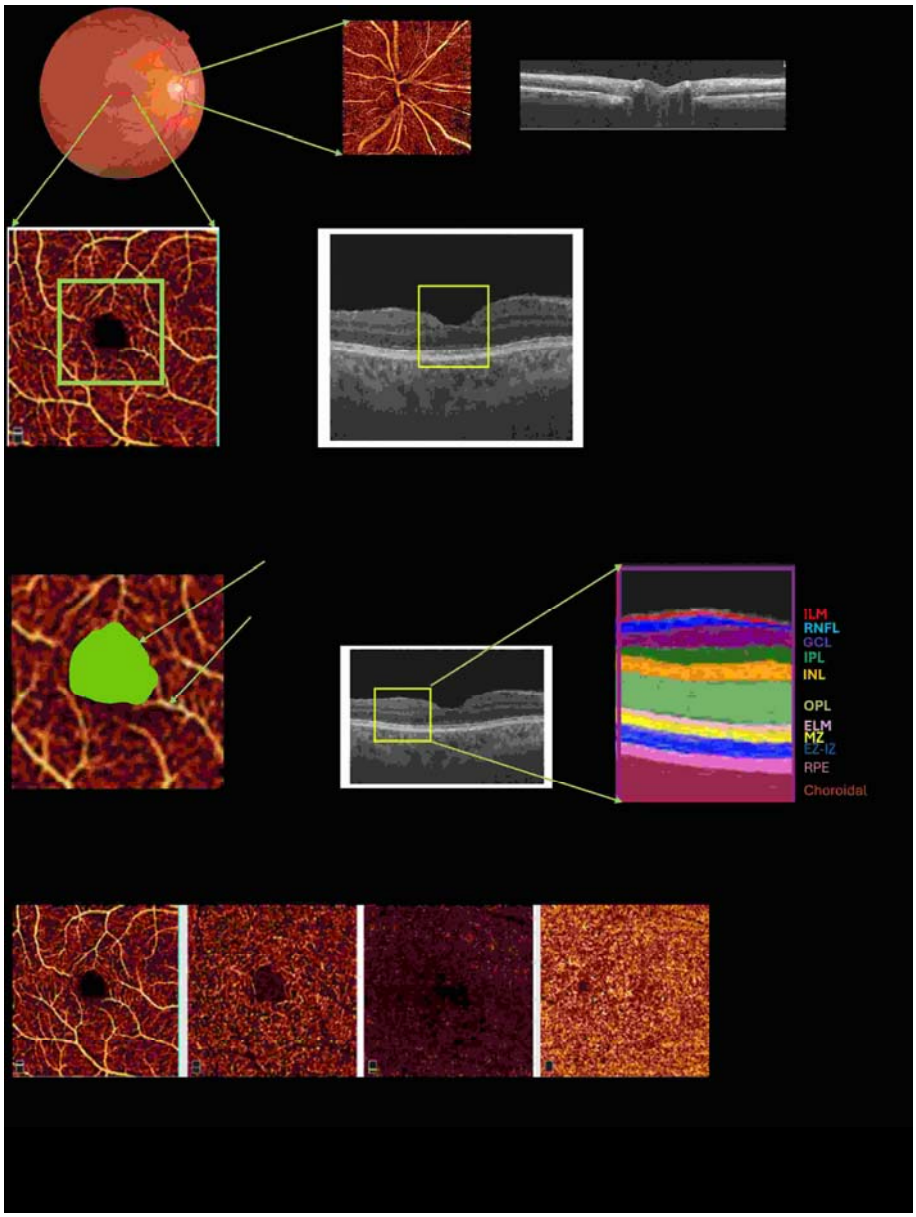


Figure 2.4: OCT and OCTA Modalities in detail. (a) Linkage of a fundus and its disk area shown by OCTA and OCT. (b) Enface OCTA image and one of the OCT slices of the macula area. (c) Enface OCTA image showing the Foveal Avascular Area (FAZ) in the middle with retinal blood and peripapillary vessels. The retinal layers are colored yellow and explained in detail. (d) OCTA enface images of four layers of the retina. Superficial Vascular Plexus (SVP): the first layers of the retina (ILM, RNFL, GCL, and IPL). Deep Vascular Plexus (DVP): deeper layers of retina (IPL,INL, OPL). The outer avascular layers of the retina.

Choriocapillaris: bottom layers of the retina (Choroidal layer in OCT).

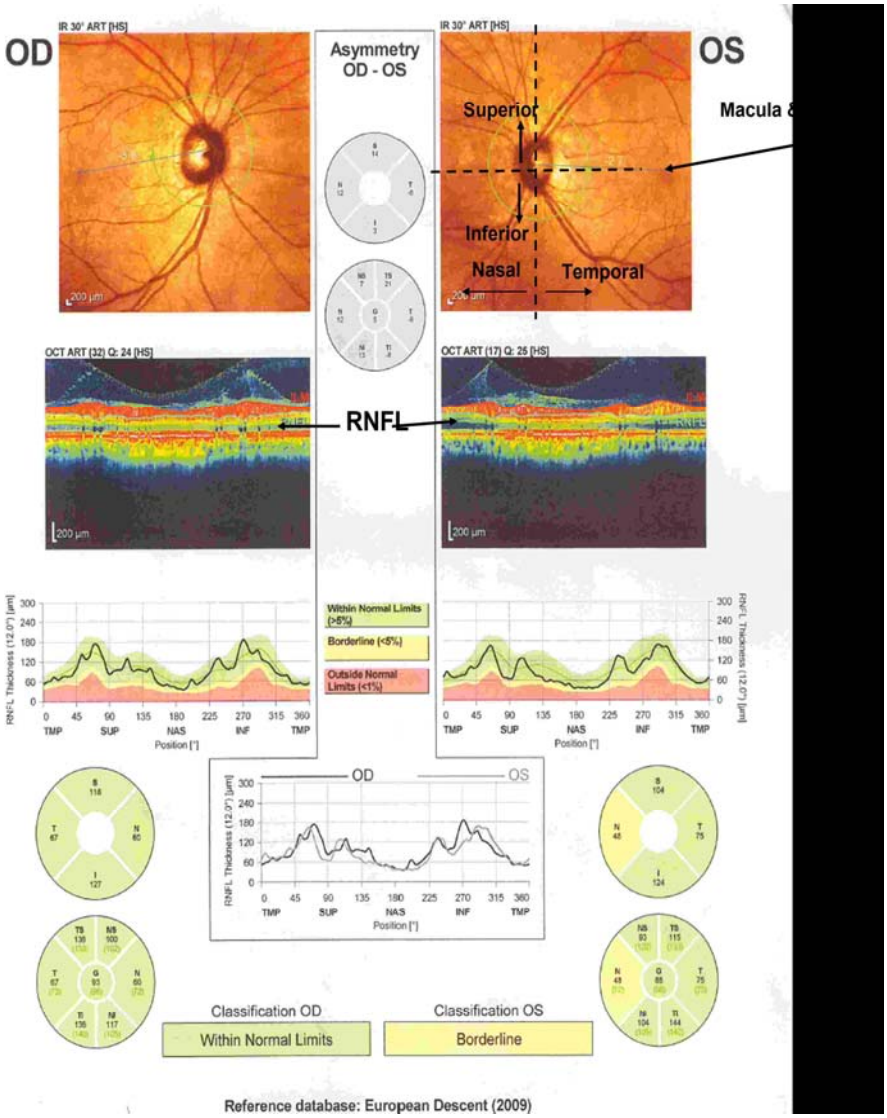


Figure 2.5: OCT output that shows the Retinal Fiber Layer Thickness decrease in the patient. OD is the right eye, and OS is the left eye. The eye has a spherical shape. The vertical area closer to the nose is called the Nasal Retina (NAS, N), and the other side is the Temporal Retina (TMP, T). The upper retina from the fovea is Superior (SUP, S), and the lower part is inferior (INF, I).

Therefore, 2D visualization requires unfolding the circle to a rectangle (from TMP to SUP, then to NAS, and back to TMP). The Retinal Nerve Fiber Layer (RNFL) layer shown in the middle image is automatically measured by

the OCT system and compared with the European Descent reference database. The results show borderline thinning of the left eye.

2.1.2 Optical Coherence Tomography Angiography (OCTA) :

OCTA is a relatively new technique, developed in 2015. It utilizes OCT devices to compute the differences between two consecutive OCT scans. These differences reveal the motion of blood within the vessels (Figure 2.4). Unfortunately, OCTA is not supported by all OCT devices. Although, it is an effective imaging method for diagnosing eye diseases. It identifies microvascular changes and abnormalities in the blood flow patterns. OCTA images are still susceptible to several distortions common in OCT, such as projection errors, motion artifacts, segmentation issues, and signal loss (Attiku et al., 2021).

In their 2021 review, (Song et al., 2021) summarized the results of several studies on OCT and OCTA. Their major findings were:

- Reduced vessel density is reported in several studies. Figure 2.4 shows the blood vessels of the retina from a patient (Ge et al., 2021).
- Reduced perfusion density is also observed together with the reduced ves-sel density. Vessel perfusion density is defined as the total area of perfused vasculature per unit area (Triolo et al., 2017).
- Increased Foveal Avascular Zone (FAZ) is another parameter reported by some studies (Song et al., 2021). The fovea is visible in Figure 2.4 as the dark spot at the center, which marks the center of the retina (Bear et al., 2016).
- Reduced peripapillary vessel density is discussed in some of the recent works. The radial peripapillary capillaries are a distinctive vascular network within the RNFL around the optic disc.

2.2 LITERATURE SURVEY

Deep learning models require substantial amounts of data for generalization. Yanagihara et al. (2020) outlined even more challenges in DL models such as lack of standardized image collection, evaluation metrics, and computational resources. However, in recent years, DL models have been favored over traditional ML counterparts in OCT and OCTA applications. In our review, we organized the analysis of existing studies according to the framework of Deep Learning (DL) workflow for diagnosing Alzheimer's disease (AD) and Mild Cognitive Impairment (MCI) using OCT and OCTA data. This framework is illustrated in Figure 2.6 (Turkan et al., 2024). We grouped the studies based on the essential components of the DL pipeline to answer questions regarding how the datasets were prepared, how the models were trained, and what validation strategies were applied.

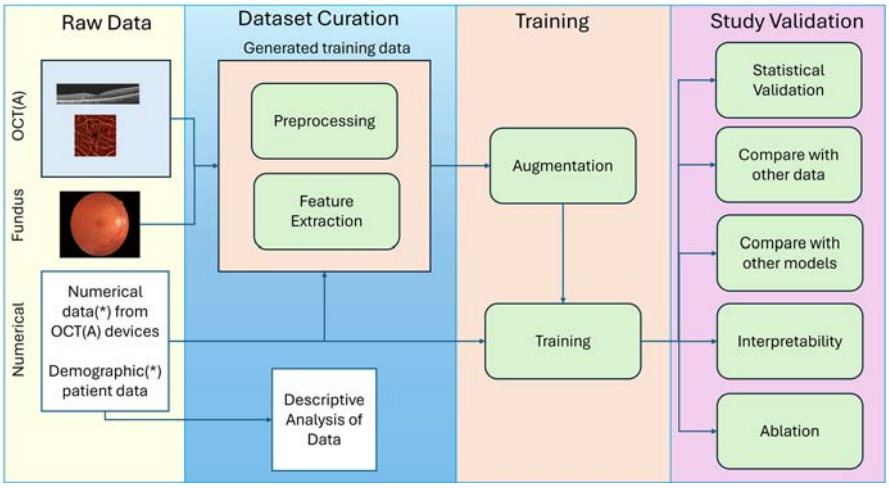


Figure 2.6: The framework of Deep Learning-driven flow for AD/MCI diagnosis in OCT/OCTA Studies showing common processes in dataset curation, training, and validation phases.

To review ML/DL-based approaches for AD or MCI diagnosis in OCT and/or OCTA scans, we followed the guidelines of the Preferred Reporting Items

for Systematic Review and Meta-Analysis (PRISMA) (Page et al., 2021).

Information Sources: We performed an exhaustive search of PubMed, Web of Science, Scopus, Google Scholar, Semantic Scholar, and CrossRef databases for relevant studies published between 2015-2025. Figure 2.7 shows the PRISMA flowchart of the systematic review.

Search Strategy: We surveyed the databases above using the following combinations of terms:

("Alzheimer's" OR "dementia" OR "cognitive impairment") AND ("optical co-herence tomography" OR "optical coherence tomography angiography" OR "reti-nal imaging") AND ("neural networks" OR "machine learning" OR "deep learn-ing")

Eligibility criteria: Articles were selected for analysis if :

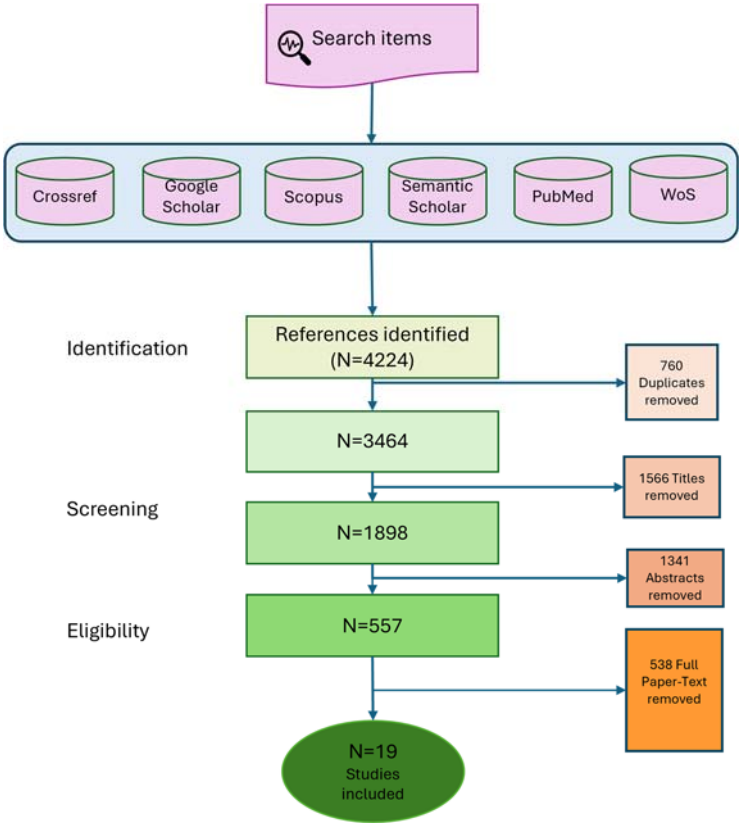


Figure 2.7: Flow diagram of eligible study selection from a search across PubMed, Web of Science, Scopus, Google Scholar, Semantic Scholar, and CrossRef databases yielded 4224 deep learning (DL) publications related to OCT and OCTA applications between 2015 to 2025 (End of Oct. 2025)

- The article was available in English.
- It was published as a primary research paper in a peer-reviewed journal or a conference papers. Duplicates, datasets, book chapters, and articles that provided only statistical analysis were excluded.
- It described an ML/DL model for AD detection, screening, or prediction using only OCT/OCTA scan images or data derived from these images. Articles related only to segmentation or image quality improvements were excluded.
- It focused solely on AD and/or MCI (not other diseases such as age-related macular degeneration, drusen, etc.).

Data extraction, quality assessment, and bias analysis: We searched the databases (PubMed, Web of Science, Scopus, Google Scholar, Semantic Scholar, and CrossRef) to identify studies that matched our search strategy. We then cross-checked and identified the studies that met our eligibility criteria. For eligible studies, a detailed analysis was performed to obtain critical information, including the number of participants, year of publication, algorithms applied and their characteristics, model prediction parameters, and performance metrics (such as accuracy, discrimination, sensitivity, and specificity rates).

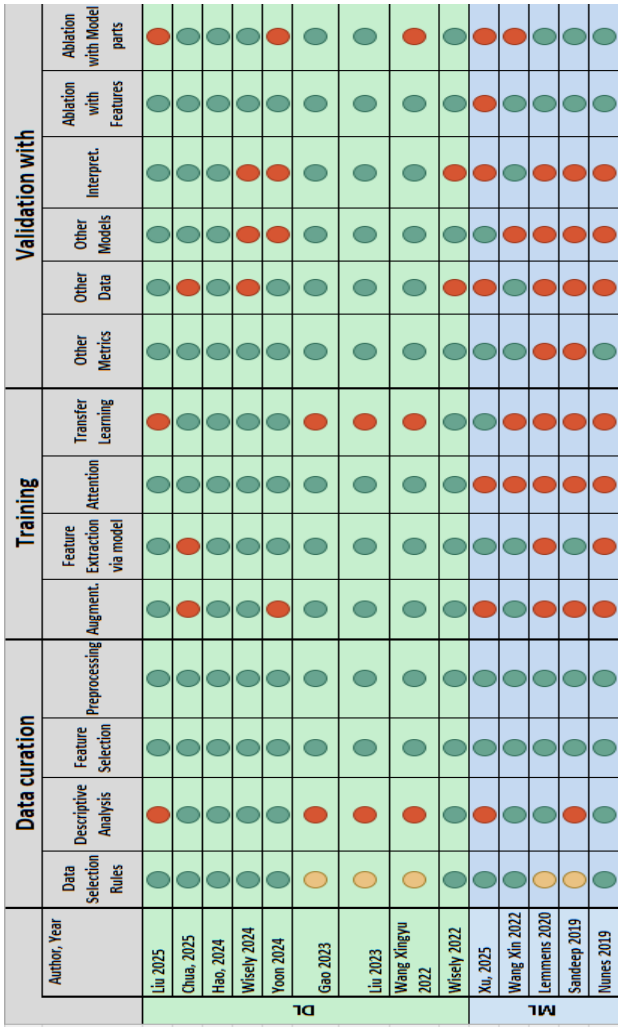
Our search initially retrieved 4224 references. After applying the elimination steps shown in Figure 2.7, 19 studies met the inclusion criteria. Two of these were previously identified by Bourkhime et al. (2022). In our survey, we identified 5 animal (mouse) and 14 human studies. These investigations were longitudinal studies involving OCT and Fundus imaging of mice aged 1-16 months. Because human studies include extensive inclusion and exclusion criteria, additional checks are necessary to ensure the sample dataset accurately represents the total population. Due to differences in design between human and mouse studies, we excluded mouse studies from this review.

Table 2.2 visually summarizes our key findings on how each study conforms to the framework for ML and DL studies. Based on this unique frame-work, we discuss our findings from the perspectives of the datasets, training, and validation.

2.2.1 Dataset Curation

Researchers primarily rely on local datasets collected from a limited number of patients in clinical settings because of the absence of publicly available

Table 2.2: High-level summary showing how the studies comply with the framework details explained in the deep learning-driven flow for AD/MCI diagnosis in OCT/OCTA. Green: the full process is observed in the study; red: the process is not mentioned or observed in the paper. Amber: no evidence of any exclusion criteria is observed even with inclusion criteria explained in the paper.



OCT and OCTA datasets labeled for Alzheimer's disease (AD) or Mild Cog-nitive Impairment (MCI). Most datasets were relatively small, and only a few studies included clearly labeled AD data. This limitation makes it difficult to build and evaluate reliable deep learning models. S. Liu et al. (2025) in their improved version of their previous study (S. Liu et al., 2023) reused the ROAD and ROMSI datasets used in the study by Hao et al. (2024). J. J. Xu et al. (2022) curated a dataset of 2000 images from other studies with various disease types, such as stroke, AD, diabetes, age-related macular degeneration, and healthy co-horts. The UK Biobank (2022) has become an important data source in this regard, providing researchers with access to a large collection of retinal images for a small fee. We observed only AD detection from Fundus image studies (R. Liu et al., 2025; Tian et al., 2021; Yousefzadeh et al., 2024), although the dataset also contained OCT scans. The dataset did not contain OCTA images.

There has been growing interest in the early diagnosis of both Alzheimer's disease (AD) and Mild Cognitive Impairment (MCI). Reflecting this shift, all studies published after 2023 have begun to include patients with MCI in their analyses (J. Chua et al., 2025; Hao et al., 2024; S. Liu et al., 2023, 2025; Wisely et al., 2024). We also observed that most of these studies focused on OCTA datasets (Hao et al., 2024; S. Liu et al., 2023, 2025; Yoon et al., 2024).

We observed a lack of standardization in data collection practices across the studies reviewed. The data inclusion and exclusion criteria differed signif-icantly among the studies. No reporting was observed in some studies, as de-picted in orange in Figure 2.2. Numerous studies have enhanced their analyses by incorporating demographic information to mitigate biases arising from small dataset sizes.

Deep learning models require significantly large dataset sizes to capture the subtle and complex features associated with AD and MCI detection. In many cases, researchers were required to discard additional scans owing to issues such as poor signal quality, image noise, or motion artifacts. Several studies have enriched their models with supplementary information to overcome these limita-tions. They added more features, such as patient demographics and quantitative

measurements extracted from OCT and/or OCTA scans, such as Retinal Nerve Fiber Layer (RNFL) thickness.

Earlier studies (Lemmens & et. al, 2020; Nunes et al., 2019; Sandeep et al., 2019) predominantly used ML models. Similar to animal studies, Nunes et al. (2019) generated Mean Value Fundus (MVF) (Bernardes et al., 2017) images from each layer, followed by the computation of a feature vector based on the gray-level co-occurrence matrix (GLCM) (Haralick et al., 1973). In later studies, DL or ML models were highly observed to be used in feature extraction, even in ML studies. A recent study by C. Xu (2025) pretrained DL model (nnU-Net) to extract vascular structure segmentation while still using GLCM features as earlier studies. They then applied various ML models for classification.

Wisely et al. (2022, 2024) explored different combinations of input data and found that the most effective single input was the Ganglion Cell–Inner Plexiform Layer (GC-IPL) thickness image derived from OCT. While combining multiple data sources, such as OCT and OCTA images, quantitative metrics, and patient data, led to only a modest performance improvement, the highest accuracy was achieved using only the retinal layer thickness measurements obtained from OCT devices. In a different approach, Yoon et al. (2024) improved model performance by integrating quantitative radiomic features and patient demographics with OCTA imaging.

S. Liu et al. (2023, 2025) adopted a different approach, using polar-transformed the superficial vascular (SVC) and deep vascular complexes (DVC) images. In their recent study (S. Liu et al., 2025) they improved their DL models called Special Extraction Module (SEM), Multi-view Model (MVM), Regional Rela-tionship Model (RRM), Polar Regional importance module (PRIM), and BiL-STM3D Model to extract more features before the classification. J. Chua et al. (2025) calculated OCT projection thickness maps to train their models, along with derived numerical features from the Fundus image.

Hao et al. (2024) and Yoon et al. (2024) used automated segmentation of the FAZ. Yoon et al. (2024) derived "randomic features" of the FAZ area in their

study, and Hao et al. (2024) added microvasculature segmentation as well to the features of their training.

2.2.2 Models and Training

Our analysis revealed a shift from classical ML pipelines to fully automated DL models for AD detection using OCT/OCTA images. We observed a shift from numerical features (i.e., layer thickness measurements and textural metrics) to the direct processing of images as inputs. This transition was most commonly observed in recent studies based on DL architectures. The latest ML study by C. Xu (2025) used a pretrained DL segmentation model for feature extraction. Furthermore, we observed a correlation between the increasing complexity of the models and the incorporation of advanced techniques. Attention mechanisms were applied in all the latest studies, except for that by Yoon et al. (2024). Transformers were used directly in the proposed models; Hao et al. (2024), Gao et al. (2023), and S. Liu et al. (2023) compared their model's classification performance with well-known transformer architectures. These models are considered to be complex black boxes. Therefore, in addition to interpretability analysis, more tools are required to test the robustness of the models. All DL studies used various ablation mechanisms to test the contribution of features and/or parts to their models' performance.

We also reviewed how these studies addressed the challenges of applying data-hungry DL models to OCT images on small datasets. To address this issue, most researchers have applied techniques such as data augmentation and transfer learning, as shown in Table 2.2. In the studies, OCT images with quality and standardization issues were mostly manually removed from the datasets. Studies used various image transformation (J. Chua et al., 2025; S. Liu et al., 2025; Sandeep et al., 2019; Wang, Li, et al., 2022), enhancement (Gao et al., 2023), and artifact removal techniques (Gao et al., 2023; Sandeep et al., 2019).

All reviewed DL studies, except those by S. Liu et al. (2023) and Wang, Li, et al. (2022) utilized transfer learning from generic networks pre-trained on ImageNet-like databases. However, OCT scan slice images differ significantly

from images in general databases because the pixel values represent the measured depth of the retinal tissue at a resolution of a few micrometers, which may limit the benefits of transfer learning. Additionally, because pretrained networks are often trained on generic RGB color image datasets, their application requires alignment with the input format. This requirement confines the analysis to a fixed-size single 2D image input, thereby limiting the utilization of the entire three-dimensional (3D) retinal volume. New, fine-tuned foundational models for OCT are now available. However, they imposed the same limitations because they were fine-tuned on generic 2D ImageNet pretrained foundation models. RETFound (Zhou et al., 2023), for instance, was trained on a substantial collection of 736 K OCT images from various datasets aimed at both segmentation and classification tasks. However, none of the reviewed DL studies used RETFound. Besides, OCTA-NET (Ma et al., 2021) and FAZ-NET (Hao et al., 2022), were developed using the ROSE dataset (Ma et al., 2021) for microvasculature and FAZ segmentation.

The inconsistencies between the metrics and datasets used in different studies made it impossible to compare and validate the classification accuracies unless we observed a repeated study or dataset sharing. The improved study by Lie et (S. Liu et al., 2025) showed that the AUC and Acc improved from 0.85 and 0.85 to 0.88 and 0.89, respectively, compared with the previous study (S. Liu et al., 2023). However, we observed no improvement in the study by Hao et al. (2024), who used the same dataset. We observed that all reviewed recent DL studies used AUC metrics in addition to other metrics such as accuracy, Kappa, Sensitivity, Specificity, and F1-Score. Wisely et al. (2022, 2024) and Lemmens and et. al (2020) preferred to use a single metric (AUC). Most studies use additional features (image-based or quantitative) to increase performance. All studies, except Wang, Jiao, et al. (2022) and Lemmens and et. al (2020) used both eyes (left and right) to double the dataset size. Nunes et al. (2019) found that the classification accuracy improved from 82% to 96% when both eyes received the same classification. Some studies, particularly those of Wang, Li, et al. (2022), Gao et al. (2023), Hao et al. (2024) and S. Liu et al. (2025), demonstrated a

more comprehensive benchmarking approach using external datasets, different models, interpretability, and ablation studies.

2.2.3 Validation

Another common practice observed in these studies was the use of a 5-fold cross-validation technique to calculate training results. In their last work, only Wisely et al. (2024) ran their models 10 repetitions and used the median performance to calculate their results. All studies used additional metrics apart from accuracy. Later studies after 2023 used standard metrics: Accuracy, Area Under Curve (AUC), Sensitivity, and Specificity. Wisely et al. (2024) and Yoon et al. (2024) added confidence calculations (p-values) to their results.

Ablation methodologies, such as testing on other datasets, reducing features, and reducing model components to verify the performance of the proposed models, have become a new trend. It is difficult to collect AD-specific data in isolation; therefore, S. Liu et al. (2023) and Wang, Li, et al. (2022) used the public OCTA-500 dataset (Li et al., n.d.), whereas Nunes et al. (2019) included Parkinson's disease in their original dataset. Hao et al. (2024) and Gao et al. (2023) used additional AD-specific datasets to verify their results. They observed relatively good results with transfer learning and fine-tuning in other domains. We observed that Hao et al. (2024) was the only study that followed all the steps of the deep learning framework.

The most frequently employed method for interpretability was Grad-CAM across all DL studies. Singh et al. (2021) evaluated 13 deep learning explainability methods for OCT scans and found that the Deep Taylor method outperformed others in diagnosing choroidal neovascularization (CNV), diabetic macular edema (DME), and Drusen. Interestingly, none of the AD-related studies in our review applied this method to explain their classification decisions. S. Liu et al. (2023) divided the image into Early Treatment of Diabetic Retinopathy Study (ETDRS) grids. They analyzed the importance of each grid for diagnosis. In a later study (S. Liu et al., 2025) they improved their explainability methods by adding regional relationship analysis. Hao et al. (2024) stood out as they ap-

plied an additional interpretability technique called importance maps. Hao et al. (2024) extracted eight parameters from the images characterizing both the retinal microvasculature and foveal avascular zone (FAZ), such as vascular length density and vascular area density. Then, they conducted two additional interpretability analyses to show the importance of these parameters in diagnosis, first at the image level with different OCTA layers and second at the region level on en face images.

S. Liu et al. (2023) and Hao et al. (2024) were notable for focusing on identifying novel biomarkers rather than merely classifying data. S. Liu et al. (2023) discovered that, in diagnosing Alzheimer's disease (AD), the choriocapillaris (CC) layer is more crucial than the deep vascular complex (DVC) layer in (OCTA), with the parafoveal region being the most critical part of the retina. In contrast, Hao et al. (2024) determined that the DVC is the most significant layer for distinguishing AD from mild cognitive impairments. They found that five out of eight parameters in the DVC showed significant differences between early onset AD and controls, whereas only two parameters were significant between mild cognitive impairment and the controls. These studies also demonstrated that utilizing more tailored interpretability techniques can enhance the identification of new biomarkers, as these tools can reveal additional information. All eligible studies compared their explainability results with those of the medical biomarkers discussed in Section 2.1.

2.2.4 Conclusion

A clear trend shift observed in recent studies is the inclusion of MCI co-horts in diagnostic study designs. This focus is vital because MCI represents the transitional stage in which interventions might be most effective.

Wisely et al. (2024) and Gao et al. (2023) specifically targeted MCI, while J. Chua et al. (2025), Hao et al. (2024), and S. Liu et al. (2025) analyzed both MCI and AD. cohorts. In their studies, they revealed a consistent performance disparity, wherein the AUC scores for MCI detection were invariably lower than those for AD. For instance, Hao et al. (2024) and (S. Liu et

al., 2025), both utiliz-ing the private ROAD and ROMCI datasets from Chinese institutions, reported significantly higher accuracy for early onset AD (AUCs of 0.90–0.93) than for MCI (AUCs of 0.80–0.88). This reduction in accuracy underscores the inherent challenge of identifying subtle, early stage pathological changes characteristic of MCI compared to the more pronounced neurodegeneration seen in established AD.

In terms of imaging modalities, OCT and OCTA have emerged as the dominant tools for non-invasive screening. Hao et al. (2024) and S. Liu et al. (2025) focused on single-modality OCTA. J. Chua et al. (2025) successfully leveraged to analyze structural thickness maps of the RNFL and GCIPL alongside quantitative anatomical parameters across multi-ethnic cohorts. Wisely et al. (2024) demonstrated the value of a comprehensive multimodal approach, combining structural GC-IPL maps and OCTA blood flow data with quantitative retinal metrics (e.g., vessel density) and demographic variables (age, sex, and education) to achieve an AUC of 0.809. Similarly Gao et al. (2023) combined fundus images with OCT B-Scans via an attention based they proposed. All multi-model studies reported higher classification performance than single-modality studies.

The inclusion and exclusion criteria for cohort selection remain inconsistent, particularly regarding age matching. While Gao et al. (2023) explicitly utilized an age-matched design, other studies used datasets which displayed significant age covariances ($p < 0.001$) that complicated comparison. J. Chua et al. (2025) showed that their performance dropped when they retrained their model with age-matched datasets in Asian population.

Finally, the most significant limitation of the studies reviewed in the survey was the inaccessibility of public data. Their research heavily relied on private and institutional cohorts that lack interoperability. For instance, Hao et al. (2024) and S. Liu et al. (2025) utilized specific hospital-based collections (ROAD/ROMCI), Wisely et al. (2024) relied on a private cohort from Duke University, and J. Chua et al. (2025) aggregated disparate datasets from Singapore and Romania. In contrast, the UK Biobank (UK Biobank, 2022) offers a unique resource for investigating the associations between retinal structure and systemic

health, comprising over 85,000 OCT scans and 170,000 fundus images linked to cognitive and health data (2010–2015). Between the initial scan and July 2023, 1,216 participants in the dataset had a dementia diagnosis and 539 had an AD diagnosis. Therefore, although the UK Biobank is somewhat limited by the absence of OCTA scans, it holds significant potential for AD prediction using fundus and OCT scans in future studies. However, UK Biobank Alzheimer's labels and imaging is not coherent which makes it challenging for a classification dataset. This dataset and related issues is further examined in detail in the following section.

Table 2.3: Summary of Recent Deep Learning Studies on AD and MCI Detection using Retinal Imaging and Quantitative Data

Study	**Targets**	**Dataset**	**Age Matched?**	**Modalities & Quantitative Inputs**	**AUC**	**AUC (OCT Only)**
Hao et al. (2024)	EOAD, MCI	Private (ROAD, ROMCI)	No ($p < 0.001$)	OCTA Images	AD: 0.936; MCI: 0.863	–
S. Liu et al. (2025)	AD, MCI	Private (ROAD, ROMCI)	No	OCTA Images	0.887; 0.880	–
J. Chua et al. (2025)	AD, MCI	Private (SERI, Bucharest)	No (Range 41–79)	OCT Maps Anatomical Parameters	0.910 (Combined)	0.820
Wisely et al. (2024)	MCI	Private (Duke)	No (MCI older)	OCT Maps+ OCTA Images Quantitative Data Demographics	0.809	0.681
Gao et al. (2023)	MCI	Private (Wenzhou)	Yes	OCT + Fundus Images	0.968	0.903

Note: CN = Cognitively Normal; EOAD = Early-Onset Alzheimer's Disease; RNFL = Retinal Nerve Fiber Layer; GCIPL = Ganglion Cell-Inner Plexiform Layer; OCTA = OCT

CHAPTER 3

3. UK BIOBANK DATASET

The UK Biobank database includes 502,386 participants, of whom 85,704 underwent OCT scans at two different times (instances 0 and 1) using the same OCT device, Topcon 3D OCT 1000 Mk 2 (S. Y. L. Chua et al., 2019). Within the whole dataset, until July 2023, 9,145 participants have a dementia diagnosis, and 3,955 have an Alzheimer's disease (AD) diagnosis. However, only 1,216 and 539 of these individuals had corresponding OCT scans.

Dementia and AD diagnoses were identified using linked electronic health records and International Classification of Diseases, 9th Revision (ICD-9) and 10th Revision (ICD-10) codes (WHO, 2014), respectively. ICD codes are the current global standards for reporting diseases and health conditions.

The study categorizes the UK Biobank participants into three distinct classes based on their longitudinal health records: **CN (Healthy Control):** Participants who remained cognitively normal throughout the follow-up period. **AD (Fu-ture Alzheimer's):** Participants who were cognitively healthy at baseline but received a future diagnosis of AD. **Dementia (Future Dementia):** Participants who were cognitively healthy at baseline but received a future diagnosis of any non-AD dementia.

We used the following ICD-9 codes: 331.0, 290.4, 331.1, 290.2, 290.3, 291.2, 294.1, 331.2, and 331.5; and the following ICD-10 codes: F00, F00.0, F00.1, F00.2, F00.9, G30, G30.0, G30.1, G30.8, G30.9, F01, F01.0, F01.1, F01.2, F01.3, F01.8, F01.9, I67.3, F02.0, G31.0, A81.0, F02, F02.1, F02.2, F02.3, F02.4, F02.8, F03, F05.1, F10.6, G31.1, and G31.8.

In our study we excluded the scans based on the following criteria:

Missing data on mean RNFL (mRNFL) and mean GCIPL (mGCIPL) values.

Image quality score (signal strength) less than 45 Db

The inner limiting membrane indicator less than 20% of the population (a measure of the minimum localized edge strength around the inner limiting membrane boundary across the entire scan; this measure can identify scans that contain regions of severe signal fading and segmentation errors).

The validity count indicator less than 20% of the population (a measure that can identify OCT scans with a significant degree of clipping in the z-axis dimension).

The motion correlation indicator less than 20% of the population and the max delta indicator greater than 80% of the population.(The motion indica-tor can be used to identify blinks, eye motion artifacts, and segmentation errors; scans with the highest degree of motion were considered of poor quality) The motion indicator was calculated from the lowest Pearson cor-relation and the highest absolute difference between the thickness of the nerve fiber layer and the total retina.

The highest and the lowest 1% of the mRNFL thickness values (to prevent spurious estimates due to outlying data points).

The spherical equivalent of a participant was less than -6 or greater than +6 diopter (as such high refractive errors can result in the spurious assessment of retinal thickness).

Intraocular Pressure (IOP) less than 21 mmHg or 0

After applying the exclusion criteria, 43,934 participants remained. Of these, 500 had dementia and 223 had AD. Table 3.1 details the sequential derivation of the analysis cohorts from the UK Biobank OCT dataset, including the full participant pool, subset with OCT imaging, high-quality imaging subset, and derived unmatched (filtered) and age-matched cohorts. For each cohort, the table reports the sample size, dementia prevalence, and Alzheimer's disease prevalence, expressed as both counts and percentages. We used standard experimental setups for data inclusion and exclusion, following the criteria outlined in previous studies (Patel et al., 2016; van der Heide et al., 2024).

3.1 STATISTICAL ANALYSIS OF THE UK BIOBANK DATASET

In the UK Biobank dataset, Alzheimer's Disease (AD) was diagnosed chronologically following the initial imaging and eye measurements. This timing is

Table 3.1: Sequential reduction of the dataset based on exclusion criteria applied to the UK Biobank OCT data. The table shows the number of participants remaining after each quality control step, along with the corresponding counts of dementia and Alzheimer's disease (AD) cases.

The exclusion criteria were missing values, poor image quality,

Exclusion Criteria	Cohort	Instance 0	Instance 1	Dementia	AD
Initial Dataset	85704	68509	19502	1216	539
Missing RNFL	82594	67129	15465	1150	506
Signal Quality < 45	75899	64023	11876	1040	456
ILM < 20 percentile	69388	59476	9912	936	412
Validity count < 20 percentile	65688	56359	9329	868	383
Motion correlation < 20 percentile	61990	53512	8478	800	356
Max Delta > 80 percentile	60162	52083	8079	774	344
Refraction < -6 and > 6	57281	49595	7686	746	333
Top, bottom 1% RNFL thickness	56396	49000	7396	724	325
Eye Surgery	52369	45643	6726	663	297
Eye Disease	47656	41602	6054	550	246
IOP > 21 mmHg or 0	43934	38371	5563	500	223

segmentation errors, and clinical or ophthalmic abnormalities. Final row indicates the cohort retained for analysis. critical, as it indicates that the ocular changes were captured during the preclin-ical phase of the disease. The analysis of this timeline shows a mean interval of 8.86 years (standard deviation of 2.70 years) between the baseline scans and the eventual clinical diagnosis. As illustrated in Table 3.2, the number of patients diagnosed with AD tends to increase over time. Specifically, the highest concen-tration of AD cases was identified in the $10 - 12$ year latency period (75 cases), underscoring the

potential of these measurements to serve as early predictive indicators.

Tables 3.3, 3.4 and 3.5 show the study cohort and variables. Table 3.3 gives the core features, such as demographic and clinical indicators and the retinal measures mRNFL and mGCIPL. Extended features, such as full retinal layer thickness in multiple regions and macular volume, are shown in Tables 3.4 and 3.4. Associations with dementia and AD were tested using Student's t-tests and chi-square tests (Student, 1908).

The average age in the dementia and AD groups was significantly higher (approximately 64.8 and 65.8 years, respectively) than that in the overall population (57 y). This age difference was statistically significant ($p < 0.001$).

Table 3.2: Cumulative incidence of Alzheimer's Disease (AD) relative to Cognitively Normal (CN) participants over time. The table details the number of incident AD cases diagnosed at specific year intervals following baseline eye measurements within the UK Biobank cohort. Percentages represent the proportion of AD cases relative to the total CN population ($n = 43,434$) at each time point. The mean time from baseline to diagnosis was 8.86 years

($SD = 2.70$ years).

	CN	AD	%
1	43434	2	0.00%
2	43434	3	0.01%
3	43434	9	0.02%
4	43434	19	0.04%
5	43434	29	0.07%
6	43434	45	0.10%
7	43434	60	0.14%
8	43434	84	0.19%
9	43434	107	0.25%
10	43434	145	0.33%
11	43434	180	0.41%
12	43434	223	0.51%

The sex distribution showed an interesting change: overall, there were

more women, but men were more common in the dementia (64% men) and AD (54% men) groups than in the general group (47% men). However, even though this difference seems noticeable and significant in dementia, it was marginal in AD.

The educational status varied significantly across groups (dementia vs. control: $p < 0.001$; AD vs. control: $p = 0.003$). Higher education was less common in dementia (36%) compared to the full cohort (42%) . However, lower secondary (30% vs. 31%) and vocational education (11% vs. 8%) were more common in dementia. In contrast, the AD group did not display a clear reduction in higher education attainment (42%, the same as the controls); however, vocational education was again more common among AD participants (10%) than among controls (8%). These findings reinforce previous evidence that lower educational status, particularly vocational or non-tertiary pathways, may contribute to a higher risk of cognitive decline and dementia (Maccora et al., 2020).

From a clinical perspective, several health markers indicate an elevated risk of dementia and AD subgroups. Diabetes was higher in the dementia group (9% vs. 4% in the general population), and systolic blood pressure was significantly higher (142–143 mmHg vs. 136 mmHg; $p < 0.001$). The use of antihypertensive medication is substantially lower among those with dementia and AD—only 49–54% of these individuals use such medications, compared to 77% in the broader population.

Lifestyle factors also differed between the groups. Although alcohol use was approximately the same, there was a clear difference in the smoking history. A higher percentage of people in the dementia and AD groups were former smokers (46%) than in the general population (35%), and this difference was statistically significant ($p < 0.001$). This suggests that a history of smoking is associated with a higher risk of cognitive decline.

Eye measurements have shown clear and consistent retinal thinning in individuals with dementia and AD. For example, RNFL was much thinner in affected people (mean: 27.7 μm vs. 29.2 μm), and similar thinning was also observed in the GCIPL.

Examination of the extended features in Table 5.3 showed that most retinal and macular measures were reduced in dementia, and the differences were statistically significant ($p < 0.05$). In AD, however, increases were detected in the ISOS–RPE central subfield ($p = 0.017$), ELM–ISOS central subfield ($p = 0.029$), disc diameter after inverse rank normal transformation ($p = 0.027$), and vertical cup-to-disc ratio ($p = 0.024$). Although most INL–ELM subfields showed slight upward shifts, these were not statistically significant for either dementia or AD. In contrast, all ELM–ISOS and ISOS–RPE subfields showed measurable reductions in dementia ($p < 0.05$), whereas in AD, only the central subfields of these layers showed differences ($p = 0.029$ and 0.017).

All macular indices and subfield thicknesses, as well as total macular volume, showed clear reductions in both dementia and AD, except for the central subfield, which displayed a small but non-significant increase. The vertical cup-to-disc ratio was elevated in both dementia and AD, while the vertical cup-to-disc ratio regressed and transformed showed consistent downward shifts in both groups (dementia: $p = 0.004$; AD: $p = 0.012$).

Table 3.3: Core demographic, clinical, and retinal features of the study population, including overall cohort, cognitively normal individuals, and participants with dementia and Alzheimer's disease (AD). Values are presented as mean ± SD, median (IQR), or count (%) as appropriate. Statistical comparisons used t-tests or chi-square tests; p-values < 0.05 are significant.

		Overall	Without	Dementia			AD		
CORE FEATURES		**N = 43,934**	**N = 43,434**	**N = 500**	T/χ^2	**p**	**N = 223**	T/χ^2	**p**
Age (mean,STD)	43934	57.09 ± 8.17	57.0 ± 8.15	64.76 ± 5.56	−30.82	∗ < 0.001	65.78 ± 4.02	−32.23	∗ < 0.001
Sex (count, %)	43934				17.08	∗ < 0.001		3.84	0.05
Women		23229,%53	23011,%53	218,%44			103,%46		
Men		20705,%47	20423,%47	282,%64			120,%54		
Educational status (count, %)	38049				24.64	∗ < 0.001		15.84	* 0.003
Higher		16035,%42	15912,%42	123,%36			65,%42		
Upper secondary		5263,%14	5220,%14	43,%13			20,%13		
Lower secondary		11675,%31	11573,%31	102,%30			36,%23		
Vocational		2880,%8	2844,%8	36,%11			16,%10		
Other		2196,%6	2158,%6	38,%11			19,%12		
Diabetes (count,%)	38371				30.85	∗ < 0.001		3.31	0.069
Without diabetes		36932,%96	36516,%96	416%0.91			192,%94		
With diabetes		1439,%04	1399,%04	40%0.09			13,%06		
Spherical equivalent (median, IQR)	43934	0.61,(0.15-1.14)	0.6,(0.14-1.13)	1.01, (0,46-1,59)	-8.83	∗ < 0.001	1.07,(0.52-1.56)	-6.6	0.91
Systolic blood pressure (mmHg) (mean,STD)	43491	136.39 ± 18.16	136.32 ± 18.15	142.29 ± 18.1	-7.32	∗ < 0.001	143.3 ± 18.92	-5.47	∗ < 0.001
Diastolic blood pressure (mmHg) (mean,STD)	43493	81.44 ± 9.97	81.44 ± 9.97	81.75 ± 9.91	−0.68	0.91	81.52 ± 10.7	-0.11	0.507
Antihypertensive medication use (count, %)	17677				76.86	∗ < 0.001		48.29	∗ < 0.001
With antihypertensive medication		13676,%77	13542,%78	134%0.54			50,%49		
Without antihypertensive medication		4001,%23	3887,%22	114%0.46			53,%51		
Alcohol Consumption (count, %)	43811				10.67	0.058		4.3	0.507
Daily		9027,%21	8937,%21	90%0.18			45,%2		
3 or 4 /week		10256,%23	10138,%23	118%0.24			53,%24		
1 or 2/week		11130,%25	11019,%25	111%0.22			46,%21		
1-3/month		5035,%11	4973,%11	62%0.12			29,%13		
Special occations		5043,%12	4977,%11	66%0.13			28,%13		
Never		3320,%08	3268,%08	52%0.1			22,%1		
Smoking Status (count, %)	43709				28.86	∗ < 0.001		14.71	∗ < 0.001
Never		24538,%56	24313,%56	225%0.46			102,%47		
Previous		15150,%35	14923,%35	227%0.46			102,%47		
Current		4021,%09	3980,%09	41%0.08			14,%06		
Body-mass index (mean, STD)	43752	27.2 ± 4.68	27.19 ± 4.67	27.75 ± 4.95	-2.49	0.924	27.16 ± 4.88	0.1	0.924
Retinal thickness indices (mean, STD)	43934								

Table 3.3 (Continuing) Core demographic, clinical, and retinal features of the study population, including overall cohort, cognitively normal individuals, and participants with dementia and Alzheimer's disease (AD). Values are presented as mean ± SD, median (IQR), or count (%) as appropriate. Statistical comparisons used t-tests or chi-square tests; p-values < 0.05 are significant.

RNFL thickness(μm), mean,SD	43934	29.17 ± 4.93	29.18 ± 4.93	27.77 ± 4.66	6.75	∗ < 0.001	27.71 ± 4.98	4.42	∗ < 0.001
GCIPL thickness(μm), mean,SD	43934	73.68 ± 6.6	73.7 ± 6.59	72.14 ± 6.97	4.96	∗ < 0.001	71.67 ± 6.4	4.71	∗ < 0.001

Table 3.4: Extended retinal features (layer-specific and subfield thickness measures) of the study population, including overall cohort, cognitively normal individuals, and participants with dementia and Alzheimer's disease (AD). Values are presented as mean ± SD, median (IQR), or count (%) as appropriate. Statistical comparisons used t-tests or chi-square tests; p-values < 0.05 are significant.

		Overall	Without	Dementia			AD		
EXTENDED FEATURES 1		**N = 43,934**	**N = 43,434**	**N = 500**	T/χ^2	**p**	**N = 223**	T/χ^2	**p**
Retinal thickness indices (mean, STD)	43934								
INL thickness(μm), mean,SD	43934	32.66 ± 2.52	32.66 ± 2.52	32.37 ± 2.69↓	2.42	0.016	32.35 ± 2.36↓	1.96	0.051
INL-ELM indices(μm), mean,SD	43934								
Average thickness	43934	80.52 ± 6.64	80.52 ± 6.63	80.64 ± 7.17↑	-0.39	0.698	80.63 ± 6.37↑	-0.27	0.787
central subfield thickness	43934	107.96 ± 10.52	107.97 ± 10.51	107.64 ± 11.33↓	0.64	0.522	107.55 ± 11.15↓	0.55	0.582
inner subfield thickness	43934	93.72 ± 7.78	93.72 ± 7.77	93.79 ± 8.33↑	-0.2	0.839	93.89 ± 7.53↑	-0.35	0.73
outer subfield thickness	43934	75.54 ± 6.58	75.54 ± 6.57	75.72 ± 7.06↑	-0.56	0.574	75.67 ± 6.27↑	-0.3	0.764
ELM-ISOS indices(μm), mean,SD	43934								
Average thickness	43934	23.76 ± 1.83	23.77 ± 1.83	23.49 ± 1.95↓	3.12	∗ < 0.001	23.57 ± 1.64↓	1.81	0.072
central subfield thickness	43934	28.33 ± 2.3	28.33 ± 2.29	27.93 ± 2.45↓	3.65	∗ < 0.001	27.99 ± 2.29↓	2.2	* 0.029
inner subfield thickness	43934	24.57 ± 1.83	24.58 ± 1.83	24.32 ± 1.96↓	2.88	∗ < 0.001	24.4 ± 1.58↓	1.65	0.101
outer subfield thickness	43934	23.35 ± 1.97	23.35 ± 1.96	23.09 ± 2.07↓	2.83	∗ < 0.001	23.16 ± 1.76↓	1.66	0.098
ISOS-RPE indices(μm), mean,SD	43934								
Average thickness	43934	37.96 ± 4.09	37.97 ± 4.09	37.43 ± 4.37↓	2.73	* 0.007	37.7 ± 4.19↓	0.95	0.345
central subfield thickness	43934	42.51 ± 6.2	42.53 ± 6.2	41.33 ± 6.06↓	4.37	∗ < 0.001	41.55 ± 6.01↑	2.41	* 0.017
inner subfield thickness	43934	38.51 ± 4.93	38.52 ± 4.93	37.98 ± 4.96↓	2.43	* 0.015	38.25 ± 4.63↓	0.88	0.38
outer subfield thickness	43934	37.62 ± 3.94	37.62 ± 3.93	37.11 ± 4.29↓	2.68	* 0.008	37.4 ± 4.13↓	0.8	0.422
INL-RPE indices(μm), mean,SD	43934								
Average thickness	43934	142.25 ± 8.62	142.25 ± 8.6	141.57 ± 9.67↓	1.58	0.116	141.9 ± 7.88↓	0.66	0.509
central subfield thickness	43934	178.8 ± 13.7	178.83 ± 13.68	176.91 ± 15.05↓	2.84	* 0.005	177.1 ± 14.5↓	1.77	0.078
inner subfield thickness	43934	156.81 ± 10.34	156.81 ± 10.33	156.09 ± 11.14↓	1.44	0.15	156.54 ± 9.48↓	0.43	0.667
outer subfield thickness	43934	136.51 ± 8.4	136.52 ± 8.39	135.92 ± 9.41↓	1.42	0.156	136.23 ± 7.61↓	0.57	0.567

Table 3.5: Extended retinal features (layer-specific and subfield thickness measures) of the study population, including overall cohort, cognitively normal individuals, and participants with dementia and Alzheimer's disease (AD). Values are presented as mean ± SD, median (IQR), or count (%) as appropriate. Statistical comparisons used t-tests or chi-square tests; p-values < 0.05 are significant.

EXTENDED FEATURES 2		Overall N = 43,934	Without N = 43,434	Dementia N = 500	T/χ^2	p	AD N = 223	T/χ^2	p
Retinal thickness indices (mean, STD)	43934								
RPE indices(μm), mean,SD	43934								
Overall thickness	43934	25.81 ± 8.17	25.81 ± 8.12	26.18 ± 11.76↑	-0.7	0.482	25.4 ± 5.46↓	1.11	0.268
central subfield	29876	26.4 ± 4.64	26.4 ± 4.64	25.78 ± 4.8↓	2.2	* 0.029	26.14 ± 5.09↓	0.61	0.541
inner inferior subfield	29876	24.73 ± 4.12	24.73 ± 4.12	24.51 ± 3.89↓	0.94	0.347	24.58 ± 3.9↓	0.45	0.656
inner nasal subfield	29876	27.16 ± 4.19	27.16 ± 4.19	27.1 ± 3.86↓	0.27	0.786	26.96 ± 3.91↓	0.59	0.554
inner superior subfield	29876	24.56 ± 4.05	24.56 ± 4.06	24.21 ± 3.84↓	1.52	0.129	24.27 ± 3.73↓	0.91	0.362
inner temporal subfield	29876	26.15 ± 4.14	26.15 ± 4.14	26.12 ± 3.93↓	0.15	0.882	26.24 ± 3.81↑	-0.27	0.787
outer inferior subfield	29876	24.0 ± 2.75	24.0 ± 2.75	23.74 ± 2.5↓	1.77	0.077	23.56 ± 2.19↓	2.34	* 0.021
outer nasal subfield	29876	26.91 ± 3.65	26.91 ± 3.65	26.53 ± 3.41↓	1.88	0.061	26.51 ± 3.37↓	1.41	0.16
outer superior subfield	29876	24.33 ± 2.86	24.34 ± 2.86	24.09 ± 2.57↓	1.59	0.112	23.94 ± 2.42↓	1.9	0.059
outer temporal subfield	29876	25.67 ± 3.49	25.67 ± 3.49	25.58 ± 3.23↓	0.47	0.641	25.39 ± 3.33↓	0.97	0.332
Macular thickness indices(μm), mean,SD									
Overall thickness	43934	277.76 ± 14.55	277.8 ± 14.51	273.86 ± 17.2↓	5.11	* < 0.001	273.64 ± 14.68↓	4.22	* < 0.001
central subfield	38218	265.56 ± 24.49	265.53 ± 24.43	267.46 ± 28.88↑	-1.39	0.164	266.6 ± 25.14↑	-0.6	0.551
inner inferior subfield	38218	310.29 ± 18.92	310.34 ± 18.87	306.13 ± 22.42↓	3.93	* < 0.001	305.4 ± 18.0↓	3.86	* < 0.001
inner nasal subfield	38218	317.45 ± 19.17	317.5 ± 19.14	313.29 ± 21.65↓	4.06	* < 0.001	312.7 ± 20.24↓	3.34	* 0.001
inner superior subfield	38218	312.97 ± 19.65	313.03 ± 19.59	307.55 ± 23.47↓	4.89	* < 0.001	306.55 ± 19.26↓	4.73	* < 0.001
inner temporal subfield	38218	301.96 ± 18.91	302.0 ± 18.77	298.45 ± 28.08↓	2.65	* 0.008	297.78 ± 19.66↓	3.02	* 0.003
outer inferior subfield	38218	263.2 ± 16.92	263.22 ± 16.88	261.43 ± 20.5↓	1.83	0.067	259.51 ± 16.5↓	3.16	* 0.002
outer nasal subfield	38218	286.63 ± 17.22	286.67 ± 17.19	282.94 ± 19.15↓	4.07	* < 0.001	282.2 ± 17.88↓	3.51	* < 0.001
outer superior subfield	38218	269.12 ± 16.57	269.16 ± 16.55	265.03 ± 17.54↓	4.92	* < 0.001	264.04 ± 17.8↓	4.05	* < 0.001
outer temporal subfield	38218	255.34 ± 15.63	255.36 ± 15.57	253.62 ± 20.05↓	1.81	0.071	252.21 ± 15.03↓	2.95	* 0.004
Total Macular Volume	29804	7.88 ± 0.37	7.88 ± 0.37	7.8 ± 0.35↓	3.93	* < 0.001	7.79 ± 0.36↓	2.78	* 0.006
Disc Diameter indices									
Mean of vertical disc diameter	35758	123.95 ± 15.37	123.93 ± 15.37	125.41 ± 15.15↑	-1.96	0.051	125.07 ± 14.03↑	-1.09	0.277
Disc diameter after inverse rank normal transformation	35758	0.05 ± 0.93	0.04 ± 0.93↓	0.17 ± 0.91↑	-2.86	* 0.005	0.18 ± 0.84↑	-2.23	* 0.027
Vertical cup to disc ratio indices									
Vertical cup to disc ratio	35601	0.32 ± 0.17	0.32 ± 0.17	0.34 ± 0.18↑	-2.47	* 0.014	0.35 ± 0.18↑	-2.28	* 0.024
Vertical cup to disc ratio regressed and transformed	35601	−0.04 ± 0.93	−0.04 ± 0.93	0.1 ± 0.98↑	-2.9	* 0.004	0.15 ± 0.99↑	-2.54	* 0.012

CHAPTER 4

4. EARLY AD PREDICTION IN UK BIOBANK DATASET

This chapter outlines the machine learning pipeline for the early prediction of Alzheimer's disease. We used the UK Biobank dataset for this study. The process began with data pre-processing. We applied filtering protocols to handle missing values and mitigate the dataset bias. Next, we trained an XGBoost classifier. This model processed feature sets that contained retinal layer measurements and clinical biomarkers. We used SHAP values to ensure model interpretability. This allowed us to quantify the feature importance for both retinal and non-retinal data. We also performed a longitudinal survival analysis using Cox proportional hazards and Nelson–Aalen estimates. These methods validated the clinical significance of the identified features over time. Ultimately, this analysis establishes a statistical baseline for the deep learning architectures presented in the subsequent sections.

4.1 DATASET PREPARATION

We implemented strict filtering and exclusion protocols to ensure data quality. However, several key features still contained null-value entries. Missing data points were observed for attributes such as education level, diabetes status, and blood pressure. The dataset also showed missing entries for the use of antihypertensive medications, alcohol consumption, and smoking status. Furthermore, we identified missing data for the body mass index (BMI) and specific ophthalmic measurements. Table 4.1 quantifies the missing values across all demographic and clinical variables in the UK Biobank.

We used a targeted exclusion strategy to address missing data without introducing bias through imputation in the dementia group. Specifically, we randomly dropped 90% of participants who were cognitively normal but had miss-

ing data. To avoid further reducing the number of dementia patients, we included all participants who had been diagnosed with dementia, regardless of missing values.

Before filtering, missing values were present in a large proportion of the 43,934 participants. After filtering 23,139 participants with fewer missing values remained. Table 4.1 shows the remaining missing data, particularly among dementia case to assess group balance.

Table 4.1: Number of missing values for key demographic, clinical, and ophthalmic features in the UK Biobank dataset before and after additional filtering. "Before" indicates the number of participants with missing values prior to filtering, while "After" refers to the number of participants with missing values after the filtering process for the first dataset. The number of missing values among participants with dementia and AD after filtering is presented in the last two columns. The initial dataset comprised 43,934 participants; after filtering to reduce bias, 23,139 participants remained.

	Dementia		AD	
Feature	**Before**	**After**	**Before**	**After**
Educational Status	5,727	702	157	66
Diabetes	5,563	630	44	18
Systolic Blood Pressure	443	42	2	2
Diastolic Blood Pressure	441	42	2	2
Antihypertensive Usage	26,257	12,692	252	120
Alcohol Consumption	123	11	1	0
Smoking Status	225	29	7	5
Body Mass Index (BMI)	182	24	2	0
Macular Thickness Subfields	5,716	660	59	24
Total Macular Volume	14,130	1,650	220	89
RPE Thickness Subfields	14,058	1,635	211	84
Mean of Vertical Disc Diameter	8,176	897	92	41
Disc Ratio	8,333	909	94	43

After initial preprocessing, substantial missing data remained, particularly in variables related to antihypertensive medication usage, educational status, and several ophthalmic measurements. Because missingness

was non-uniform across diagnostic groups (e.g., a higher proportion of missing educational status values occurred among dementia cases), explicit strategies were required to assess its impact on predictive modeling.

XGBoost (Ryu et al., 2020) implements a native strategy for handling missing data during tree construction. When evaluating a split on a feature, the algorithm not only determines the optimal thresholds for observed values but also learns a default direction (left or right branch) for instances with missing values. This default path is chosen to maximize the gain function, ensuring that the treatment of missing values is optimized jointly with the split criterion. During inference, any sample with a missing value in a split feature is routed along this learned default path. This approach removes the necessity for separate imputation steps.

Participants with dementia or AD were substantially older than the general population, creating the risk that predictive models could exploit age-related variation rather than disease-specific signals. To control for this confounding factor, we constructed age-matched control cohorts for both dementia and AD. Previous studies have demonstrated that RNFL, macular, and photoreceptor thickness decline naturally with age (Alamouti & Funk, 2003; Khawaja et al., 2020), necessitating disentanglement of ageing effects from pathological changes.

We implemented stratified sampling by dividing participants into discrete age bins and randomly selecting 10 cognitively normal (CN) participants per dementia or AD case from the same bin. This ensured comparable age distributions across groups while maintaining a relatively balanced and sufficient sample size for model training. To account for the variance introduced by random sampling, as described in the next section, we repeated this procedure five times. Retinal feature analyses—including measurements of the RNFL, GCIPL, and other retinal layers—were conducted using both unmatched and age-matched datasets. The age-matched datasets enabled more accurate isolation of the effects of dementia and AD from age-related changes.

4.2 METHODS

XGBoost has been successfully applied in numerous medical research stud-ies. For instance, Zhang et al. (2024) trained a multiclass XGBoost model us-ing blood-based biomarkers and showed that it was successful in differentiating between individuals with AD, cognitively normal (CN), and mild cognitive im-pairment (MCI) (Zhang et al., 2024). Nguyen et al. (2023) used MRI features to create a three-class XGBoost-based classification model (AD vs. early mild cognitive impairment (EMCI) vs. cognitively normal). We selected XGBoost as the primary model for this study based on its proven performance.

We used the XGBoost classifier with the following hyperparameter set-tings: learning_rate = 0.1, max_depth = 3, eta = 0.3, gamma=0, lambda=1, alpha=0, and n_estimators = 100. To perform binary classification, the evaluation metric was defined as "logloss," and the objective parameter was assigned as "binary:logistic." In addition, enable_categorical was set to True to permit the native handling of categorical features.

Our dataset consisted of cognitively normal (CN), Alzheimer's disease (AD), and dementia patients. We conducted two separate binary classification tasks: (1) dementia vs. CN and (2) AD vs. CN. For each task, we evaluated the model performance using different strategies for handling missing data, using XGBoost as the base classifier.

The data were split in each iteration, with 80% allocated to training and 20% to testing. For statistical robustness (Bouckaert, 2003), trainings were re-peated five times using different random seeds. Each train–test split was rerun five times on the resampled datasets. The reported results are averages over the 5×5 experiments. Analyses were performed on both the unmatched dataset and an age-matched version, in which cognitively normal (CN) participants were randomly subsampled to match the age distribution of the AD and dementia groups. Each classification task was performed using both the core and extended feature sets. The model performance was evaluated using the area

under the receiver operating characteristic curve (AUC) (Fawcett, 2006). The mean and standard deviation of the AUC values were derived from 5×5 repeated runs. To visualize the performance, ROC curves were generated using the vertical av-eraging of true positive rates (TPRs) at fixed false-positive rate (FPR) intervals across repetitions, as recommended for the consistent comparison of multiple models (Fawcett, 2006). In addition, we reported the true positive rate (TPR) at a fixed false-positive rate (FPR) of 0.15, which is a commonly accepted thresh-old in biomarker evaluation studies (Khoury & Ghossoub, 2019; Schindler et al.,2024).

Model Interpretability

We applied and compared the following interpretability methods to investigate which features were the most important for the model's decision-making. In the context of model explainability, global interpretability refers to under-standing how the model behaves across the entire dataset, as opposed to local interpretability, which focuses on the individual predictions of the model.

Global Interpretability through XGBoost Feature Importance: XG-Boost provides global interpretability by generating feature importance scores derived from its ensemble of decision trees. For every feature in-volved in a split, XGBoost calculates how much that split improves the model's objective function (in our case, binary:logistic). These improve-ments, known as gains, are then averaged across all splits and trees, illus-trating the total impact of the features on minimizing loss. This approach allowed us to gain a deeper understanding of the data by providing an overall ranking of the features and highlighting the variables that the model prioritizes during training.

Local Interpretability through Shap: SHAP is a model-agnostic, game-theoretic method that quantifies the contribution of each feature to individ-ual predictions (Lundberg & Lee, 2017). Unlike global feature importance, SHAP provides instance-level explanations, which are essential in medi-cal

applications requiring interpretable decisions. For example, Yi et al. (2023) used SHAP with XGBoost to build an interpretable Alzheimer's disease classification model on imbalanced datasets. In this study, we ap-plied SHAP for two purposes: (1) generating beeswarm plots from the vertically concatenated (pooled) SHAP values across 25 test runs to il-lustrate how features influenced predictions in both direction and magni-tude relative to XGBoost and statistical feature importance results; and (2) producing waterfall plots to demonstrate a clear explanation of individual early predictions within a high-dimensional feature space.

Further Validation of Identified Important Features

We also performed conventional statistical analyses to further interpret our findings and assess the clinical significance of individual biomarkers. These experiments had two aims: (1) to determine whether the features identified as important by the machine learning model were also statistically associated with AD outcomes and (2) to quantify the effect of these features on the risk and timing of AD diagnosis in our cohort.

We employed the Cox proportional hazards model (Cox, 1972), a standard method in survival analysis, to examine the relationship between selected retinal or clinical features and the time of Alzheimer's disease onset. The Cox model is a semi-parametric model that estimates the hazard function. It models the effect of each feature with a log-linear coefficient, which indicates the proportional change in hazard for a one-unit change in that feature, while holding the others constant. The Cox model also accounted for censored data, that is, participants who did not develop AD during the follow-up period or who were lost to follow-up, by including them in the analysis up to the point they were last observed. This allowed us to estimate how features increase or decrease the risk of developing AD over time, even when not all participants experienced these events. However, we note that AD diagnosis records and dates in the dataset may not be entirely precise, which could affect the timing estimates.

Additionally, we utilized the Nelson-Aalen (Aalen, 1978) estimator to calculate cumulative hazard curves for groups stratified by key features: "thin" (lower than median value) and "thick" (larger than median value). This non-parametric estimator provides an interpretable visualization of how the cumula-

tive risk of disease changes over time for different subgroups of the dataset with respect to a particular feature.

Finally, we used Spearman's pairwise complete correlation coefficient matrix (Spearman, 1904) to examine the relationships between the top features identified in unmatched and age-matched datasets. The relationships among the top features were examined to assess whether they showed meaningful patterns in the data. This analysis added a further check on the reliability of the model's feature selection.

4.3 RESULTS

In this section, we present the main results of our study based on the steps described in the Methods section. First, we report the performance of the XG-Boost models in predicting dementia and AD. We then focused on AD and exam-ined which features had the most significant impact on the model's predictions using both overall and individual explanation methods. We then assessed the rel-evance of these features using statistical approaches, including survival analysis. Finally, we explored the relationships among the top features to better understand their interdependencies and the effects of age matching. All the experimental re-sults presented below are the average of 5×5 repeated random $80/20$ train–test splits.

4.3.1 Early Prediction Performance

The predictive performance of the XGBoost models for both dementia versus non-dementia and AD versus non-AD classification tasks is summarized in Table 4.2, with the corresponding ROC curves shown in Figures 4.1 and 4.2.

XGBoost with extended features achieved the highest performance for both dementia and AD detection versus cognitively normal controls, with unmatched mAUCs of 0.892 and 0.894 and age-matched mAUCs of 0.782 and 0.763, respectively. Performance declined significantly in age-matched cohorts, highlighting the impact of age.

Table 4.2: Mean AUC (mAUC) and mean TPR (mTPR) (at FPR = 0.15) of 25 test runs for XGBoost models on dementia and AD classification tasks. Results are shown for core and extended (includes core as well) feature sets, and with/without age-matching.

Task	Feature Set	Unmatched		Age-matched	
		mAUC	mTPR	mAUC	mTPR
Dementia vs CN	Core	0.864 ± 0.015	0.719 ± 0.042	0.698 ± 0.028	0.462 ± 0.042
	Extended	**0.892 ± 0.013**	**0.754 ± 0.029**	**0.782 ± 0.021**	**0.646 ± 0.047**
AD vs CN	Core	0.867 ± 0.019	0.667 ± 0.076	0.68 ± 0.031	0.419 ± 0.078
	Extended	**0.894 ± 0.018**	**0.746 ± 0.056**	**0.763 ± 0.034**	**0.576 ± 0.059**

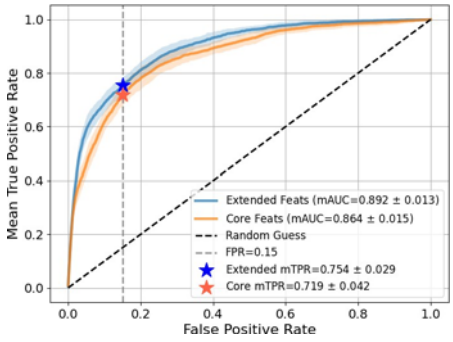
4.3.2 Model Interpretability Results

To search for AD-related potential biomarkers, we analyzed the top 15 mean feature importances across 25 test runs. These feature importance values were identified by the best-performing model using the built-in feature importance algorithm of XGBoost. We visualized their contributions through SHAP beeswarm plots, as shown in Figures 4.3, 4.4, and 4.5. This analysis focused on the model trained with XGBoost's native handling of missing features for AD vs. Non-AD prediction.

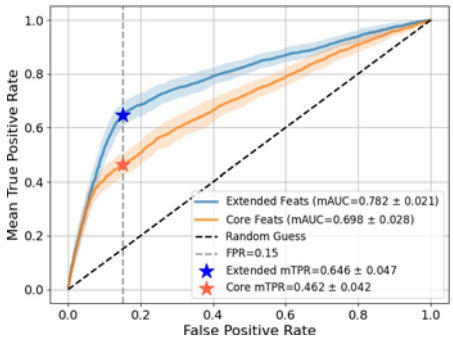
XGBoost Feature Importances:

Figure 4.3 (a) shows the feature importance ranking for tests with unmatched datasets, and Figure 4.3 (b) presents the results for tests with age-matched datasets for AD prediction, where the missing values were handled automatically by XGBoost.

In the unmatched datasets, educational status was the most dominant feature (mean importance, 0.086), followed by age and sex. In contrast, when age effects were removed through matching, total macular volume became the top feature (0.085), and the vertical cup-to-disc ratio regressed and transformed; several macular and RPE thickness measurements gained relative importance.



(a) Unmatched Datasets



(b) Age-Matched Datasets

Figure 4.1: Mean ROC curves of 25 train/test runs for dementia vs healthy classification where XGBoost handles missing values.

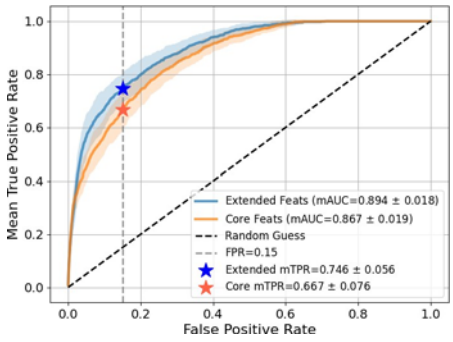
Both analyses showed that several RPE thickness measures (inner superior, outer superior, outer nasal, outer temporal) and macular structural metrics (macular thickness, total macular volume, mRNFL, mGCIPL) ranked highly. These features remained top predictors in unmatched and age-matched datasets.

Furthermore, educational status remained an important non-retinal factor in both analyses, and the importance of antihypertensive use increased when age was controlled.

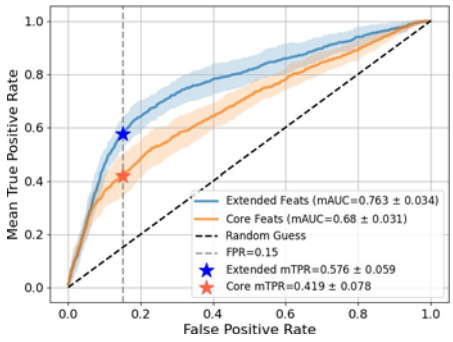
SHAP Importances:

Figures 4.4 and 4.5 show the SHAP beeswarm plots illustrating the contribution of individual features to AD vs. non-AD classification using unmatched

and age-matched datasets, respectively.



(a) Unmatched Datasets



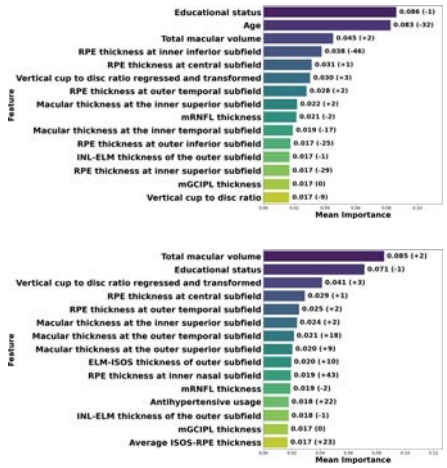
(b) Age-Matched Datasets

Figure 4.2: Mean ROC curves of 25 train/test runs for AD vs healthy classification. (a) Unmatched Datasets. (b) Age-Matched Datasets.

As expected, in the unmatched datasets (Figure 4.4), age was the most influential feature, with the widest SHAP value distribution observed. Several macular and RPE thickness measurements (e.g., inner inferior, outer nasal, outer inferior, and superior subfields) showed moderate SHAP value. Other ocular features reflecting global ocular geometry, such as disc diameter and spherical equivalent, contributed to the classification decisions.

For the age-matched datasets, the dominance of age diminished considerably, and the SHAP distribution became more balanced across the retinal features (Figure 4.5). RPE thickness at the outer nasal and inferior subfields, average RPE

thickness, mGCIPL thickness, and macular thickness gain were relatively important. These findings highlight macular- and RPE-related measures as central discriminative features independent of age. Educational status, alcohol consumption, and antihypertensive use showed smaller but measurable contributions.



(a) Unmatched Datasets (b) Age-Matched Datasets

Figure 4.3: Top 15 mean feature importances across 25 test runs from the XGBoost model for AD vs non-AD classification where XGBoost handles missing values. Numbers in parentheses indicate the change in feature ranking between unmatched and age-matched datasets.

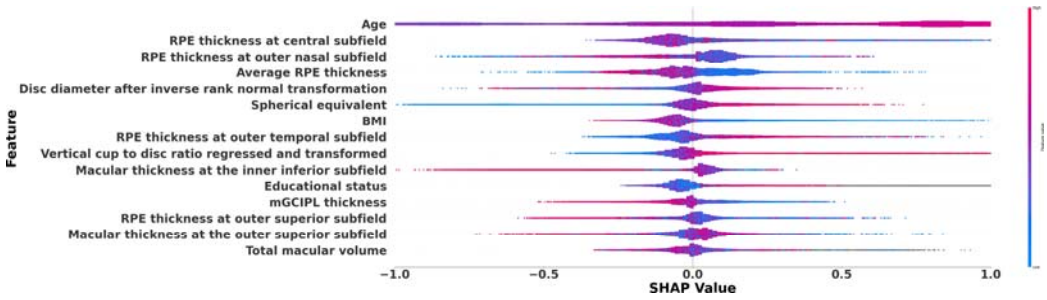


Figure 4.4: SHAP beeswarm plots of 25 test runs showing the top 15 features for AD vs non-AD classification where XGBoost handles missing values.

To illustrate how individual features influence the predictions at the individual level, Figures 4.6 and 4.7 show the SHAP waterfall plots for the representative cases. In the AD-positive example (Figure 4.6 (a)), macular thickness at the inner nasal (+0.46) and outer nasal (+0.25) subfields, together with RPE thickness at the central subfield (+0.18), were the dominant drivers shifting the prediction toward AD, outweighing the protective effects of RPE thickness at the outer temporal (−0.26), mGCIPL thickness (−0.17), and total macular volume (−0.15). In contrast, in the CN case (Figure 4.6 (b)), negative contributions from the mean vertical disc diameter (−0.15), central macular thickness (−0.14), and outer temporal RPE thickness (−0.13) outweighed the smaller AD-driving effects from the INL–ELM thickness (+0.21) and mRNFL thickness (+0.12), resulting in a correct classification as cognitively normal. Together, these case level explanations complement the cohort-level beeswarm plots by showing how combinations of retinal and systemic features influence individual-predictions.

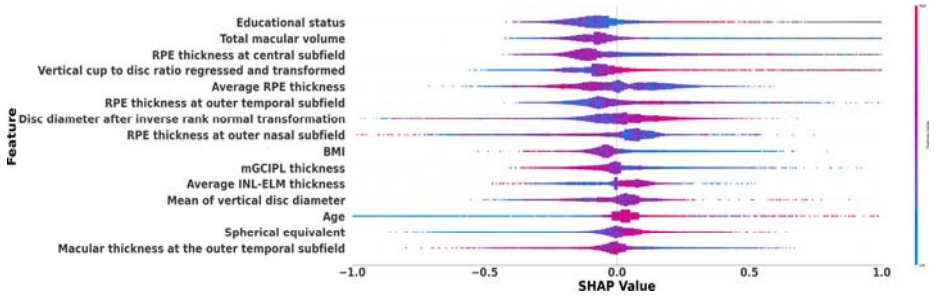


Figure 4.5: SHAP beeswarm plots of 25 test runs showing the top 15 features for AD vs non-AD classification for the age-matched case where XGBoost handles missing values.

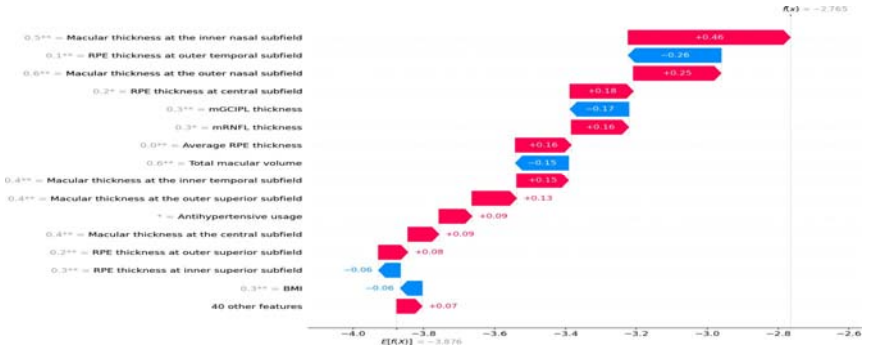
Figure 4.7 shows two representative misclassified cases. In the first exam-ple (AD misclassified as CN; Figure 4.7 (c)), protective signals such as higher total macular volume (-0.20) and educational status (-0.14) outweighed the smaller AD-driving effects from RPE thickness at the inner superior subfield (+0.12) and macular thickness at the outer inferior subfield (+0.13). The net effect shifted the prediction below the AD threshold despite the true AD label. In the CN misclassified as an AD case (Figure 4.7 (d)), outer temporal RPE (+0.69), central RPE (+0.57), and average RPE (+0.34) strongly drove the prediction toward AD. Smaller protective effects from total macular volume (-0.16) and mRNFL (-0.08) could not counterbalance them.

4.3.3 Survival Analysis and Cumulative Hazard Estimates

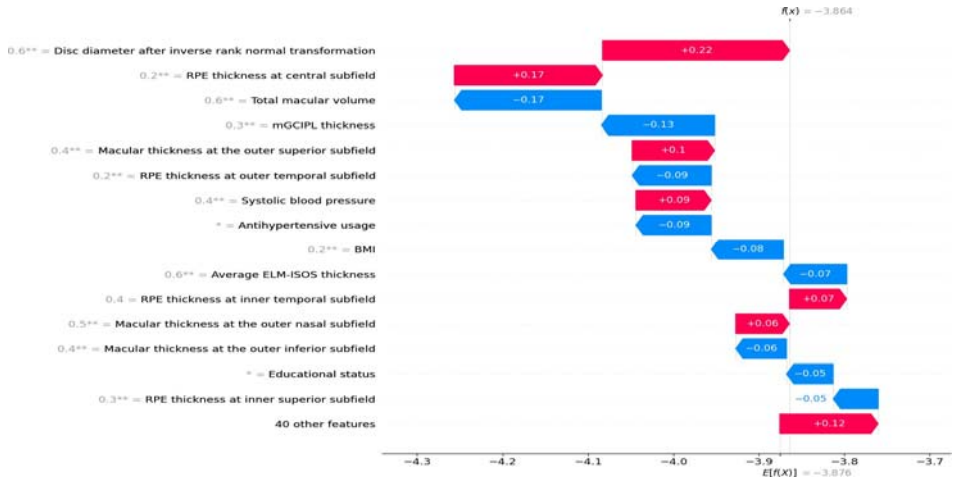
To test the clinical importance of the retinal and clinical features identified by our machine learning models, we performed a survival analysis using the Cox proportional hazards model in age-matched datasets. For this calculation, we used KNN-imputed datasets because Cox regression cannot be performed with missing values.

Figure 4.8 shows the corresponding forest plot of the average log hazard ratios with 95% confidence intervals. Significant associations ($p < 0.05$) were found between total macular volume, alcohol consumption, age, educational sta-

tus, average RPE thickness, overall macular thickness, and INL–ELM thickness in the outer subfields. These results were consistent with the XGBoost and SHAP results, which also ranked macular and RPE-related metrics highest and identi-

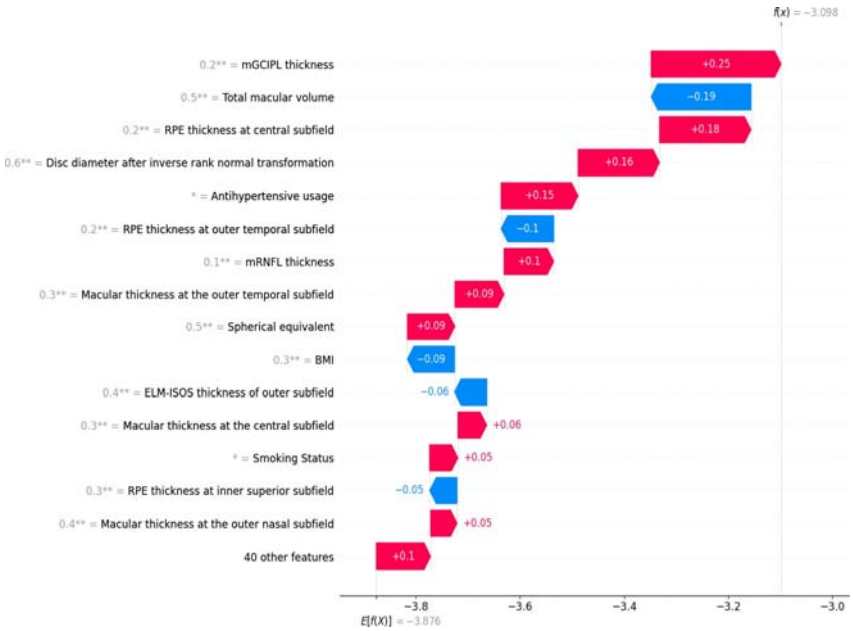


(a) AD correctly classified as AD

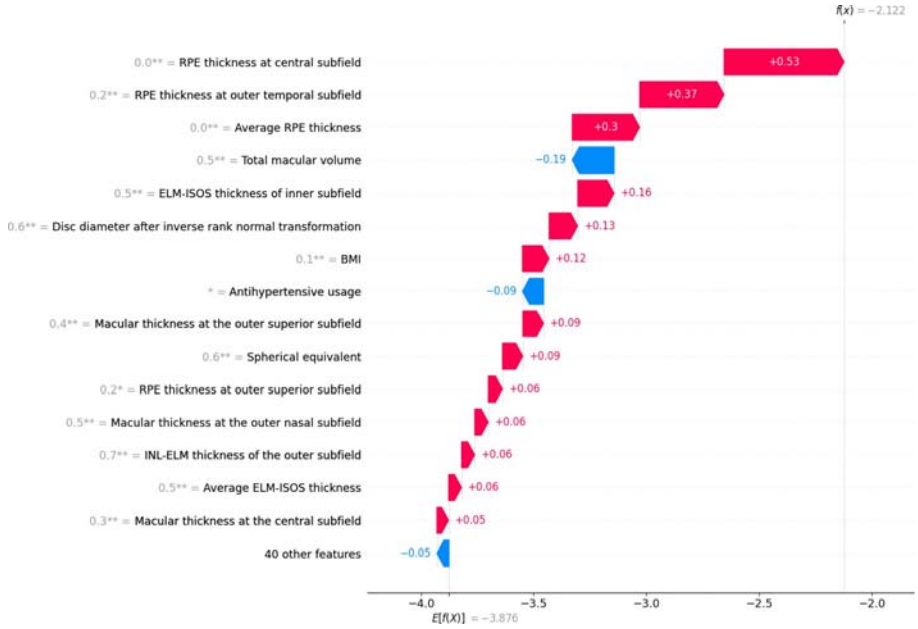


(b) CN correctly classified as CN

Figure 4.6: SHAP explainability results for correctly classified samples in the AD vs. Healthy cohort. Feature values were masked to protect participant privacy; retinal layer thicknesses were limited to the first significant digit (identical in all patients), and all demographic variables were completely removed.



(c) AD misclassified as CN



(d) CN misclassified as AD

Figure 4.7: SHAP explainability results for misclassified samples in the AD vs. Healthy cohort. Feature values were masked to protect participant privacy; retinal layer thicknesses were limited to the first significant digit (identical in all patients), and all demographic variables were completely removed.

-fied educational status as the most influential non-retinal variable.

To visualize risk relationships over time, mean Nelson–Aalen cumulative hazard plots were generated for the five datasets and a subset of the most informative retinal features (Figure 4.9). For mGCIPL thickness, mRNFL thickness, average RPE thickness, and vertical cup-to-disc ratio (regressed and transformed), participants in the "thin" group (values below the median) showed a higher cumulative hazard of Alzheimer's disease over time compared with the "thick" group. In contrast, for total macular volume, the "thick" group (values above the median) showed a higher cumulative hazard. These temporal patterns were consistent with the results of Cox proportional hazards analysis.

4.3.4 Feature Correlations in Age-Matched Datasets

To better understand how the top features relate to each other and to AD, we calculated the Spearman's correlation coefficients between the top 15 XGBoost and SHAP features identified by five runs of XGBoost with 5 different age-matched datasets. Figure 4.10 illustrates the correlation matrix for the age-matched datasets.

Two primary structural clusters were observed. The first was a macular cluster consisting of the total macular volume, macular thickness at the inner superior subfield, outer temporal subfield, outer superior subfield, INL-ELM thickness of the outer subfield, and mGCIPL thickness measurements. These features were highly correlated (r = 0.63–0.84). The mRNFL thickness was moderately correlated with total macular volume (r = 0.41).

The second cluster included RPE measures of the central, outer temporal, and inner nasal subfields that were moderately to highly correlated with each other (r = 0.36–0.76) and showed weak correlations with the total macular volume (r = -0.04 – 0.00).

In the age-matched dataset, age showed only weak correlations with both retinal and non-retinal features. Core retinal biomarkers, including mRNFL thickness (p = –0.04), mGCIPL thickness (p = –0.17), and total macular volume (p = –0.12), demonstrated weak negative associations with age. Similarly, the

macular subfields showed minor negative correlations, such as the inner superior thickness (p = –0.14), outer superior thickness (p = –0.11), and outer temporal thickness (p = –0.07). These results were consistent with slight age-related thinning, even among the matched groups. RPE measurements were largely independent of age, with very weak correlations observed in the central (p =–0.06), outer nasal (p =–0.04), and outer superior (p =–0.03) subfields. For non-retinal variables, most showed negligible relationships with age, including BMI (p = 0.01), antihypertensive use (p = 0.14), and educational status (p = 0.06). In contrast, the spherical equivalent (p = 0.25) showed a moderate positive correlation with age. This finding suggests a shift toward hyperopia with increasing age, even within the matched groups.

4.4 DISCUSSION

This study provides a combined machine learning and statistical evaluation of retinal features for early prediction of Alzheimer's disease (AD). Our results indicate that extended retinal measurements—beyond commonly used metrics such as mean retinal nerve fiber layer (mRNFL) and mean ganglion cell–inner plexiform layer (mGCIPL)—enhanced the predictive performance of AD classification models.

The study aimed to evaluate the predictive performance of the extreme gradient boosting (XGBoost) model, identify the most influential features, and assess the contribution of retinal changes while accounting for age. To handle severely imbalanced, high-dimensional data with missing values, we implemented appropriate preprocessing strategies. We applied two complementary explainability methods—XGBoost feature importance and SHapley Additive exPlanations (SHAP)—to quantify how individual features influenced model predictions. To support and validate these machine learning findings, we performed statistical analyses, including survival modeling and pairwise correlation, to link model-identified features to clinically interpretable outcomes.

Table 4.3: Comparison of AUC performance for dementia and Alzheimer's disease (AD) prediction across risk models. Best performer in bold.

Model	Dementia mAUC (± Std)	AD mAUC (± Std)
UKB-DRP (You et al., 2022)	0.848 ± 0.007	0.862 ± 0.015
CAIDE (Kivipelto et al., 2006)	0.705 ± 0.008	–
DRS (Walters et al., 2016)	0.752 ± 0.007	–
ANU-ADRI (Anstey et al., 2013)	–	0.584 ± 0.017
Ours (Unmatched)	0.892 ± 0.013	0.894 ± 0.018
Ours (Age-matched)	0.782 ± 0.021	0.763 ± 0.034

The early prediction of Dementia or Alzheimer's disease has been inves

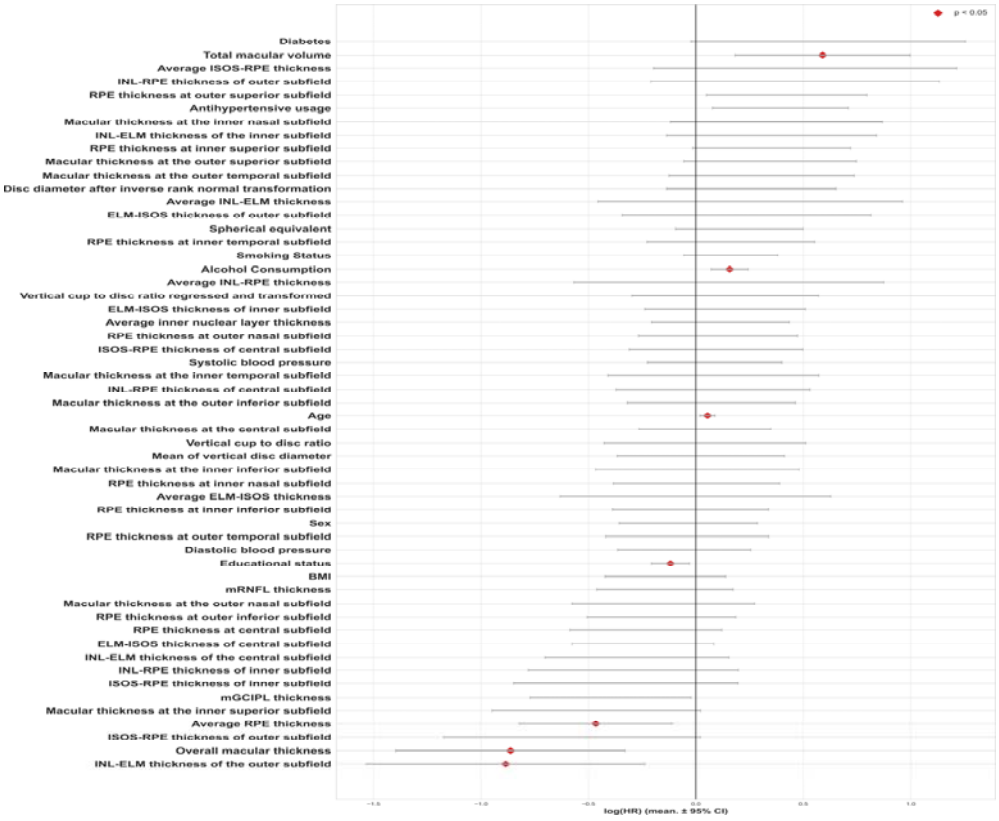
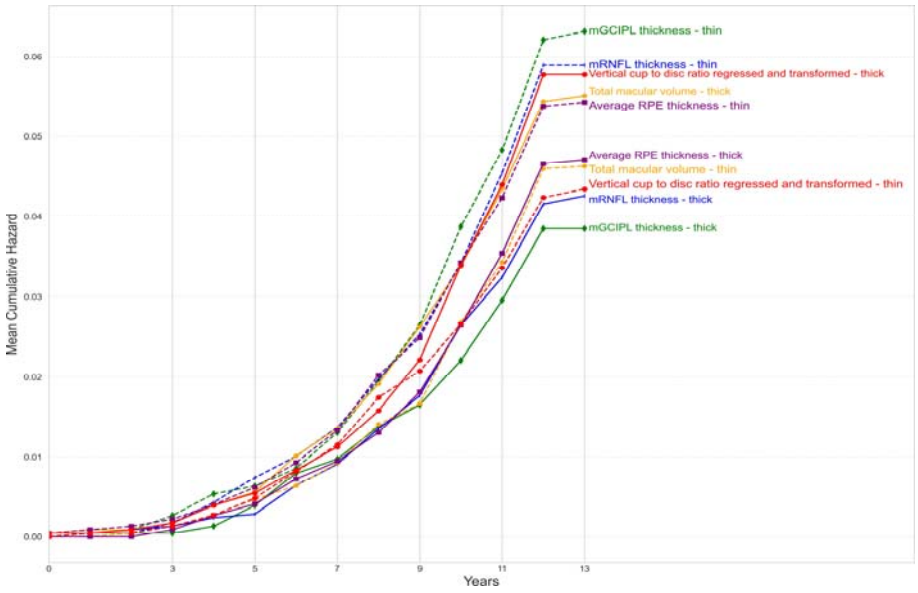


Figure 4.8: Forest plot of average Cox proportional hazards model coefficients (log hazard ratios) with 95% confidence intervals across five data subsets. Each point represents the mean log(HR) for a given retinal or clinical feature. Red diamonds indicate statistically significant associations ($p < 0.05$). Features with negative coefficients (left of zero) are associated with reduced hazards

(protective), whereas features with positive coefficients (right of zero) indicate increased hazards. Several macular and RPE thickness parameters, along with total macular volume and systemic factors (e.g., diabetes, alcohol consumption), show significant associations with Alzheimer's disease hazard.



(a) Mean Nelson–Aalen Cumulative Hazard Estimates

	After 0 Years	After 3 Years	After 5 Years	After 7 Years	After 9 Years	After 11 Years	After 13 Years
mGCIPL thickness - thin							
At risk	2280	2258	2224	2180	2107	1990	1795
Events	0	6	20	49	107	211	345
mGCIPL thickness - thick							
At risk	2282	2268	2243	2204	2152	2083	1896
Events	1	2	11	33	70	135	219
mRNFL thickness - thin							
At risk	2281	2261	2223	2178	2113	2008	1808
Events	0	4	21	52	108	206	332
mRNFL thickness - thick							
At risk	2282	2264	2244	2207	2146	2066	1883
Events	1	4	10	30	69	140	232
Total macular volume - thin							
At risk	2280	2260	2230	2193	2144	2049	1867
Events	0	3	13	34	71	146	245
Total macular volume - thick							
At risk	2282	2265	2237	2191	2115	2024	1824
Events	1	5	18	48	106	200	319
Average RPE thickness - thin							
At risk	2279	2256	2226	2179	2109	2018	1834
Events	1	6	20	50	105	197	314
Average RPE thickness - thick							
At risk	2284	2269	2241	2206	2150	2055	1856
Events	0	2	11	32	72	149	250
Vertical cup to disc ratio regressed and transformed - thin							
At risk	2241	2227	2201	2158	2093	2022	1849
Events	0	3	14	39	84	156	248
Vertical cup to disc ratio regressed and transformed - thick							
At risk	2322	2299	2267	2227	2166	2051	1841
Events	1	5	17	43	93	190	316

(b) Mean Nelson–Aalen Risk Table

Figure 4.9: Mean Nelson–Aalen cumulative hazard estimates for thin and thick

(< median) subgroups of selected features on the age-matched case using KNN imputation. (a) shows the mean cumulative hazard curves over time, while (b) presents the mean number of participants at risk and the mean number of events at different time points.

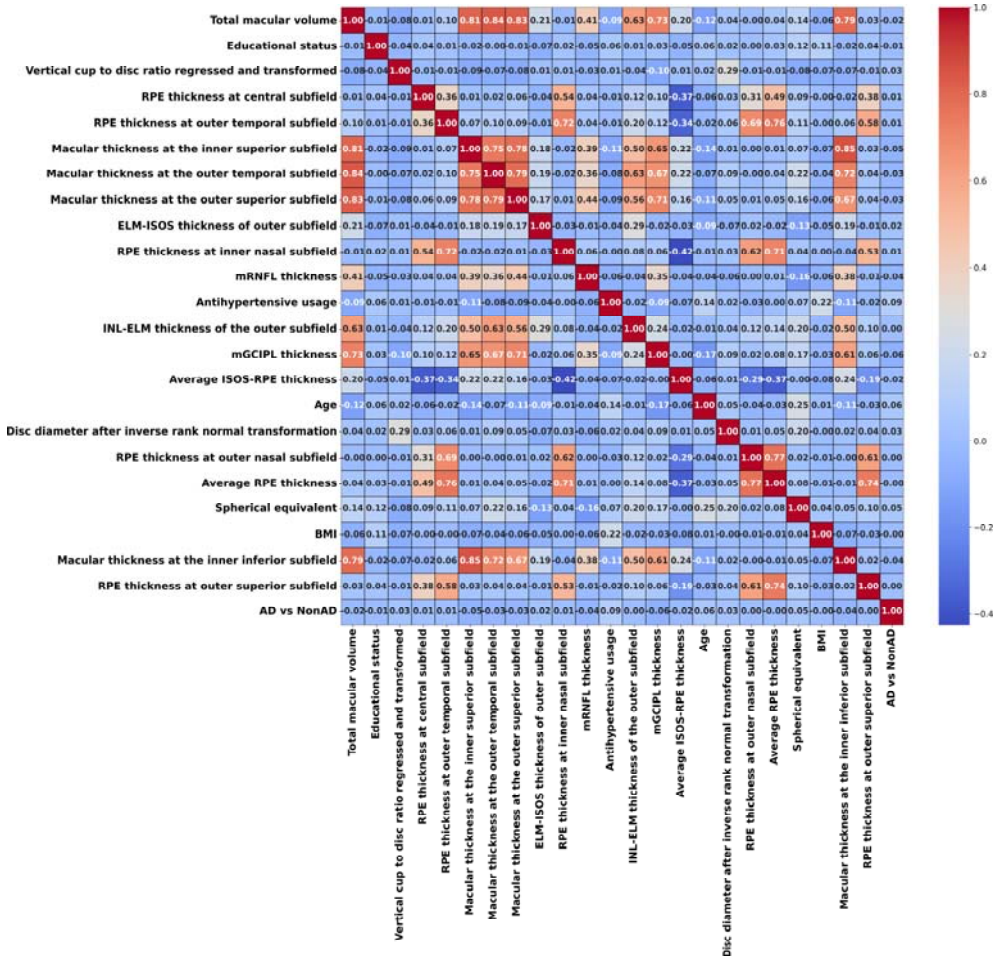


Figure 4.10: Mean feature correlations for AD vs non-AD classification for the age-matched case using the datasets with missing values that were directly used by the XGBoost model. Spearman's pairwise complete correlation is used to calculate the correlations.

-tigated in various contexts. The UK Biobank Dementia Risk Prediction (UKB-DRP) model (You et al., 2022) reported high discriminative accuracy for de-mentia (AUC 0.848 ± 0.007) and even higher accuracy for Alzheimer's

disease (AUC 0.862 ± 0.015) detection. Incorporating age and nine additional (automat-ically selected) features, the UKB-DRP outperformed established risk scores, including Cardiovascular Risk Factors, Aging, and Incidence of Dementia Risk Score (CAIDE (Kivipelto et al., 2006); AUC 0.705 ± 0.008), Dementia Risk Score (DRS (Walters et al., 2016); AUC 0.752 ± 0.007), and Australian National University Alzheimer's Disease Risk Index (ANU-ADRI) (Anstey et al., 2013); AUC 0.584 ± 0.017). The full feature set included age,

ApoE $\epsilon 4$, pair-matching time, leg fat percentage, number of medications taken, reaction time, peak expi-ratory flow, mother's age at death, long-standing illness, and mean corpuscular volume.

Table 4.3 summarizes the results of previous early prediction approaches, including risk scores. Using the UK Biobank, our method achieved the highest predictive accuracies. This study was designed to evaluate retinal features as potential biomarkers, with analyses performed on both unmatched and age-matched datasets to ensure their robustness. In practical early prediction settings, age must be included as a feature, as confirmed by our age-matched analyses. Age is a strong predictor of AD and dementia; however, experiments on unmatched datasets distort the age distribution and lead to overly optimistic estimates of predictive performance.

However, differences in datasets and sampling strategies restrict direct comparisons across studies. Incorporating additional variables, such as the ApoE $\epsilon 4$ genotype, may further improve our results. Nonetheless, OCT imaging remains a noninvasive and cost-effective alternative to genetic testing.

A critical challenge addressed in this study was the handling of incomplete feature data, which are common in large-scale biobank studies. We compared imputation strategies and found that XGBoost's native handling of missing data outperformed both KNN and Random imputation methods.

Table 4.4 summarizes the results of our feature importance and explainability experiments, along with their significance in Cox proportional hazards modeling, Nelson-Aalen cumulative hazard, correlations with age and AD, and

results from the t-test group comparison.

Our combined interpretability and statistical analyses in age-matched datasets highlighted several retinal features that consistently contributed to Alzheimer's disease (AD) classification beyond the traditionally reported mRNFL and mG-CIPL metrics. Although XGBoost and SHAP produced different full top-15 rankings, they agreed on the top four features: total macular volume, educa-tional status, vertical cup-to-disc ratio (regressed and transformed), and central RPE thickness. These overlapping features suggest that retinal structural changes can be detected in patients with AD.

Total macular volume emerged as the top-ranked feature in XGBoost and was ranked third in SHAP, highlighting its strong and consistent contribution across the different explainability approaches. It also showed statistical significance in both Cox proportional hazards modeling ($p = 0.044$) and t-test group comparisons ($p = 0.006$). These results reinforce its association with Alzheimer's disease (AD). Interestingly, the Nelson–Aalen cumulative hazard and Cox analyses suggested that greater macular volume was linked to a higher AD risk, which contrasted with the general pattern of thinning-related features and the direction implied by the t-test results.

Several other macular subfield thicknesses also appeared in the top ranks: inner superior (rank 6), outer temporal (rank 7), and outer superior —with strong significance in the t-tests ($p < 0.005$). These subfield measures were highly correlated with total macular volume ($r = 0.81$–0.84). They add predictive value individually, but when combined with macular volume, they may provide overlapping information.

mGCIPL thickness (rank 14) correlated strongly with total macular volume ($r = 0.73$). In contrast, mRNFL thickness (rank 11) had a moderate correlation ($r = 0.41$). This means mGCIPL is more linked to macular volume, while mRNFL reflects other structural changes.

The introduction of total macular volume into the feature set appeared to lower the relative importance of traditional retinal neurodegeneration metrics (mGCIPL and mRNFL) in predicting visual function. In this analysis, both

mGCIPL and mRNFL were ranked outside the top 10 in XGBoost and SHAP, whereas previous studies often reported them as top predictors when macular volume was excluded. This suggests that the total macular volume acts as an umbrella metric capturing a significant portion of the variance explained by these standard features, potentially reshaping feature prioritization in multimodal AD risk models.

Educational status ranked second in the XGBoost and first in the SHAP analyses. These results place it among the most influential features for both explainability approaches. It was also statistically significant in Cox proportional hazards modeling ($p = 0.011$) and t-tests ($p = 0.030$), indicating a measurable association with AD risk. Unlike structural retinal measurements, educational status is a demographic factor that reflects the concept of cognitive reserve. Its high ranking in both models demonstrates the importance of including key de-mographic variables and ocular biomarkers to improve predictive performance.

The vertical cup-to-disc ratio (VCDR) regressed and transformed ranked within the top four predictors in both XGBoost (3rd) and SHAP (2nd) analyses, showed significant group differences in t-tests ($p = 0.012$), but was not significant in the Cox proportional hazards model ($p = 0.371$). While most often associated with glaucoma, increased VCDR has also been reported in Alzheimer's disease (AD) cohorts (Chan et al., 2019; den Haan et al., 2018), suggesting that alterations in the optic nerve head may be involved in neurodegenerative processes such as AD. VCDR performed strongly and consistently in both interpretability rankings and statistical tests. These results suggest that VCDR may be a novel retinal biomarker of AD. It should be further tested in independent groups and studies that follow participants over time to confirm its validity.

The RPE thickness in the central subfield ranked fourth in both the XGBoost and SHAP models. Although it was not significant in the Cox proportional hazards model ($p = 0.380$) or in the t-tests ($p = 0.541$), its position among the top-ranked predictors suggests its potential relevance to AD-related retinal changes. Other RPE-related measures, such as thickness at the outer temporal

subfield (XGBoost rank 5, SHAP rank 8) and average RPE thickness (SHAP rank 9, significant in Cox at $p = 0.044$), also appeared within the top predictive features, reinforcing the role of RPE morphology in the modeling framework. RPE thickness in the central subfield had almost no correlation with total macular volume ($r = 0.01$). This suggests that RPE measures, particularly those from the central and peripheral subfields, may be independent and promising biomarkers for AD. They could add useful information to the macular and inner retinal measurements in the prediction models.

Several other features in the ranking fall outside the top-4 predictors and are not directly related to macular or RPE morphology. Among these, mRNFL thickness (XGBoost rank 11) and mGCIPL thickness (rank 14) are standard retinal neurodegeneration metrics that showed significant group differences in t-tests ($p < 0.001$ for both) but were deprioritized in the presence of total macular volume and related measures, likely because of the shared variance. Antihypertensive usage (rank 12) also reached significance in t-tests ($p < 0.001$), reflect-ing systemic vascular factors that may influence the risk of AD independently of ocular structure. Inner retinal segmentation measures, such as INL–ELM thickness of the outer subfield (rank 13, significant in Cox at $p = 0.010$) and ELM–ISOS thickness of the outer subfield (rank 9), appeared as mid-ranked predictors, suggesting that photoreceptor-adjacent and middle retinal layers may contribute complementary signals to the AD classification. Additional non-structural or less directly interpretable predictors, including spherical equivalent (SHAP rank 10), BMI (SHAP rank 12), and disc diameter after inverse rank transformation (SHAP rank 6). These features reflect systemic or anatomical co-variates rather than primary disease-driven retinal changes. They are unlikely to work as biomarkers on their own, but adding them to a multivariate model can improve accuracy by capturing risk factors beyond macular and RPE changes.

We also examined local SHAP explanations for a few correctly classified and misclassified cases. The waterfall plots illustrate how retinal and systemic features can shift an individual prediction toward or away from AD, occasionally

deviating from global feature patterns such that key features—such as macular volume, RPE thickness, and disc diameter—are important at the individual level, though their influence varies across patients. This can be usefull in practice, for example, a clinician reviewing a patient's waterfall plot can see that reduced macular volume, increased RPE thickness at the central subfield, and a smaller disc diameter collectively drive the model's prediction toward AD. Such a visualization allows the clinician to understand which biomarkers are contributing most to a patient's predicted risk and to discuss these findings in a personalized context with the patient. This provides an explanation for prediction scores in a clinically interpretable manner, improving transparency, clarifying model behavior, and highlighting both consistent and patient-specific AD risk classification.

This study had several limitations. The primary limitation was the incomplete nature of the dataset, as not all retinal features were available for every participant. Although we used XGBoost-based imputation to address this, residual uncertainty may remain, and future studies with more comprehensive data are required to confirm these findings. The analysis reflects the UK population and includes individuals of mixed but predominantly European ancestry; therefore, the findings may not be generalizable to other populations. Survival analysis provides limited information on longitudinal changes, and diagnosis records and dates may be imprecise, as they could reflect occasional visits, hospital encounters for related discomfort, or unrelated symptoms. Age matching reduced confounding factors; however, other factors, such as vascular or ocular condi-tions, may have influenced the results. The dataset was also imbalanced, with far fewer dementia and AD cases than cognitively normal participants, which may have affected the model training despite the use of balanced accuracy metrics. Furthermore, because the study focused on retinal biomarkers, we did not ex-plore alternative machine learning models such as LightGBM or CatBoost, nor did we incorporate genetic features or genetic risk scores, including ApoE $\epsilon 4$, which are established predictors of dementia and AD. Future studies should val-idate these findings in independent, more diverse populations and models that integrate retinal, genetic, and clinical features.

4.5 CONCLUSION

In this study, we explored the potential of high-dimensional retinal features from UK Biobank OCT images for classifying early Alzheimer's disease using gradient-boosting machine learning models. By integrating extended retinal metrics beyond the standard RNFL and GCIPL layers, implementing robust strategies for handling incomplete feature data, and applying age matching to control for confounding factors, we achieved strong classification performance and identified key retinal features with potential clinical relevance.

In AD vs. non-AD classification, the unmatched dataset gave high performance (AUC 0.894 ± 0.018). Accuracy declined in the age-matched dataset (AUC 0.763 ± 0.034), since age was no longer a strong driver. Retinal features still carried predictive value after age effects were removed.

Building on these findings, the subsequent phase of this research seeks to determine if similar or superior conclusions can be derived directly from raw OCT imagery using Deep Learning. While the gradient-boosting models demonstrated the utility of pre-defined retinal metrics, Deep Learning offers the distinct advantage of learning latent, complex representations directly from high-dimensional scans without relying on explicit feature extraction. The following chapters will investigate this hypothesis, exploring whether end-to-end deep learning architectures can independently identify Alzheimer's-associated biomarkers and validate the predictive power of the retina in an image-based context.

Table 4.4: Overlap and ranking of retinal and non-retinal features identified by XGBoost and SHAP analyses, with corresponding Nelson–Aalen cumulative hazard plots, Cox proportional hazards model significance, t-test group comparison results, and correlation with Alzheimer's disease (AD) and age. The rankings for each machine learning method are shown. Directional markers indicate interpretation: ↑means higher values of the feature are associated with increased AD risk, ↓means lower values are associated with increased AD risk, and → indicates no conclusive result from that method. Statistical measures highlight features with potential as novel retinal biomarkers beyond the mRNFL and mGCIPL.

Retinal Feature	XGBoost Rank	SHAP Rank	Nelson-Aalen Cum. Hazard	Cox Significance	t-tests Significance	Correlation with AD	Correlation with Age
Total Macular Volume	1	3 ↓	thick↑	∗0.044↑	∗0.006↓	-0.02↓	-0.12
Educational Status	2	1 ↓		∗0.011↓	*0.030 ↓	-0.01↓	0.06
Vert. cup to disc ratio regressed and transformed	3	2 ↑	thick↑	0.371↑	∗0.012↑	0.03↑	0.02
RPE thickness in the central subfield	4	4↓		0.380↓	0.541 ↓	0.01↑	-0.06
RPE thickness at the outer temporal subfield	5	8↑		0.388↓	0.332↓	0.01↑	-0.02
Macular thickness at the inner superior subfield	6	28		0.068↓	∗ < 0.001↓	-0.05↓	-0.14
Macular thickness at the outer temporal subfield	7	14 ↓		0.181↑	∗0.004↓	-0.03↓	-0.07
Macular thickness at the outer superior subfield	8	31		0.538↑	∗ < 0.001↓	-0.03↓	-0.11
ELM-ISOS thickness of outer subfield	9	39		0.4312↑	0.098↓	0.02↑	-0.09
RPE thickness at inner nasal subfield	10	26		0.800↑	0.554↓	0.01↑	-0.01
mRNFL thickness	11	16	thin↓	0.410↓	∗ < 0.001↓	-0.04↓	-0.04
Antihypertensive usage	12	22		0.211↑	∗ < 0.001	0.09↑	0.14
INL-ELM thickness of the outer subfield	13	36		*0.010↓	0.764↑	0.00 →	-0.01
mGCIPL Thickness	14	11 ↓	thin↓	0.069↓	∗ < 0.001↓	-0.06↓	-0.17
Average ISOS-RPE thickness	15	40		0.179↑	0.345↓	-0.02↓	-0.06
Age	33	5 ↑		∗0.003↑	∗ < 0.001↑	0.06↑	1.00
Disc diam. after inverse rank normal transform.	47	6 ↑		0.223↑	∗0.027↑	0.03↑	0.05
RPE thickness at outer nasal subfield	49	7 ↓		0.374	0.160↓	0.00 →	-0.04
Average RPE thickness	39	9 ↓	thin↓	*0.044↑	0.268↓	−0.00 →	-0.03
Spherical equivalent	43	10 ↑		0.204↑	0.91↑	0.05↑	0.25
BMI	30	12 ↓		0.339↓	0.924↓	−0.00 →	0.01
Macular thickness at the inner inferior subfield	52	13 ↓		0.778↑	∗ < 0.001↓	-0.04↓	-0.11
RPE thickness at the outer superior subfield	53	15 ↓		0.111↑	0.059↓	0.00 →	-0.03

CHAPTER 5

5. EARLY AD DETECTION FROM RETINAL OCT B-SCANS

This chapter explains our work on the Deep Learning based early Alzheimer detection using OCT B-scnas. We will first explain the OCT image cohort and dataset then the method, followed by our results.

5.1 OCT IMAGING STUDY DATASET

We used the selected data from UK Biobank described in Chapter 3. The OCT scans in the UK Biobank are in FDS and FDA formats. FDS bulk files contain the entire, 16-bit raw OCT volume for both the left and right eyes, providing the complete dynamic range and unprocessed scan data necessary for custom image processing or resegmentation. In contrast, the FDA bulk files are an 8-bit downsampled representation derived directly from FDS volumes; they include four segmentation layers but do not permit recovery of the original 16-bit voxel intensities. The differences between the FDA and FDS image formats are shown in Figure 5.1 (a) and (b).

This study utilized a targeted **4-year window** by selecting participants diagnosed with AD within four years of their baseline assessment. This specific timeframe was chosen to isolate and investigate biomarkers associated with MCI and early neurodegenerative changes that precede a formal clinical diagnosis.

To ensure that any observed anatomical differences were driven by disease pathology rather than confounding variables, the AD group was matched by age, sex, eye, and instance with a Healthy Control group ($N = 30$). We utilized both eyes of the AD patients if both met the eligibility quality standards. For each qualifying B-scan in the AD group, a custom candidate pool was generated from the extensive UK Biobank healthy cohort. This pool consisted of thousands of control eyes that strictly matched the specific age, sex, eye orientation (left or right), and scan instance of the target AD eye. From this filtered subset, a single control scan was randomly selected to maintain a balanced dataset and mitigate selection bias.

During the course of our study, one AD patient withdrew their consent from the UK Biobank; accordingly, their data (2 B-scans) were excluded, and all statistical tests were rerun using the final dataset (28 AD B-scans vs. 30 control B-scans). Table 5.1 illustrates the demographic alignment achieved through our selection process, highlighting the transition from the initial large-scale disparities in the original UK Biobank cohort to the precisely balanced study groups.

Table 5.1: Comparison of Demographic Characteristics between the Original and Curated Datasets.

Group	**N**		**Age (Years)**		**Sex (Woman %)**	
	Full	4-year	Full	4-year	Full	4-year
AD	223	19	65.78	66.79	46%	21%
CN	43,434	30	57.09	66.20	53%	23%

The analysis in Table 5.2 confirmed the efficacy of this strategy, showing no statistically significant differences between the groups regarding **Age** ($p = 0.657$) or **Sex distribution** ($p = 1.000$).

The analysis in Table 5.2 confirmed the efficacy of this curation strategy; there were **no statistically significant differences** between the cohorts regarding **Age** ($p = 0.657$) or **Sex distribution** ($p = 1.000$). However, a significant difference was observed in **Educational Status** ($p = 0.008$), with the AD group exhibiting a higher prevalence of lower education levels. Structural analysis revealed a statistically significant thinning of the **mGCIPL** in the AD group ($69.18 \pm 6.38\ \mu$m) compared to the control group ($72.95 \pm 6.29\ \mu$m, $p = 0.050$), while, the **mRNFL** did not show significant thinning ($p = 0.629$).

The extended analysis in Table 5.3 provided a more granular view. The **Total Macular Volume** was significantly lower in the AD group (7.64 mm^3 vs. 7.96 mm^3, $p = 0.028$). Significant thinning was heavily localized to specific subfields of the macula, particularly the Inner Superior ($p = 0.002$), Inner Temporal ($p = 0.028$), and Outer Superior ($p = 0.008$) sectors. Other retinal layers,

such as the Inner Nuclear Layer and the Retinal Pigment Epithelium, remained structurally stable between groups ($p > 0.05$).

Table 5.2: Demographic and Eye-related Features Analysis for 4-Year Dataset

Feature	**Total** (N=49)	**AD** (N=19)	**Healthy** (N=30)	**p-value**
Age (years), mean ± SD	66.43 ± 4.38	66.79 ± 4.63	66.20 ± 4.29	0.657
Sex (count, %)				1.000
0	11 (22%)	4 (21%)	7 (23%)	
1	38 (78%)	15 (79%)	23 (77%)	
Educational status (count, %)				**0.008***
1	14 (39%)	6 (55%)	8 (32%)	
2	4 (11%)	4 (36%)	0 (0%)	
3	8 (22%)	1 (9%)	7 (28%)	
4	2 (6%)	0 (0%)	2 (8%)	
5	4 (11%)	0 (0%)	4 (16%)	
6	4 (11%)	0 (0%)	4 (16%)	
Diabetes (count, %)				0.276
Without	36 (90%)	15 (100%)	21 (84%)	
With	4 (10%)	0 (0%)	4 (16%)	
Spherical equivalent, median (IQR)	1.35 (0.60-1.67)	1.30 (0.41-1.50)	1.47 (0.88-1.81)	0.226
Systolic BP (mmHg), mean ± SD	141.56 ± 14.03	145.79 ± 14.86	138.88 ± 13.02	0.106
Diastolic BP (mmHg), mean ± SD	82.34 ± 9.94	83.45 ± 13.24	81.63 ± 7.32	0.589
Antihypertensive use (count, %)				1.000
No	21 (68%)	7 (64%)	14 (70%)	
Yes	10 (32%)	4 (36%)	6 (30%)	
Alcohol Consumption (count, %)				0.143
1	10 (20%)	3 (16%)	7 (23%)	
2	14 (29%)	4 (21%)	10 (33%)	
3	11 (22%)	3 (16%)	8 (27%)	
4	6 (12%)	4 (21%)	2 (7%)	
5	5 (10%)	2 (11%)	3 (10%)	
6	3 (6%)	3 (16%)	0 (0%)	
Smoking Status (count, %)				0.189
0	23 (47%)	8 (42%)	15 (50%)	
1	24 (49%)	9 (47%)	15 (50%)	
2	2 (4%)	2 (11%)	0 (0%)	
BMI (kg/m^2), mean ± SD	26.09 ± 3.63	26.31 ± 4.56	25.95 ± 2.99	0.761
Retinal thickness indices				
RNFL thickness (μm)	27.72 ± 4.52	28.16 ± 5.64	27.45 ± 3.73	0.629
GCIPL thickness (μm)	71.49 ± 6.53	69.18 ± 6.38	72.95 ± 6.29	**0.050***

Note: Percentages are based on valid responses.

* Statistically significant difference ($p \leq 0.05$)

5.2 METHODS

We used the OCTExplorer software, which utilizes the Iowa Reference Algorithms (Retinal Image Analysis Lab, Iowa Institute for Biomedical Imaging, Iowa City, IA) (Abramoff et al., 2010; Garvin et al., 2009) for the segmentation of the fds files (Figure 5.1 (c)).

This study focused on ImageNet-pretrained 2D models. Therefore, the middle slice (B-scans) of the raw image volumes (128 × 512 × 650) was selected as the input. The OCT image was rectified using the bottom contour, as shown

in Figure 5.1 (c).

In Optical Coherence Tomography (OCT) imaging, rectification is a critical preprocessing step used to correct the natural curvature of the retina and the "tilting" effects caused by the scanning process. This procedure creates a flat, standardized baseline, which is essential for obtaining accurate layer thickness measurements.

Table 5.3: Extended Retinal Features Analysis for 4-Year Dataset

Feature	**Total** (N=49)	**AD** (N=19)	**Healthy** (N=30)	p-value
INL thickness				
INL thickness (μm)	32.79 ± 2.07	32.42 ± 2.05	33.03 ± 2.08	0.315
INL-ELM indices (μm)				
Average thickness	81.26 ± 6.47	80.82 ± 6.89	81.53 ± 6.29	0.720
Central subfield	108.11 ± 9.28	107.23 ± 11.04	108.66 ± 8.13	0.630
Inner subfield	94.04 ± 7.23	93.43 ± 8.21	94.44 ± 6.66	0.655
Outer subfield	76.39 ± 6.54	75.91 ± 6.94	76.70 ± 6.38	0.691
ELM-ISOS indices (μm)				
Average thickness	23.56 ± 1.58	23.47 ± 1.40	23.62 ± 1.70	0.730
Central subfield	28.18 ± 1.57	27.76 ± 1.74	28.46 ± 1.42	0.150
Inner subfield	24.44 ± 1.28	24.51 ± 1.23	24.39 ± 1.33	0.740
Outer subfield	23.14 ± 1.77	23.02 ± 1.59	23.22 ± 1.90	0.706
ISOS-RPE indices (μm)				
Average thickness	38.27 ± 3.81	37.29 ± 3.80	38.90 ± 3.74	0.154
Central subfield	42.71 ± 5.43	41.33 ± 4.94	43.58 ± 5.63	0.149
Inner subfield	39.24 ± 4.39	38.04 ± 3.97	40.00 ± 4.53	0.119
Outer subfield	37.82 ± 3.67	36.92 ± 3.83	38.40 ± 3.50	0.182
INL-RPE indices (μm)				
Average thickness	143.09 ± 7.73	141.58 ± 8.04	144.05 ± 7.51	0.289
Central subfield	179.00 ± 12.29	176.31 ± 13.78	180.70 ± 11.15	0.252
Inner subfield	157.72 ± 8.91	155.98 ± 9.60	158.82 ± 8.42	0.297
Outer subfield	137.36 ± 7.67	135.85 ± 7.96	138.31 ± 7.47	0.288
RPE indices (μm)				
Overall thickness	24.99 ± 2.21	25.07 ± 2.78	24.95 ± 1.82	0.868
Central subfield	24.02 ± 3.95	26.13 ± 4.73	23.24 ± 3.44	0.176
Inner inferior	23.73 ± 3.72	25.91 ± 4.80	22.93 ± 3.01	0.164
Inner nasal	26.80 ± 3.79	29.46 ± 3.90	25.82 ± 3.34	0.055
Inner superior	22.71 ± 3.37	24.86 ± 4.74	21.91 ± 2.41	0.159
Inner temporal	24.91 ± 3.28	25.64 ± 4.18	24.64 ± 2.97	0.574
Outer inferior	23.84 ± 1.91	25.14 ± 2.34	23.35 ± 1.52	0.098
Outer nasal	27.39 ± 4.19	28.06 ± 1.44	27.15 ± 4.85	0.466
Outer superior	24.43 ± 2.25	24.82 ± 3.07	24.28 ± 1.95	0.679
Outer temporal	25.77 ± 3.37	27.13 ± 2.78	25.26 ± 3.49	0.180
Macular thickness indices (μm)				
Overall thickness	275.10 ± 12.32	271.34 ± 11.56	277.49 ± 12.37	0.085
Central subfield	264.92 ± 22.73	261.45 ± 23.76	266.87 ± 22.40	0.491
Inner inferior	307.82 ± 16.61	300.55 ± 15.94	311.89 ± 15.84	**0.042***
Inner nasal	315.08 ± 19.65	307.14 ± 19.85	319.53 ± 18.46	0.067
Inner superior	307.90 ± 19.66	295.33 ± 17.76	314.94 ± 17.25	**0.002***
Inner temporal	299.58 ± 16.95	291.66 ± 15.74	304.01 ± 16.24	**0.028***
Outer inferior	262.63 ± 14.68	259.45 ± 14.60	264.41 ± 14.71	0.319
Outer nasal	284.42 ± 15.45	278.00 ± 15.93	288.01 ± 14.25	0.062
Outer superior	265.41 ± 14.14	257.65 ± 12.34	269.76 ± 13.39	**0.008***
Outer temporal	255.07 ± 11.90	249.58 ± 11.28	258.14 ± 11.31	**0.031***
Total Macular Volume (mm^3)				
Total macular volume	7.88 ± 0.36	7.64 ± 0.24	7.96 ± 0.37	**0.028***
Disc Diameter indices				
Mean vertical disc diameter	126.86 ± 13.02	129.40 ± 11.04	125.11 ± 14.17	0.266
Transformed diameter	0.27 ± 0.89	0.38 ± 0.73	0.19 ± 1.00	0.474

Table 5.3 (Continuing) Extended Retinal Features Analysis for 4-Year Datase

Vertical cup to disc ratio indices				
Vertical cup to disc ratio	0.36 ± 0.21	0.36 ± 0.24	0.36 ± 0.18	0.952
Transformed ratio	0.09 ± 1.11	0.02 ± 1.32	0.15 ± 0.96	0.714

Note: All values are Mean ± Standard Deviation. p-values are from independent t-tests.

* Statistically significant difference ($p \leq 0.05$).

The process of rectifying an OCT B-scan using the bottom contour involves the following steps:

- Bottom Contor OB_RPE is selected.
- Due to the eye's spherical shape and potential axial misalignment, this contour usually appears as a curve. We calculated the vertical offset (pixel distance) of each point on this contour relative to a flat horizontal line (the reference baseline).
- Each vertical column (A-scan) of the image is shifted up or down based on its calculated offset. For example, if the bottom contour at a specific pixel is 5 pixels too low, that entire column of pixels is shifted upward by 5 pixels.
- Once the shifts are applied, the bottom contour becomes a perfectly straight, horizontal line. All retinal layers above it—such as the **mGCIPL** and **mRNFL**—are also flattened, preserving their structural relationships against the bottom contour.
- Post-rectification, the images were not globally aligned to a fixed vertical center. This step was intentionally omitted because our training pipeline incorporates vertical shifting augmentation.

This rectification process was necessary for our study to ensure that the measurements of the AD group ($N = 28$) and Healthy Controls ($N = 30$) were derived from a consistent geometric baseline. It guarantees that the observed thinning corresponds to actual tissue loss rather than artifacts arising from retinal curvature.

The areas above the top contour and below the bottom contour were re-

moved to eliminate noise at the top and bottom of the scans. An additional layer-masked image was generated (Figure 5.1 (d)). Finally, all pre-processed images of size 512 × 650 were cropped to 512 × 512.

To adapt the single-channel OCT images to the three-channel input requirements of the pretrained models, we constructed a composite 3-channel representation for each sample, as shown in Figure 5.2. Specifically, the original grayscale OCT B-scan was assigned to the first channel of the network. The second channel contained a layer-masked version of the image, where each retinal layer was selectively enhanced to highlight the structural regions of interest (Eren et al., 2024). The third channel consisted of a binary image encoding the retinal layer contours, providing explicit anatomical boundaries as auxiliary input. This multichannel representation was designed to enrich the input with both intensity.

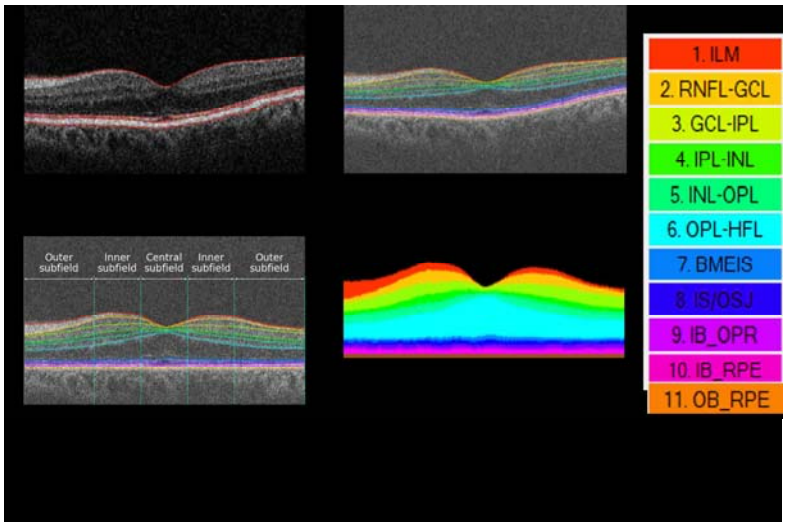


Figure 5.1: Overview of the preprocessing pipeline and retinal layer annotations in OCT B-scans. (a) Original grayscale OCT scan with inner and outer retinal boundaries. (b) Retinal layer segmentation with 11 color-coded contours representing the anatomical boundaries. (c) Alternate view of the same segmented B-scan for visualization consistency. (d) Pixel-wise retinal layer mask used as input for the second channel of the model. The legend on the right maps each color to a specific retinal layer, from the inner limiting membrane (ILM) to the outer boundary of the retinal pigment epithelium (OB_RPE). This

multichannel representation encodes both intensity and anatomical structure, providing richer input for deep learning models. *Input OCT B-scan was used with permission from the UK Biobank under Application Number 82266.*

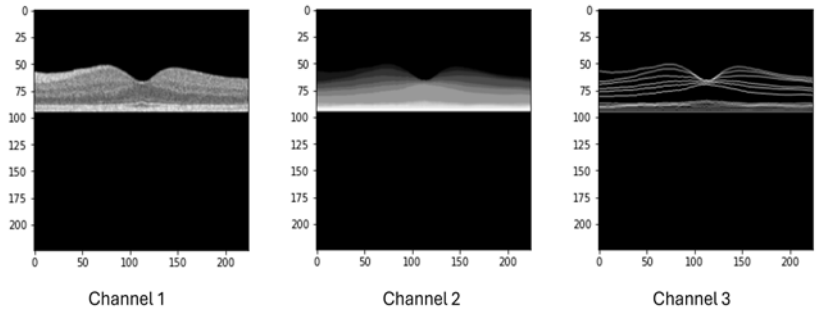


Figure 5.2: Composite RGB representation of a single OCT B-scan used as model input. (a) First channel: original grayscale OCT image. (b) The second channel: the layer-masked version, where the retinal layers are selectively enhanced to highlight the structural features. (c) Third channel: binary retinal layer contours providing anatomical boundary information. *The input OCT B-scan was used with permission from the UK Biobank under Application Number 82266.*

based and structural information. This facilitates improved feature extraction in downstream convolutional-based architectures.

5.2.1 Training Augmentation

Deep learning models are prone to overfitting, particularly when they are trained on small datasets. To mitigate this issue and improve generalization, we applied extensive augmentation strategies to artificially expand the training set and introduce greater variability, thereby enhancing the robustness of the model and reducing the risk of overfitting. During training, the augmentation techniques were randomly selected and applied one at a time. Custom image augmentations, such as flipping, horizontal and vertical shift, center crop (excluding shear and rotation in rectified images because of the nature of these images), were applied to all channels. In addition, OCT-specific augmentations such as

occlusion, contrast adjustment, artificial vascular patterns, and noise addition were employed only in the first channel of the model.

Training Models and Parameters

To investigate the effectiveness of deep learning for OCT-based classification, we employed a pretrained convolutional neural network (CNN) and transformer architecture. Specifically, we used ResNet (depths of 18, 34, 50, and 101) and VGG (depths of 8 and 11).

In addition to these standard architectures, we also tested a domain-specific RETFound (Zhou et al., 2023), a vision transformer pretrained on a large dataset (over a million) of OCT and fundus images. RETFound has demonstrated strong performance in retinal imaging tasks (Nazlı et al., 2025) and was therefore in-cluded for comparative evaluation in our Alzheimer's disease classification frame-work. We implemented two distinct training strategies: **RETFound-S**, trained using three identical mid-scan OCT images as input, and **RETFound-C**, trained end-to-end using the composite 3-channel representation described in Figure 5.2.

We also developed and trained a custom convolutional model to serve as a baseline and explored architectural variations tailored to the specific characteristics of OCT data.

All models were trained using a batch size of 4, which was selected to maintain stable gradient updates while accommodating the constraints of small dataset size. We experimented with a range of learning rates (0.001, 0.0001, and 0.000027) to explore the optimal convergence behavior across different model architectures. The AdamW optimizer was used for optimization in all experiments. We employed extensive image augmentation techniques (as detailed in the previous section) to address the risk of overfitting.

To adapt the pretrained backbones for binary classification under low-data conditions, we appended a lightweight classification head consisting of a fully connected layer (from the feature dimension to 64 units), Layer Normalization, followed by a ReLU activation, dropout ($p = 0.4$), and a final linear layer pro-

jecting to two output classes. This minimal regularization structure was selected to balance the expressiveness and overfitting control.

We employed a year-weighted loss function (Nazlı et al., 2025), assigning higher importance to samples temporally closer to the clinical diagnosis of Alzheimer's disease to recognize disease progression dynamics.

Due to the extensive data augmentation and regularization techniques applied to mitigate overfitting, the training and validation accuracy exhibited considerable fluctuations, making it difficult to identify an appropriate stopping epoch for the model. We trained the models for 100 epochs and then applied Stochastic Weight Averaging (SWA) (Izmailov et al., 2019) starting from epoch 80 to stabilize the training and improve convergence.

5.2.2 Validation Comparing Results

We employed a nested cross-validation strategy (Zhong et al., 2023) for a robust and unbiased performance estimation. The data were split into five outer folds, each of which served as a held-out test set. Within the remaining training data, three-fold inner cross-validation was used to tune the hyperparameters. The selected model was evaluated using the corresponding test fold. The predictions from the five outer folds were pooled to compute a single AUC per run. This procedure was repeated five times with different random splits, yielding five AUC values per model (5 runs × 5 outer folds × 3 inner folds).

We applied this procedure separately to ResNet (18, 34, 50, and 101), VGG (8 and 11), RetFound (S and C), and the CNN baseline models. For each architecture, we calculated the mean AUC across five runs and selected the top-performing variant within each family. These best variants were then compared using a calibrated paired t-test (Bouckaert, 2003), with adjusted degrees of freedom to account for dependence on repeated cross-validation.

Explainability

We chose the Grad-CAM method to explain our best-performing model. Instead of using the standard image overlay format, we used 10 OCT-specific layers and the central subfield of the macula. We applied a 0.8 threshold to the Grad-CAM results to generate a highly focused explainability image. The con-

tours of the OCT scan were overlaid on this image. To evaluate the class-level performance, we used the Intersection Over Union (IoU), Dice Score, and Filling Ratio on each layer, as detailed in the study by Nazlı et al., 2025. Finally, we added the center subfield region of the OCT b-scans, where the overlaid Grad-CAM results were highlighted.

Running with another sample set

We generated another training dataset with different age, sex, and instance-matching CNs while keeping the AD population the same. For the best-performing model, the same experiments were performed to determine whether the results were consistent with different datasets.

Ablation Studies

The model trained with only one of the channels was replicated for all three channels to determine the impact of the segmentation and contour information on the training results.

5.3 RESULTS

We evaluated multiple deep learning models using nested cross-validation, 5 outer folds, and 3 inner folds to assess the classification performance in predicting Alzheimer's disease (AD) using retinal OCT scans acquired 4 years before diagnosis. Table 5.4 lists the top AUC results obtained for each model.

The highest performing model was ResNet-34, which achieved a mean AUC of $\mathbf{0.624 \pm 0.060}$ across five nested cross-validation runs. This model was used as a reference point for pairwise comparisons with alternative architectures. VGG-11 (mAUC $= 0.581 \pm 0.017$) and RETFound-C (mAUC $= 0.540 \pm 0.037$) showed reduced performance relative to ResNet-34, but the differences were not statistically significant (corrected paired t-tests $p = 0.1458$ and $p = 0.0845$). The custom CNN model (mAUC $= 0.519 \pm 0.026$) also performed worse than ResNet-34, with significance under the standard t-test ($p = 0.0266$), but not after correction ($p = 0.0364$). In contrast, RETFound-S (mAUC $= 0.459 \pm 0.068$) was significantly worse than ResNet-34 under both tests (standard $p = 0.0043$;

corrected $p = 0.0202$).

We further validated the ResNet architecture on an independent sample set (Dataset 2), which achieved a very similar performance (mAUC = 0.652 ± 0.058), This supports the robustness of the model.

Finally, ablation experiments were conducted by rerunning ResNet-34 with reduced feature inputs. When trained with masked images only (ResNet[1], mAUC = 0.452 ± 0.042), OCT images replicated to three channels (ResNet[2], mAUC= 0.561 ± 0.061), or layer contour inputs (ResNet[3], mAUC = 0.529 ± 0.071), The performance decreased significantly compared to that of the full multichan-nel representation. These results confirm that combining raw OCT, masked, and contour information provides the strongest feature representation for early AD prediction.

To investigate the classification decision of the model and ensure that it learns clinically relevant features, we generated and analyzed saliency maps us-

Table 5.4: Model accuracies on 4-Year dataset

Model	mAUC	f1-score	Precision	Sensitivity	Specificity	t-test p Value	corrected t-test p Value
ResNet	**0.624 ± 0.060**	**0.552 ± 0.135**	**0.583 ± 0.086**	**0.680 ± 0.018**	0.486 ± 0.159		
VGG	0.581 ± 0.017	0.527 ± 0.064	0.568 ± 0.055	0.633 ± 0.062	0.500 ± 0.076	0.1354	0.1458
RETFound-C	0.540 ± 0.037	0.524 ± 0.061	0.557 ± 0.039	0.607 ± 0.037	**0.507 ± 0.081**	0.0728	0.0845
Custom CNN model	0.519 ± 0.026	0.490 ± 0.039	0.553 ± 0.028	0.600 ± 0.041	0.464 ± 0.051	*0.0266	*0.0364
RETFound-S	0.459 ± 0.068	0.468 ± 0.072	0.461 ± 0.046	0.446 ± 0.122	0.479 ± 0.117	*0.0043	*0.0202
ResNet[1]	0.452 ± 0.042	0.424 ± 0.018	0.463 ± 0.021	0.520 ± 0.038	0.407 ± 0.020	*0.0133	*0.0208
ResNet[2]	0.561 ± 0.061	0.523 ± 0.095	0.531 ± 0.090	0.527 ± 0.086	0.536 ± 0.107	*0.0464	0.0577
ResNet[3]	0.529 ± 0.071	0.498 ± 0.091	0.517 ± 0.081	0.533 ± 0.071	0.500 ± 0.104	0.0643	0.0759
ResNet[4]	**0.652 ± 0.058**	**0.613 ± 0.037**	**0.604 ± 0.051**	0.593 ± 0.094	**0.614 ± 0.030**	0.6815	0.6773

Note: ResNet[1]: the ablation test run with masked images replicated to 3 channels. ResNet[2]: ablation test run with OCT images replicated into three channels. ResNet[3]: The ablation test was performed with layer contours replicated into three channels. ResNet[4]: the ResNet model trained with the dataset 2. Corrected *p*-values were com-puted using a calibrated paired *t*-test to account for dependence induced by repeated cross-validation (Bouckaert, 2003).

ing GradCAM. The results are shown in Figure 5.3. First, we aggregated the top 5% most salient pixels from all test images for each class to identify the most consistently important regions for classification (Figure 5.3(a). Furthermore, we analyzed individual cases to connect the model's attention to its performance on specific examples (Figure 5.3(b).

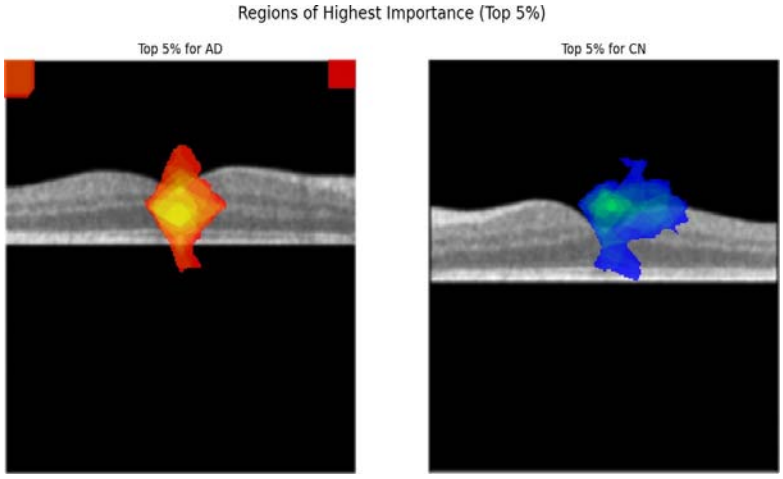
The saliency analysis in Table 5.5 shows that the model gave minimal attention to the RNFL (Filling Ratio < 6% for AD) and instead focused on the

central macular region (34.9% Filling Ratio for AD), especially the BMEIS and IS/OSJ layers.

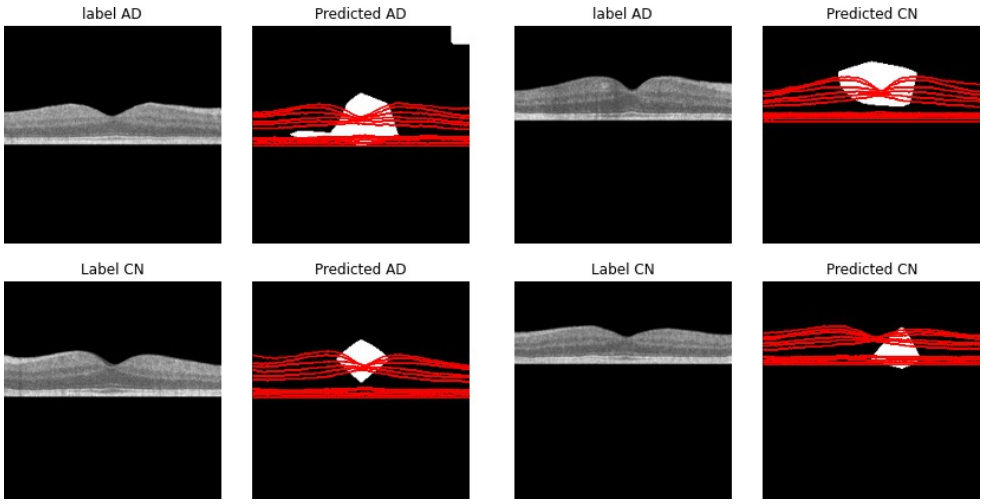
5.4 DISCUSSION

This study presents the first end-to-end deep learning framework for predicting Alzheimer's disease (AD) from UK Biobank Optical Coherence Tomography (OCT) B-scans up to four years before diagnosis. Our best-performing model, ResNet-34, achieved a mean AUC (mAUC) of 0.624 ± 0.060 on a rigorously selected age, sex, and instance-matched AD and control sample.

Earlier studies using data from the UK Biobank predicted early AD from fundus images. Tian et al (Tian et al., 2021). achieved an accuracy of 0.824 using fundus images with vessel segmentation. Wisely et al. (2022) obtained an AUC of 0.625 using only OCTA images and 0.681 with GC-IPL projection maps. However, when they used quantitative data, including age and sex, their model achieved an AUC of 0.96. When all the information (image and quantitative)



(a) Aggregated Top 5% Saliency Regions



(b) Saliency Maps for Individual Test Samples

Figure 5.3: Model interpretability analysis. (a) The aggregated top 5% most salient pixels for the AD (left, red/yellow) and CN (right, blue/green) classes highlight the model's consistent focus on distinct anatomical regions. (b) Individual examples show model attention for True Positives (TP), True Negatives (TN), False Negatives (FN), and False Positives (FP). *Input OCT image was used with permission from the UK Biobank under Application Number 82266.*

Table 5.5: Quantitative Overlap of Saliency Maps with Retinal Layers

Layer	IoU		Dice		Fill (%)	
	CN	**AD**	**CN**	**AD**	**CN**	**AD**
RNFL-GCL	.020	.013	.039	.025	10.6	5.6
GCL-IPL	.038	.017	.070	.033	15.8	7.4
IPL-INL	.048	.028	.089	.054	15.7	8.6
INL-OPL	.052	.021	.097	.041	22.7	10.7
OPL-HFL	.067	.034	.124	.065	23.7	12.6
BMEIS	**.164**	**.103**	**.273**	**.177**	**26.3**	**18.6**
IS/OSJ	.022	.015	.042	.029	16.8	13.4
IB_OPR	.022	.017	.042	.033	15.7	14.9
IB_RPE	.019	.015	.036	.029	11.6	10.9
OB_RPE	.016	.015	.030	.028	9.5	9.9
Macula	.249	.192	.379	.304	41.3	34.9

Note: "Fill (%)" is the Filling Ratio. The filling Ratio is defined as $|\text{Saliency} \cap \text{Layer}|/|\text{Layer}|$. The bold values indicate the highest overlap for each class within the specific layers.

It was combined, the AUC was 0.809. Chua et al (J. Chua et al., 2025). obtained an AUC of 0.82 when using a GC-IPL projection map dataset; however, their performance was reduced to 0.76 when trained with age-matched subjects. These results show that age and sex can be strong confounding factors in AD prediction models. There is no OCTA image dataset in the UK Biobank; therefore, these studies (J. Chua et al., 2025; Wisely et al., 2022) used private OCTA image datasets. Our goal was to study only the structural retinal changes related to AD. As a result, our model performance was moderate, but it likely reflects AD-related structural features in the retina more directly, as it was not influenced by demographic confounders.

When comparing the architectures, ResNet-34 outperformed both VGG-11 and OCT-pretrained transformer models. VGG-11 achieved a mAUC of 0.581 ± 0.017, which was not significantly different from that of ResNet-34 after correction. RETFound-C reached an mAUC of 0.540 ± 0.037, which was not signif-icantly different. In contrast, RETFound-S performed considerably

worse, with an mAUC of 0.459 ± 0.068, and the difference relative to ResNet-34 remained statistically significant even after correction for multiple comparisons. These findings suggest that in low-sample settings, convolutional networks may pro-vide more stable representations than transformer-based models, despite being pre-trained on OCT data. To further test robustness, we repeated the analysis using the same AD cohort and a randomly matched control group. Performance was preserved (AUC = 0.652 ± 0.058), indicating that the results were stable against variations in the control selection.

Ablation experiments further emphasized the importance of the three-channel input strategy. When ResNet was trained using only masked images, only repli-cated grayscale OCT, or only retinal layer contours, performance dropped sub-stantially compared to the full multichannel input. This demonstrates that the combination of raw OCT intensity, layer-specific masking, and anatomical bound-ary (contour) information provides the most informative feature representations.

Explainability analysis showed that the model allocated very little attention to the RNFL (Filling Ratio <6% for AD), despite its frequent use as a biomarker in previous studies. Instead, attention was concentrated on the central macular region (34.9% for AD), with the highest overlap observed in the BMEIS and IS/OSJ layers. This indicates that the network relies on features within the pho-toreceptor and RPE complex, highlighting regions that may act as early biomark-ers of AD. These results further support the potential of retinal OCT combined with AI for noninvasive and scalable early AD risk prediction.

Despite these promising results, this study has several limitations. The dataset size, particularly in the 4-year diagnostic group, constrained the statisti-cal analyses and may have limited the generalizability of the findings to broader populations. One of the main contributing factors to the modest prediction accu-racy was the study's dependency on a single dataset, which limited the model's ability to generalize and increased the risk of overfitting. Although our nested

cross-validation framework was designed to mitigate overfitting and provide a robust internal estimate of model performance, external validation remains a critical next step. However, to the best of our knowledge, no publicly available OCT datasets exist that include both retinal imaging and longitudinal Alzheimer's disease outcomes, making such validation infeasible. Future collaborations with clinical centers or research initiatives will be essential to acquire independent cohorts for further validation. Establishing shared datasets and benchmarks in this domain would greatly enhance reproducibility and comparability across studies and facilitate the translation of OCT-based biomarkers into clinical use. Additionally, while nested cross-validation mitigates overfitting, further improvements could be achieved through ensembling, multimodal integration (e.g., OCT angiography, cognitive testing), or temporal modeling of longitudinal scans. Finally, clinical validation against gold-standard biomarkers (e.g., amyloid PET and CSF tau) is necessary to establish OCT's role in preclinical AD screening.

5.5 CONCLUSION

This study presents the first deep learning framework applied to UK Biobank retinal OCT data for early prediction of Alzheimer's disease up to four years before diagnosis. We proposed an anatomically guided preprocessing pipeline with multichannel OCT input, hybrid augmentation, and nested cross-validation. ResNet-34 achieved the best performance, with a mean AUC of 0.624, and ro-bustness was confirmed with a re-matched control dataset. Ablation experiments showed that combining raw, masked, and contour inputs improved performance, while saliency maps highlighted the macular BMEIS and IS/OSJ layers. Our findings provide a reproducible baseline for OCT-based AD prediction, highlight the challenges of detecting subtle retinal biomarkers years before AD diagnosis, and point to the need for larger datasets and multimodal approaches.

CHAPTER 6

6. EARLY AD DETECTION FROM RETINAL OCT C-SCANS

A significant limitation of previous OCT analyses is that standard B-scans represent only a single cross-sectional "slice" of a much larger 3D volume. Although these slices provide high-resolution axial details, they inherently fail to capture the full spatial context of retinal pathology. Consequently, a critical question arises: can we extract more diagnostic information by considering the entire 3D volume rather than isolated slices?

While 3D analysis is theoretically superior, processing full volumetric data and modeling the complex associations between various retinal layers are computationally expensive and require more data. To overcome these costs, en-face OCT imaging, as depicted in Figure 6.1 can be used to generate a 3D-informed representation within a 2D format.

An en face image is a frontal view of the retina created by reconstructing 3D data along a specific anatomical plane. Instead of looking at a vertical "cut," the en-face view looks "down" onto the retinal surface, similar to a traditional fundus photograph but with depth-specific capabilities. By projecting the 3D layers into an en-face representation, we can visualize comprehensive vascular and structural patterns, such as vessel density or geographic atrophy, that are often invisible or difficult to quantify on individual B-scans. Fig. 6.1 shows a mean-projection en face image made by taking 3D OCT stack (a series of x–z B-scans) and, for each x–y pixel, averaging its intensity across all depths (the z-axis). This “flattens” the volume into a single 2D map, where bright, consistent structures throughout the depth stand out and random speckles are smoothed away. In contrast, a B-scan is just one of those original cross-sections (showing depth vs. one lateral line) and provides information about layer thicknesses and boundaries at that slice. The thickness projection map, as depicted in Figure 6.2 (c), was calculated by considering the height difference between pixels of two

specific boundary layers (e.g., between the ILM and RPE).

We found that the model focused mostly on the macular region to make its decisions; therefore, we concentrated on the inner circle thickness projection maps, as shown in Figure 6.2 (b). This specific focus was derived from an interpretability analysis performed in our preliminary B-scan study.

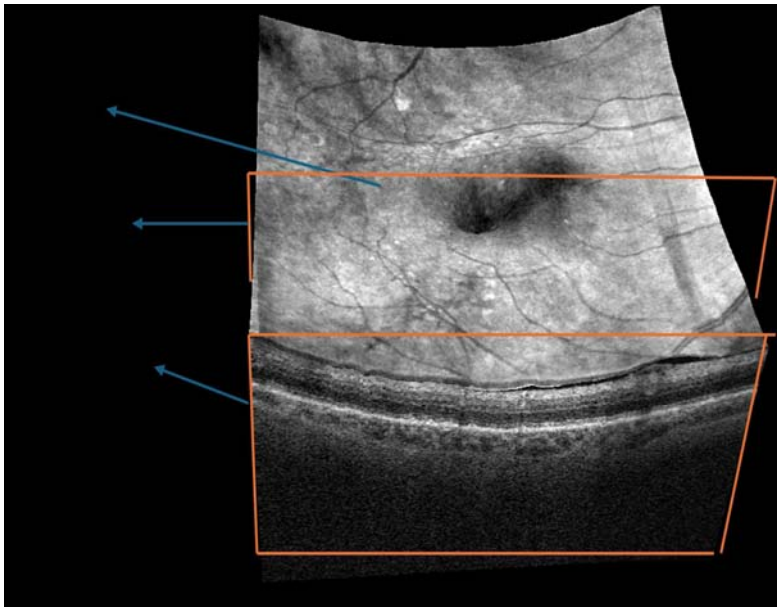


Figure 6.1: Visualization of a volumetric OCT scan showing the en face projection (top) and a corresponding B-scan (bottom) through the macular region. The orange boxes indicate the spatial locations of the B-scan slices within an en face view. The blue arrows indicate the key regions of interest along the x, y, and z dimensions, highlighting the correspondence between the 2D cross-sectional (B-scan) and 3D surface (en face) features. This alignment provides anatomical context and enables precise localization of retinal layer structures for downstream analysis and model input.

By applying saliency mapping to individual B-scans, we observed a consistent concentration of high-importance features in the central slice. Based on these interpretability results, we concluded that the peripheral data contributed negligible predictive value compared with that of the middle sections of the scan. To translate these cross-sectional findings into a more com-

putationally efficient format, we selected an inner circle (3 mm). Thickness projection maps were used as the input. This approach ensured that the input data were pre-filtered to include only the anatomically relevant regions identified by the interpretability analysis.

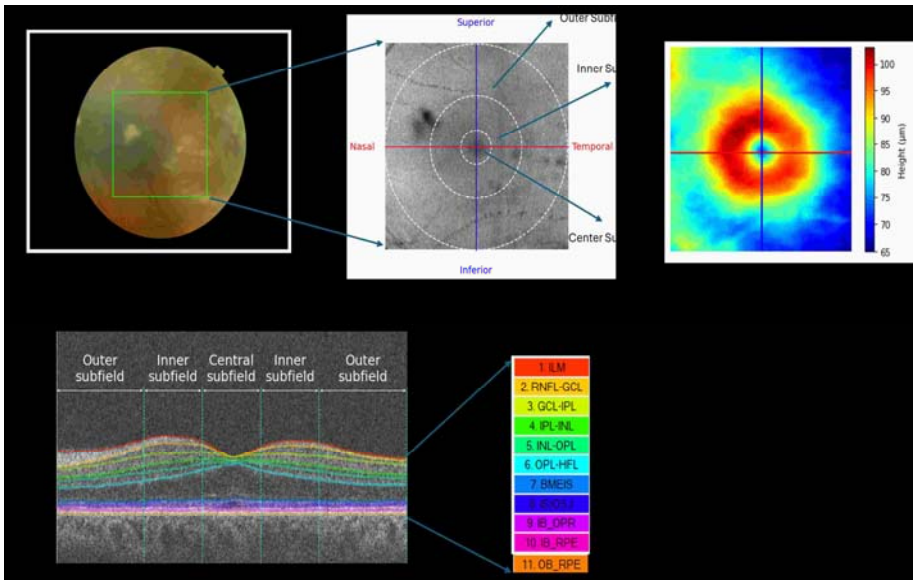


Figure 6.2: OCT Scan Details: (a) 6mm×6mm macular OCT scan overlay on fundus image. (b) The ETDRS grid subdivides the macula into three concentric rings: central (1 mm), inner (3 mm), and outer (6 mm). (c) Macular Thickness shown on a projection map. (d) Retinal layers visualized in the OCT b-scan with ETDRS overlays. ILM: Inner Limiting Membrane, RNFL: Retinal Nerve Fiber Layer, GCL: Ganglion Cell Layer, IPL: Inner Plexiform Layer, INL: Inner Nuclear Layer, OPL: Outer Plexiform Layer, ONL: Outer Nuclear Layer, HFL: Henle's Nerve Fiber Layer, BMEIS: Boundary of Myoid and Ellipsoid of Inner Segments, IS: Inner Segment, OSJ: Outer Segment Junction, IB_OPR: Inner Boundary of the Outer Photoreceptor, IB_RPE: Inner Boundary of the Retinal Pigment Epithelium, OB_RPE: Outer Boundary of the Retinal Pigment Epithelium. Images obtained from the UK Biobank resource (© UK Biobank).

6.1 METHOD

In this study, we used 2D architectures based on ImageNet. To standardize the input, all retinal layer thickness projection maps were cropped to the 3 mm inner circle region and resized to 512×512 pixels, as shown in Figure 6.3. To meet the three-channel input requirements of the 2D models, each single-channel OCT thickness map was replicated across the RGB channels. Because these images represent specific anatomical thickness data rather than natural scenes, the models were trained from scratch without using the pretrained ImageNet weights.

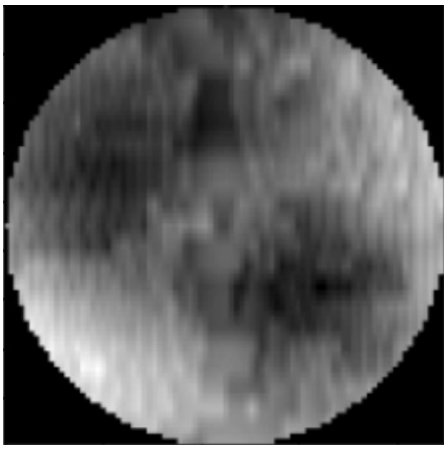


Figure 6.3: Retinal Thickness Projection Map

The images represent specific anatomical thickness data; therefore, we restricted the data augmentation to techniques that preserve the underlying morphology. Specifically, we applied horizontal flipping, translation, and occlusion of random patches (cutouts) to enhance the model robustness without compromising the clinical integrity of the retinal maps.

We used ResNet-34 as the backbone architecture to evaluate the predictive performance of the individual retinal layer thickness maps. To ensure stable gradient updates within the constraints of our limited dataset size, all models were trained with a batch size of four. Optimization was performed using the

AdamW optimizer, and we explored the convergence behavior by testing a range of learning rates (0.001, 0.0001, and 0.000027).

To adapt the 2D backbone for binary classification under low-data conditions, we replaced the standard output layer with a lightweight classification head. This head consists of a fully connected layer (mapping the feature dimension to 64 units), followed by Layer Normalization, a ReLU activation, and a dropout layer ($p = 0.4$). A final linear layer projects the features onto two output classes. This streamlined regularization structure was designed to balance the model expressivity with robust overfitting control.

Finally, to account for the temporal dynamics of disease progression, we employed a year-weighted loss function (Nazlı et al., 2025). This approach assigns greater weight to samples collected closer to the clinical diagnosis of Alzheimer's disease, prioritizing the most pathologically relevant data points during training.

We implemented a nested 5-fold cross-validation strategy to ensure the reliability and reproducibility of our results. Each experiment was repeated five times using different random seeds to ensure a diverse data distribution across the folds. We reported the final performance as the mean validation AUC, with the standard deviation accounting for the variability across the independent runs. Following the identification of the optimal retinal layer thickness map, we evaluated the generalizability of our findings by extending our analysis to VGG and Swin Transformer architectures. These models were selected to compare the performance of traditional deep convolutional neural networks with that of modern attention-based vision transformers under the same experimental conditions. We included the Swin Transformer in our study because a previous study successfully used this model to detect MCI.

Standard Convolutional Neural Networks (CNNs), such as ResNet or VGG, process images by examining small, local areas (receptive fields) one at a time. In contrast, the Swin Transformer uses a hierarchical "shifted-window" attention mechanism. This design allows the model to connect distant parts of the image and understand the entire retinal map simultaneously. Theoretically,

this abil-ity to capture long-range dependencies is useful for detecting subtle, widespread signs of early stage neurodegeneration. Therefore, we selected this architecture to test whether its global attention capabilities could outperform those of tradi-tional CNNs in analyzing retinal thickness maps.

We chose the Grad-CAM method to explain our best-performing model.

To evaluate the predictive capacity of the retinal thickness maps at various stages before disease onset, we performed a longitudinal sensitivity analysis. The dataset was partitioned into temporal subsets based on the interval between the initial OCT scan and the date of clinical Alzheimer's disease (AD) diagnosis. This interval ranged from 4 to 12 years, allowing us to assess how far in advance pathological changes were detectable in the patients.

For each annual time point, we used the optimized VGG-19 architecture to perform binary classification (healthy vs. future AD cohorts). This process was conducted separately for the Ganglion Cell Layer (GCL) and Retinal Nerve Fiber Layer (RNFL) thickness maps to compare their respective diagnostic utilities. To ensure statistical robustness, we maintained our 5-fold cross-validation protocol for each temporal window and calculated the mean AUC and its standard devi-ation. This approach allowed us to quantify the rate at which the morphological markers in each layer converged toward the 0.5 chance baseline, effectively iden-tifying the "diagnostic horizon" for each biomarker.

6.2 RESULTS

The performance of the ResNet-34 model across different retinal layer thickness maps is shown in Table 6.1. The Ganglion Cell Layer (GCL) thick-ness map achieved the highest predictive accuracy, yielding a Mean AUC of 0.715 ± 0.028. This significantly outperformed the other layers. In contrast, the RNFL (0.385 ± 0.074) and IS/OSJ (0.412 ± 0.068) showed performance lev-els below the chance thresholds. Mid-range performance was observed in the BMEIS (0.591 ± 0.032) and INL (0.578

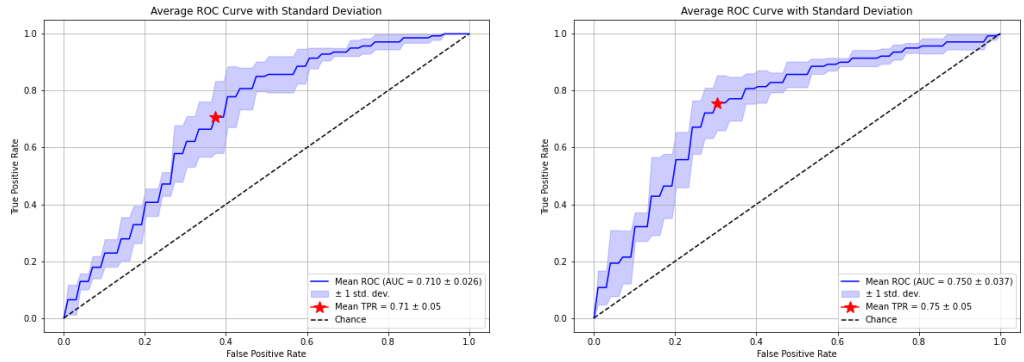
$\pm$ 0.059) regions.

Table 6.1: Validation AUC Results per Layer

Layer	Mean AUC
Total Macula	0.493 $\pm$ 0.039
RNFL	0.385 $\pm$ 0.074
GCL	0.715 $\pm$ 0.028
IPL	0.568 $\pm$ 0.039
INL	0.578 $\pm$ 0.059
OPL	0.470 $\pm$ 0.058
HFL	0.569 $\pm$ 0.016
BMEIS	0.591 $\pm$ 0.032
IS/OSJ	0.412 $\pm$ 0.068
IB OPR	0.423 $\pm$ 0.018
IB RPE	0.527 $\pm$ 0.038

Although the Swin Transformer was evaluated to leverage its attention-based spatial modeling, the architecture faced significant overfitting issues, likely because of the limited dataset size. The transformer models failed to achieve reasonable predictive performance because they consistently converged toward the training noise rather than generalizable features.

In contrast, the VGG-19 model demonstrated superior robustness and performance compared to the ResNet-34 baseline. Utilizing the GCL thickness maps, VGG-19 achieved a mean AUC of 0.750 $\pm$ 0.037. To further illustrate these findings, we provide the Receiver Operating Characteristic (ROC) curves for both the ResNet-34 and VGG-19 models in Figure 6.4. These plots confirmed the stability and improved sensitivity of the VGG-19 architecture in identifying the target pathology.



(a) ResNet-34 (b) VGG-19

Figure 6.4: Comparative ROC analysis for the GCL thickness map using different architectures. **(a)** ROC curve of the ResNet-34 model (Mean AUC: 0.715 ± 0.028). **(b)** ROC curve of the VGG-19 model, which achieved the highest predictive stability and performance (Mean AUC: 0.750 ± 0.037).

Results are averaged over 5-fold cross-validation.

We analyzed model performance relative to the time elapsed before a formal AD diagnosis to evaluate the clinical utility of retinal thickness maps for early screening, As illustrated in Figure 6.5, a distinct trend was observed exclusively within the Ganglion Cell Layer (GCL). The predictive accuracy was highest for scans taken closer to the point of diagnosis (4 years prior) and exhibited steady degradation as the temporal horizon extended.

The GCL's performance converged toward the 0.5 chance baseline by the 12-year mark. Its performance remained consistently near the 0.5 baseline for the 12-year period.

6.3 DISCUSSION

The results of this study offer important insights into the application of en face retinal thickness maps for the early detection of Alzheimer's

disease (AD).

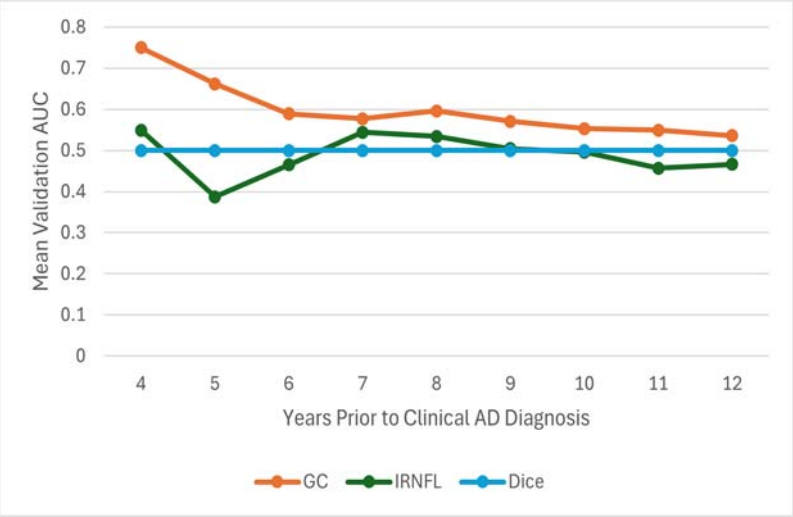


Figure 6.5: Longitudinal performance comparison between the GCL and RNFL thickness maps across a 12-year pre-diagnostic window. The y-axis represents the Mean AUC, while the x-axis denotes the years prior to a clinical AD diagnosis. The GCL demonstrates a unique temporal degradation in predictive accuracy, whereas the RNFL consistently fluctuates near the 0.5 chance baseline, indicating a lack of predictive significance for early detection in this cohort.

The outcomes validated the transition from two-dimensional B-scans to three-dimensional informed projection maps and highlighted a notable difference in the diagnostic utility of the specific retinal layers.

The most notable finding is the performance of the Ganglion Cell Layer (GCL), which achieved a peak mean AUC of 0.750 ± 0.037 using the VGG-19 architecture. Although the Retinal Nerve Fiber Layer (RNFL) is traditionally the most studied biomarker in glaucoma and various neurodegenerative diseases, it did not provide a higher predictive signal in this cohort, fluctuating near the 0.5 chance baseline. These results suggest that in early AD, neurodegeneration may manifest more prominently as the loss of retinal ganglion cell bodies (in the

GCL) rather than the loss of axons (in the RNFL). This observation is consistent with recent histopathological findings indicating that thinning of the cell body in the macula may precede axonal loss in the peripapillary region.

Our comparative analysis of the architectures revealed a clear performance hierarchy. Although Swin Transformers represent the state-of-the-art for large-scale natural image datasets, they exhibited severe overfitting in these experiments. This outcome confirms that for specialized medical imaging tasks with limited sample sizes, the inductive bias of convolutional neural networks surpasses that of global attention mechanisms in transformers.

The VGG-19 model outperformed ResNet-34, likely because its simpler and deeper sequential structure was more effective in recognizing subtle, low-frequency textural changes inherent in thickness projection maps. In contrast, the residual connections of ResNet are often better suited for higher-frequency edge detection in natural-image processing.

Longitudinal analysis (Figure 6.5) identified a "diagnostic horizon" for retinal biomarkers. The GCL showed a clear and steady decline in predictive power as the interval from clinical diagnosis increased. Convergence to a 0.5 AUC at the 12-year mark indicates that, although retinal changes are "early" markers, they may only become detectable by current deep learning methods approximately 4 to 8 years before clinical symptoms emerge. The absence of any RNFL signal across this 12-year window further highlights the importance of the GCL in developing screening tools for preclinical AD.

6.4 CONCLUSION

This study evaluated the predictive utility of 3D-informed en-face retinal thickness maps from the UK Biobank for identifying Alzheimer's disease (AD) up to 12 years before clinical diagnosis. An anatomically guided preprocessing pipeline was proposed, focusing on the 3 mm inner macular region and utilizing multichannel inputs and morphology-preserving data augmentation. Among the tested architectures, VGG-19 achieved the highest performance on the Ganglion Cell Layer (GCL), yielding a mean AUC of 0.750 ± 0.037. Longitudinal sensitivity analysis demonstrated a distinct 'diagnostic horizon,' with the GCL between 4 and 8 years before diagnosis, while the traditionally utilized RNFL remained near the chance baseline. These findings suggest that the GCL is a superior morphological biomarker for preclinical AD screening and provides an effective platform for long-term OCT analyses.

In the next chapter we investigated whether combining GCL thickness with B-scans and numerical data, such as demographic information and OCT system-driven results, could improve the diagnostic performance or extend the diagnostic horizon beyond four years.

CHAPTER 7

7. MULTI-MODAL SOFT-VOTING ENSEMBLE

The previous chapters established the efficacy of individual modalities in early AD prediction. Although each model achieved respectable performance individually, they operated independently. In clinical practice, a diagnosis is rarely formed from a single data point; rather, it is a consensus derived from multiple sources of evidence. To emulate this clinical decision-making process, this study proposes a **Multi-Modal Soft-Voting Ensemble**. Unlike complex fusion architectures that require retraining, this approach leverages the probabilistic outputs of previously trained models (from the 5-fold cross-validation phases) to derive a final, robust classification through statistical averaging. To ensure methodological consistency across the ensemble components, the XGBoost model was retrained on the 4-year dataset with the same cross-validation folds as the VGG and ResNet models.

7.1 METHODS

The proposed framework does not concatenate feature vectors. Instead, it operates at the decision level, aggregating the confidence scores (probabilities) generated by the three distinct classifiers.

The ensemble shown in Figure 7.1 integrates the following three streams, as optimized in previous chapters:

Stream 1: ResNet34 (B-Scan Mid-Layers): This model targets cross-sectional structural details. For each patient, the five models generated during the 5-fold cross-validation in Chapter 5 provided a probability score indicating the likelihood of future AD based on B-scans.

Stream 2: VGG19 (GCIPL Thickness Projection Maps): This model analyzes the topographical thickness maps. Similarly, the probabilities from

the five validated VGG19 models were used.

Stream 3: XGBoost (Clinical & Demographic Data): This model processes structured data.

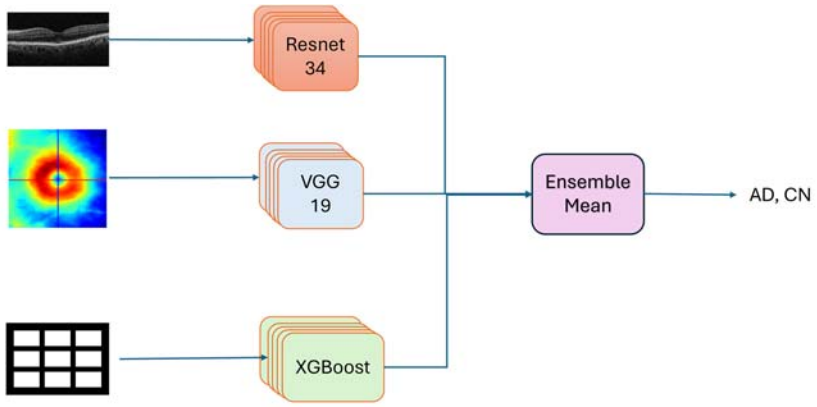


Figure 7.1: The Ensemble model over individual architectures. The core mechanism of this ensemble is Soft Voting. Let $P_{V\,GG}$, P_{ResNet}, and P_{XGB} be the predicted probabilities of the positive class (AD) from the three streams. The final ensemble probability, $P_{Ensemble}$, is calculated as the unweighted mean of these streams as follows:

$$P_{Ensemble} = \frac{1}{3}(P_{V\,GG} + P_{ResNet} + P_{XGB}) \qquad (7.1$$

While the unweighted mean (Soft Voting) assumes that each modality is equally reliable, we also investigated a Logistic Regression Ensemble (Stacked Generalization) to account for the varying predictive strengths of the three streams. In this configuration, the probabilistic outputs $P_{V\,GG}$, P_{ResNet}, and P_{XGB} serve as input features for a meta-classifier. The final prediction is defined as:

$P_{Ensemble} = \sigma(\beta_0 + \beta_1 P_{V\,GG} + \beta_2 P_{ResNet} + \beta_3 P_{XGB})$(7.2) where σ is the sigmoid function, and β_i represents the learned coefficients. This allows the ensemble to "weight" the VGG19 stream more heavily if it con-sistently

demonstrates a higher empirical accuracy than the ResNet34 stream.

To ensure stability, the system first aggregates the predictions from the 5 repetitions of 5-fold cross-validation models within each stream. For a given input sample, the 5 variations of the VGG19 model produce 5 probability scores; their highest score is taken as the representative P_{VGG}. This process was repeated for the ResNet and XGBoost streams. Finally, the three representative probabilities were averaged to produce a final decision.

This method was designed to mitigate "confident wrong" predictions. For instance, if the visual model (VGG) detects a false positive due to an image artifact, but the clinical model (XGBoost) reports a low probability due to healthy RNFL thickness, the averaging process suppresses the false positive, resulting in a correct classification.

To test our models, we used two types of 4-year datasets: the B-Scan dataset and the Cohort dataset. Because the dataset was small, we included both eyes of the AD patients when possible. However, to prevent data leakage, we ensured that both eyes of the same patient were always in the same group (either training or validation). This method allowed us to increase the number of samples from 49 to 58. For the Cohort dataset, if we had results for both eyes, we calculated the final score by taking the average of the two probabilities.

7.2 RESULTS

The ensemble was evaluated on the folds of the test set described in the Methodology section. The threshold for classification was set at 0.5; if $P_{Ensemble} > 0.5$, the subject was classified as having AD.

Table 7.1 presents the performance metrics of the multimodal ensemble compared to the best individual predictors from the previous chapters. As illustrated, the ensemble approach yielded a higher accuracy than any single modality.

Table 7.1: Performance Comparison of Uni-Modal Models vs. Multi-Modal Soft-Voting Ensemble

Model	Dataset	Clinical Data	B-Scan	GCIPL Map	AUC
XGBoost	Cohort	✓			0.70
ResNet34	Scan		✓		0.62
VGG19	Scan			✓	0.75
Ensemble-mean	**Cohort**	✓	✓	✓	**0.85**
Ensemble-logistic reg.	Cohort	✓	✓	✓	0.84
Ensemble (ablation)	Scan		✓	✓	0.82
Ensemble (ablation)	Cohort		✓	✓	0.84

Uni-Modal Performance Among the individual models, the VGG19 model trained on GCIPL maps achieved the highest performance, with an AUC of 0.75. The XGBoost model, which utilized only tabular clinical data, had an AUC of 0.70. The ResNet34 model trained on B-scan images showed the lowest performance, with an AUC of 0.62.

Ensemble Performance The results demonstrate that combining these modalities significantly improves the classification accuracy. The Multi-Modal Mean Ensemble, which integrates Clinical Data, B-Scans, and GCIPL Maps, achieved the highest overall AUC of **0.85**. This result was 0.10 higher than that of the best single-modality model (VGG19). As shown in Table 7.1, the Logistic Regression Ensemble achieved an AUC of 0.84, which was slightly lower than the 0.85 achieved by the simple mean. This indicates that although the meta-classifier attempted to optimize the influence of each model, the unweighted average provided better generalization for this specific dataset.

Ablation Study We also performed an ablation study to test the ensemble using only image datasets (B-Scan and GCIPL Maps) without clinical data. This image-only ensemble achieved an AUC of 0.84 at the cohort level. This indicates that while the combination of retinal images provides strong predictive power, the addition of clinical data slightly improves the model performance.

7.2.1 Discussion

The results of this study strongly support the hypothesis that combining multiple data sources improves the detection of early Alzheimer's disease (AD). While individual models showed promise, the multimodal ensemble achieved the highest accuracy (AUC 0.85). It significantly outperformed the best single-modality model (VGG19, AUC = 0.75).

The primary reason for the success of the ensemble was the complementary nature of these data streams. Each model "looks" at the patient from a different perspective.

The **VGG19 model** analyzes the GCIPL thickness maps, identifying topographical patterns of atrophy.

The **ResNet34 model** examines the B-Scans, focusing on the cross-sectional integrity of retinal layers.

The **XGBoost model** evaluates clinical risk factors like blood pressure and education level.

By averaging these inputs, the ensemble mimics the clinical consensus. For example, if the B-scan model produces a false positive due to image noise, the Clinical model (which might indicate low risk) and the GCIPL model can correct this error. This explains why the ensemble is more robust than any single model that operates in isolation. It is notable that the ResNet34 model (B-Scan) had the lowest individual performance (AUC 0.62). However, the ablation study showed that combining B-scans with GCIPL maps (ensemble ablation) raised the AUC to 0.84, which is much higher than that of the GCIPL model alone (0.75). This suggests that although the B-scan model struggles to make a diagnosis on its own, it contains valuable structural information that reinforces the predictions of the thickness maps.

The addition of clinical data provided a final performance improvement, increasing the AUC from 0.84 to 0.85. Although this increase appeared small, it demonstrated that non-visual markers (such as demographics and systemic

health) help to refine visual predictions. In a clinical setting, where every percentage point of accuracy matters for early diagnosis, this integration is valuable. The superior performance of the unweighted mean to the logistic regression ensemble can likely be attributed to the relatively small sample size (N = 58). While Logistic Regression is a more sophisticated aggregation method, it in-troduces additional parameters (β coefficients) that must be estimated. The sim-ple soft-voting approach acts as a natural regularizer, reducing the variance of in-dividual "confident but wrong" predictions without the risk of over-parameterization.

The proposed soft-voting framework offers practical advantages for clinical applications. Unlike complex "black box" fusion models that require training from scratch, this method uses the outputs of existing systems. This made it modular; if a better B-scan model is developed in the future, it can simply replace the current stream without changing the entire system design.

However, this study had limitations in terms of sample size. Although we successfully expanded the dataset to 58 samples by utilizing both eyes and applying strict patient-level splitting, it remained relatively small in size. Although the results were promising and statistically valid within this cohort, future studies with larger populations are necessary to confirm the generalizability of our ensemble model.

7.2.2 Conclusion

In conclusion, this chapter successfully established a multimodal soft-voting ensemble that significantly outperformed the individual detection methods. By statistically averaging the probabilistic outputs of the ResNet34 (B-scan), VGG19 (GCIPL), and XGBoost (clinical) models, the system achieved a robust AUC of 0.85, confirming that fusing diverse evidence streams captures a more complete picture of early neurodegeneration. This approach not only mimics the collab-orative nature of clinical diagnostics but also provides a flexible, noninvasive framework for improving the accuracy of early Alzheimer's disease detection.

CHAPTER 8

8. DISCUSSION

This study represents the first application of a deep learning framework to the UK Biobank's retinal OCT image data for predicting Alzheimer's disease up to four years before clinical diagnosis.

The diagnostic performance reported in recent studies for Mild Cognitive Impairment (MCI) generally ranges from 0.809 to 0.968 (see Table 8.1). Numerically, our single-modality peak Mean AUC of 0.750 ± 0.037 using VGG-19 on GCL thickness maps is lower than the benchmarks set by Gao et al. (2023) (0.968 for MCI) and Hao et al. (2024) (0.863 for MCI). However, this disparity highlights the profound difference in task difficulty:

Diagnosis vs. Prediction: The cohorts in the literature were already symptomatic at the time of imaging. In contrast, our UK Biobank cohort was cognitively normal during the initial scan, with AD appearing only years later.

2D vs. 3D: When comparing structural-only results, our 0.750 for future prediction is remarkably competitive. For instance, Wisely et al. (2024) re-ported an AUC of only 0.681 when restricted to structural OCT for current MCI diagnosis. This suggests that the 3D-informed en-face GCL mapping used in this study captures a more robust neurodegenerative signal than traditional B-scan or tabular feature analyses.

A critical advancement presented in this thesis is the development of a multimodal soft-voting ensemble (Chapter 7). While individual modalities showed respectable performance, operating in isolation limits their diagnostic potential. By integrating clinical biomarkers (XGBoost), cross-sectional structural data (ResNet34 on B-scans), and topographical thickness maps (VGG19 on GCIPL), our ensemble model achieved a peak AUC of **0.85**.

This result is significant for two reasons:

Surpassing Single-Modality Limits: The ensemble outperformed the best

single-modality model (VGG19) by a margin of 0.10 (AUC 0.85 vs. 0.75). This confirms that non-visual clinical risk factors and structural B-scan data still contain complementary information that reinforces the topographical signals observed in the GCL maps.

Competitive early AD detection baselines: With an AUC of 0.85, our predictive model for *pre-symptomatic* patients begins to rival the per-formance of models designed for MCI diagnosis (e.g., Wisely et al., 0.809), effectively bridging the gap between early screening and clinical confirma-tion.

Table 8.1: Consolidated Comparison: Symptomatic Literature (MCI/AD) vs. This Thesis (Pre-symptomatic AD)

Study	Target Stage	Dataset	Primary Modality	Max AUC	AUC (OCT Only)
Gao et al. (2023)	MCI	Private	OCT + Fundus (DuCAN)	0.968	0.903
Hao et al. (2024)	EOAD, MCI	Private	OCTA (Graph-based)	0.936 (AD)	–
Liu et al. (2025)	AD, MCI	Private	OCTA (PolarNet+)	0.887 (AD)	–
Chua et al. (2025)	AD, MCI	Private	OCT Maps + Anatomy	0.910	0.820
Wisely et al. (2024)	MCI	Private	OCT + OCTA + De-mog.	0.809	0.681
This The-sis (2025)	**Future AD**	**UK Biobank**	**Structural OCT (GCL En-face)**	**0.750**	**0.750**
This The-sis (2025)	**Future AD**	**UK Biobank**	**Multi-Modal Ensem-ble**	**0.850**	**0.840***

*AUC of 0.840 represents the Ensemble ablation using only OCT images (B-Scan + GCIPL) without clinical data.

Our findings consistently identified the Ganglion Cell Layer (GCL) as the most potent early biomarker, outperforming the traditionally emphasized RNFL. This aligns with the biological evidence that the loss of retinal ganglion cell bodies (GCL) may precede the significant loss of their axons (RNFL) in the early stages of amyloid-beta accumulation. While Chua et al. (2025) achieved high performance by combining the RNFL and GCIPL, our layer-specific ablation

experiments confirmed that isolating the GCL optimizes the signal-to-noise ratio for the preclinical detection.

Furthermore, the ablation study in our ensemble chapter revealed that combining B-scans with GCL maps raised the AUC from 0.75 to 0.84. This suggests that while B-scans alone (AUC 0.62) are noisy predictors, they contain latent structural features—likely subtle mid-layer disruptions—that validate and strengthen the signals observed in the en-face maps.

Unlike the private hospital-based datasets used in the literature (e.g., ROAD, Duke, and Wenzhou), which may suffer from limited generalizability, our use of the UK Biobank provides a reproducible baseline for population-level screen-ings. Our study revealed that the GCL based retinal signal was most discrimina-tive 4–8 years prior to diagnosis, after which the AUC converged toward the 0.5 chance baseline.

CONCLUSION AND SUGGESTIONS

This thesis explored if retinal OCT images can predict Alzheimer's disease (AD) early. We used a dataset from the UK Biobank for our analyses. We specifically focused on patients four years before diagnosis. This allowed us to build a system that detects signs of disease before the symptoms appear.

This thesis had three main phases. The Multi-Model Ensemble helped us build a strong prediction model.

Statistical Analysis: First, we looked at the data using statistics. We confirmed that the macula looks different in patients with future AD. Specifically, the **Ganglion Cell-Inner Plexiform Layer (mGCIPL)** was significantly thinner ($p = 0.050$). We also tested clinical data, like education and blood pressure, using a model called XGBoost. This gave an AUC of 0.70. This showed that clinical data is useful but not enough on its own.

Deep Learning on Images: We tested two deep learning methods. First, we used **ResNet34** on B-scan images (cross-sections of the eye). It achieved an AUC of 0.62. This was a modest result, but it proved that the model focused on the correct central layer. Second, we created 2D thickness maps of the retina. We used the **VGG-19** model for these maps. This worked much better and achieved an AUC of **0.75**. This proves that the thickness patterns contain strong disease-related signals.

Multi-Modal Ensemble: The most important result was the **Multi-Modal Ensemble**. We combined the predictions from three models: clinical (XGBoost), B-scans (ResNet34), and Thickness Maps (VGG-19). By averaging their scores, we achieved a peak **AUC of 0.85.** This mimics how doctors work. Doctors examine different types of evidence to make a diagnosis. Our model performs the same function and performs much better than any single method.

This study makes three key contributions to the literature.

The GCL is the Best Early Biomedical Marker We found that the Ganglion Cell Layer (GCL) is the most important biomarker for early AD. Re-searchers

often look at the Nerve Fiber Layer (RNFL). However, our results showed that the GCL was the earliest impacted layer. The RNFL remained healthy in our early-stage patients. This suggests that cell bodies die before nerve fibers in the early stage.

The "Diagnostic Horizon" We identified the optimal time for detecting these changes. The retinal signal is strongest **4–8 years before diagnosis**. This is a critical insight for this thesis. This suggests that eye scans are most useful as a screening tool in mid-life rather than a test for late-stage patients.

High Accuracy for Pre-Symptomatic Patients We compared our results to existing studies. Other studies have shown high accuracy ($AUC > 0.80$) in patients with existing symptoms. Our Ensemble model achieved an AUC of **0.85** for patients who were *pre-symptomatic*. This is a major improvement. It shows we can detect AD years before cognitive decline starts.

The results are promising, but there are limitations. The main limitation is the small sample size ($N = 58$). This was due to our strict rules for data selection and the 4-year window.

While the current findings are promising, several tests remain for future exploration to enhance the robustness and generalizability of the proposed ensemble:

Expansion of the Control Cohort: Future work should focus on rerun-ning the training with more controls. Increasing the sample size of the healthy cohort would allow for a more granular understanding of the physiological baseline and help the model better distinguish subtle pathological deviations from healthy aging.

Robustness Testing via Un-matched Controls: It can also be tested to rerun the training with un-matched control groups. Evaluating the model's performance on datasets where subjects are not strictly matched by demographics or systemic health factors will provide insights into the system's resilience and its potential for deployment in diverse, real-world clinical environments.

In conclusion, this thesis proves that the eye is a window to the brain. We

used deep learning to analyze retinal images before AD diagnoses. We built a robust framework that detects early warning signs of Alzheimer's. This brings us closer to a future where a simple eye scan can help doctors diagnose AD before it-causes-damage.

REFERENCES

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, *6*(4), 701–726.

Abramoff, M. D., Garvin, M. K., & Sonka, M. (2010). Retinal imaging and image analysis. https://doi.org/10.1109/RBME.2010.2084567

Alamouti, B., & Funk, J. (2003). Retinal thickness decreases with age: An oct study. *Br J Ophthalmol*, *87*, 899–901. www.bjophthalmol.com

Alber, J., Arthur, E., Sinoff, S., Cabrera Debuc, D., Chew, E. Y., Douquette, L., Hatch, W. V., Hudson, C., Kashani, A., Lee, C. S., Montaquila, S., Mozdbar, S., Cunha, L. P., Tayyari, F., Stavern, G. V., & Snyder, P. J. (2020). A recommended "minimum data set" framework for SD-OCT retinal image acquisition and analysis from the Atlas of Retinal Imaging in Alzheimer's Study (ARIAS). *Alzheimer's & Dementia*. https://doi.org/ 10.1002/dad2.12119

Anstey, K. J., Cherbuin, N., & Herath, P. M. (2013). Development of a new method for assessing global risk of alzheimer's disease for use in popu-lation health approaches to prevention. *Prevention Science*, *14*(4), 411–421. https://doi.org/10.1007/s11121-012-0313-2

Attiku, Y., He, Y., Nittala, M., & Sadda, S. (2021). Current status and future possibilities of retinal imaging in diabetic retinopathy care applicable to low- and medium-income countries. *Indian Journal of Ophthalmology*, *69*(11), 2968.

Aumann, S., Donner, S., Fischer, J., et al. (2019). Optical coherence tomography (oct): Principle and technical realization. In B. JF (Ed.), *High resolution imaging in microscopy and ophthalmology: New frontiers in biomedical optics*. Springer. https://doi.org/10.1007/978-3-030-16638-0_3

Bear, M. F., Connors, B. W., & Paradiso, M. A. (2016). *Neuroscience , Exploring*

the Brain. Wolters Kluwer.

Bernardes, R., Silva, G., Chiquita, S., Serranho, P., & et al. (2017). Retinal biomarkers of Alzheimer's disease: Insights from transgenic mouse models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10317 LNCS*, 541–550. https://doi.org/10.1007/978-3-319-59876-5_60

Blazes, M., & Lee, C. S. (2021). Understanding the Brain through Aging Eyes. *Advances in Geriatric Medicine and Research*. https://doi.org/10.20900/ agmr20210008

Bouckaert, R. R. (2003). Choosing between two learning algorithms based on calibrated tests. *ICML-Proceedings of the Twentieth International Con-ference on Machine Learning*.

Bourkhime, H., Qarmiche, N., Omari, M., Bahra, N., Tachfouti, N., Fakir, S. E., & Otmani, N. (2022). Machine learning and novel ophthalmologic biomark-ers for Alzheimer's disease screening: Systematic Review. *ITM Web of Conferences*, *43*, 01009. https://doi.org/10.1051/ITMCONF/20224301009

Carelli, V., Morgia, C. L., Ross-Cisneros, F. N., & Sadun, A. A. (2017). Optic neuropathies: the tip of the neurodegeneration iceberg. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddx273

Chan, V. T. T., Sun, Z., Tang, S., Chen, L. J., Wong, A., Tham, C. C., Wong, T. Y., Chen, C., & Mok, V. C. T. (2019). Optical coherence tomography in alzheimer's disease: A review of current literature. *Eye*, *33*, 1271–1278.

Chua, J., Li, C., Antochi, F., Toma, E., Wong, D., Tan, B., Garhöfer, G., Hi-lal, S., Cherecheanu, A. P., Chen, C. L. H., & Schmetterer, L. (2025). Utilizing deep learning to predict alzheimer's disease and mild cognitive impairment with optical coherence tomography. *Alzheimer's and Demen-tia: Diagnosis, Assessment and Disease Monitoring*, *17*. https://doi.org/ 10.1002/dad2.70041

Chua, S. Y. L., Thomas, D., Allen, N., Lotery, A., Desai, P., Patel, P., Muthy,

Z., Sudlow, C., Peto, T., Khaw, P. T., Foster, P. J., Zheng, Y., Aslam, T., Barman, S. A., Barrett, J. H., Bishop, P., Blows, P., Bunce, C., Carare, R. O., . . . Yip, J. (2019). Cohort profile: Design and methods in the eye and vision consortium of UK Biobank. *BMJ Open*, *9*(2), 1–13. https://doi.org/10.1136/bmjopen-2018-025077

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–220.

Cunha, L. P., Pires, L. A., Cruzeiro, M. M., Almeida, A. L. M., Martins, L. C., Martins, P. N., Shigaeff, N., & Vale, T. C. (2022). Optical coherence to-mography in neurodegenerative disorders. *Arquivos de Neuro-Psiquiatria*, *80*(2), 180–191. https://doi.org/10.1590/0004-282X-ANP-2021-0134

den Haan, J., Morrema, T. H., Verbraak, F. D., de Boer, J. F., Scheltens, P., Rozemuller, A. J., Bergen, A. A., Bouwman, F. H., & Hoozemans, J. J. (2018). Amyloid-beta and phosphorylated tau in post-mortem Alzheimer's disease retinas. *Acta neuropathologica communications*, *6*(1), 147. https://doi.org/10.1186/s40478-018-0650-x

Dening, T., & Sandilyan, M. B. (2015). Dementia: definitions and types. *Nursing standard (Royal College of Nursing (Great Britain) : 1987)*, *29*(37), 37–42. https://doi.org/10.7748/ns.29.37.37.e9405

Early Treatment Diabetic Retinopathy Study Research Group. (1991). Grading diabetic retinopathy from stereoscopic color fundus photographs—an ex-tension of the modified airlie house classification: Etdrs report number 10. *Ophthalmology*, *98*(5 Suppl), 786–806. https://doi.org/10.1016/S0161-6420(13)38012-9

Eren, Ö., Tek, F. B., & Turkan, Y. (2024). Segmentation based classification of retinal diseases in oct images. *UBMK 2024 - Proceedings: 9th Inter-national Conference on Computer Science and Engineering*, 890–895. https://doi.org/10.1109/UBMK63289.2024.10773527

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Gao, H., Zhao, S., Zheng, G., Wang, X., Zhao, R., Pan, Z., Li, H., Lu, F., & Shen, M. (2023). Using a dual-stream attention neural network to characterize mild cognitive impairment based on retinal images. *Computers in Biol-ogy and Medicine*, *166*. https://doi.org/10.1016/j.compbiomed.2023. 107411

Gardner, M. R., Baruah, V., Vargas, G., Motamedi, M., Milner, T. E., & Rylander, H. G. (2020). Scattering angle resolved optical coherence tomog-raphy detects early changes in 3xTg Alzheimer's disease mouse model. *Translational Vision Science and Technology*, *9*(5), 1–14. https://doi.org/ 10.1167/TVST.9.5.18

Garvin, M. K., Abràmoff, M. D., Wu, X., Russell, S. R., Burns, T. L., & Sonka, M. (2009). Automated 3-d intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. *IEEE Transac-tions on Medical Imaging*, *28*, 1436–1447. https://doi.org/10.1109/TMI. 2009.2016958

Gaugler, J., James, B., Johnson, T., Reimer, J., Solis, M., Weuve, J., Buckley R. F., & Hohman, T. J. (2022). 2022 alzheimer's disease facts and figures. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, *18*, 700–789. https://doi.org/10.1002/alz.12638

Ge, Y. J., Xu, W., Ou, Y. N., Qu, Y., Ma, Y. H., Huang, Y. Y., Shen, X. N., Chen, S. D., Tan, L., Zhao, Q. H., & Yu, J. T. (2021). Retinal biomarkers in Alzheimer's disease and mild cognitive impairment: A systematic review and meta-analysis. *Ageing Research Reviews*, *69*(April). https://doi.org/ 10.1016/j.arr.2021.101361

Hadoux, X., Hui, F., Lim, J. K., Masters, C. L., Pébay, A., Chevalier, S., Ha, J., Loi, S., Fowler, C. J., Rowe, C., Villemagne, V. L., Taylor, E. N., Fluke, C., Soucy, J. P., Lesage, F., Sylvestre, J. P., Rosa-Neto, P., Matho-taarachchi, S., Gauthier, S., . . . van Wijngaarden, P. (2019). Non-invasive in vivo hyperspectral imaging of the retina

for potential biomarker use in Alzheimer's disease. *Nature Communications*, *10*(1). https://doi.org/10. 1038/S41467-019-12242-1

Hao, J., Kwapong, W. R., Shen, T., Fu, H., Xu, Y., Lu, Q., Liu, S., Zhang, J., Liu, Y., Zhao, Y., Zheng, Y., Frangi, A. F., Zhang, S., Qi, H., & Zhao, Y. (2024). Early detection of dementia through retinal imaging and trust-worthy ai. *npj Digital Medicine*, *7*. https://doi.org/10.1038/s41746-024-01292-5

Hao, J., Shen, T., Zhu, X., Liu, Y., Behera, A., Zhang, D., Chen, B., Liu, J., Zhang, J., & Zhao, Y. (2022). Retinal structure detection in octa image via voting-based multitask learning. *IEEE Transactions on Medical Imaging*, *41*, 3969–3980. https://doi.org/10.1109/TMI.2022.3202183

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-3*(6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. G. (2019). Averaging weights leads to wider optima and better generalization. *arxiv*. http://arxiv.org/abs/1803.05407

Khawaja, A. P., Chua, S., & et al. (2020). Comparison of associations with different macular inner retinal thickness parameters in a large cohort: The uk biobank. *Ophthalmology*, *127*, 62–71. https://doi.org/10.1016/j.ophtha. 2019.08.015

Khoury, R., & Ghossoub, E. (2019). Diagnostic biomarkers of alzheimer's disease: A state-of-the-art review. *Biomarkers in Neuropsychiatry*, *1*, 100005.

Kivipelto, M., Ngandu, T., Laatikainen, T., Winblad, B., Soininen, H., & Tuomile-hto, J. (2006). Risk score for the prediction of dementia risk in 20 years among middle aged people: A longitudinal, population-based study. *TheLancet Neurology*, *5*(9), 735–741. https://doi.org/https://doi.org/10.1016/ S1474-4422(06)70537-3

Lemmens, S., & et. al. (2020). Combination of snapshot hyperspectral retinal imaging and optical coherence tomography to identify Alzheimer's disease patients. *Alzheimer's Research and Therapy*, *12*(1). https://doi.org/10.1186/s13195-020-00715-1

Li, M., Huang, K., Xu, Q., Yang, J., Zhang, Y., Ji, Z., Xie, K., Yuan, S., Liu, Q., & Chen, Q. (n.d.). Octa-500: A retinal dataset for optical coherence tomography angiography study. https://ieee- dataport.org/open- access/octa-500.

Liu, R., Yang, S., Zhong, X., Zhu, Z., Huang, W., & Wang, W. (2025). Metabolomic signature of retinal ageing, polygenetic susceptibility, and major health outcomes [Query date: 2025-05-06 08:36:53]. *The British journal of oph-thalmology*, *109*(5), 619–627. https://doi.org/10.1136/bjo-2024-325846

Liu, S., Hao, J., Xu, Y., Fu, H., Guo, X., Liu, J., Zheng, Y., Liu, Y., Zhang, J., & Zhao, Y. (2023). Polar-net: A clinical-friendly model for alzheimer's disease detection in octa images. *arXiv*. https://doi.org/10.1007/978-3-031-43990-2_57

Liu, S., Zhang, Z., Gu, Y., Hao, J., Liu, Y., Fu, H., Guo, X., Song, H., Zhang, S., & Zhao, Y. (2025). Beyond the eye: A relational model for early dementia detection using retinal octa images. *Medical Image Analysis*, *102*. https://doi.org/10.1016/j.media.2025.103513

London, A., Benhar, I., & Schwartz, M. (2013). The retina as a window to the brain - From eye research to CNS disorders. *Nature Reviews Neurology*, *9*(1), 44–53. https://doi.org/10.1038/NRNEUROL.2012.227

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Ma, Y., Hao, H., Xie, J., Fu, H., Zhang, J., Yang, J., Wang, Z., Liu, J., Zheng, Y., & Zhao, Y. (2021). ROSE: A Retinal OCT-Angiography Vessel Segmentation Dataset and New Model. *IEEE Transactions on Medical Imaging*, *40*(3), 928–939. https://doi.org/10.1109/TMI.2020.3042802

Maccora, J., Peters, R., & Anstey, K. J. (2020). What does (low) education mean

in terms of dementia risk? a systematic review and meta-analysis highlighting inconsistency in measuring and operationalising education. *SSMPopulation Health*, *12*. https://doi.org/10.1016/j.ssmph.2020.100654

Mokhtari, A., Maris, B. M., & Fiorini, P. (2025). A survey on optical coherence tomography—technology and application. *Bioengineering*, *12*(1), 65. https://doi.org/10.3390/bioengineering12010065

Nazlı, M. S., Turkan, Y., Tek, F. B., Toslak, D., Bulut, M., Arpacı, F., & Öcal, M. C. (2025). Retinal disease diagnosis in oct scans using a foundational model. *Lecture Notes in Computer Science*, *15618 LNCS*, 208–220. https://doi.org/10.1007/978-3-031-88220-3_15

Nguyen, K., Nguyen, M., Dang, K., Pham, B., Huynh, V., Vo, T., Ngo, L., & Ha, H. (2023). Early alzheimer? s disease diagnosis using an xg-boost model applied to mri images. *Biomedical Research and Therapy*, *10*(9), 5896–5911.

Nunes, A., Silva, G., & et al. (2019). Retinal texture biomarkers may help to discriminate between Alzheimer's, Parkinson's, and healthy controls. *PLoS ONE*, *14*(6). https://doi.org/10.1371/JOURNAL.PONE.0218826

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T.,Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), 1–11. https : / / doi. org/ 10 . 1186/s13643-021-01626-4

Patel, P. J., Foster, P. J., Grossi, C. M., Keane, P. A., Ko, F., Lotery, A., Peto, T., Reisman, C. A., Strouthidis, N. G., & Yang, Q. (2016). Spectral-domain optical coherence tomography imaging in 67 321 adults: Associations with macular thickness in the uk biobank study. *Ophthalmology*, *123*, 829–840. https://doi.org/10.1016/j.ophtha.2015.11.009

Ryu, S.-E., Shin, D.-H., & Chung, K. (2020). Prediction model of dementia risk

based on xgboost using derived variable extraction and hyper parameter optimization. *IEEE Access*, *8*, 177708–177720. https://doi.org/10.1109/ACCESS.2020.3025553

Sandeep, C. S., Sukesh Kumar, A., Mahadevan, K., & Manoj, P. (2019). Analysis of retinal OCT images for the early diagnosis of Alzheimer's disease. *Advances in Intelligent Systems and Computing*, *749*, 509–520. https://doi.org/10.1007/978-3-319-74808-5_43

Schindler, S. E., Galasko, D., Pereira, A. C., Rabinovici, G. D., Salloway, S., Suárez-Calvet, M., Khachaturian, A. S., Mielke, M. M., Udeh-Momoh, C., Weiss, J., et al. (2024). Acceptable performance of blood biomarker tests of amyloid pathology—recommendations from the global ceo initiative on alzheimer's disease. *Nature Reviews Neurology*, *20*(7), 426–439.

Singh, A., Balaji, J. J., Rasheed, M. A., Jayakumar, V., Raman, R., & Lakshminarayanan, V. (2021). Evaluation of explainable deep learning methods for ophthalmic diagnosis. *Clinical Ophthalmology*, *15*, 2573–2581. https://doi.org/10.2147/OPTH.S312236

Snyder, P. J., Alber, J., Alt, C., Bain, L. J., Bouma, B. E., Bouwman, F. H., Cabrera Debuc, D., Campbell, M. C. W., Carrillo, M. C., Chew, E. Y., Cordeiro, M. F., Dueñas, M. R., Fernández, B. M., Koronyo-Hamaoui, M., & Snyder, H. M. (2020). Retinal imaging in Alzheimer's and neurodegenerative diseases. *Alzheimer's & Dementia*. https:// doi. org/ 10 . 1002/alz.12179

Song, A., Johnson, N., Ayala, A., & Thompson, A. C. (2021). Optical coherence tomography in patients with alzheimer's disease: What can it tell us? *Eye and Brain*, *13*, 1–20. https://doi.org/10.2147/EB.S235238

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Student. (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25. Retrieved April 11, 2024, from http://www.jstor.org/stable/2331554

Tian, J., Smith, G., Guo, H., Liu, B., Pan, Z., Wang, Z., Xiong, S., & Fang, R.

(2021). Modular machine learning for Alzheimer's disease classification from retinal vasculature. *Scientific Reports* |, *11*, 238. https://doi.org/10.1038/s41598-020-80312-2

Triolo, G., Rabiolo, A., Shemonski, N. D., Fard, A., Di Matteo, F., Sacconi, R., Bettin, P., Magazzeni, S., Querques, G., Vazquez, L. E., Barboni, P., & Bandello, F. (2017). Optical coherence tomography angiography macular and peripapillary vessel perfusion density in healthy subjects, glaucoma suspects, and glaucoma patients. *Investigative Ophthalmology and Visual Science*, *58*(13), 5713–5722. https://doi.org/10.1167/iovs.17-22865

Tsang, S. H., & Sharma, T. (2018). *Electroretinography*. Springer International Publishing. https://doi.org/10.1007/978-3-319-95046-4_5

Turkan, Y., Tek, F. B., Arpaci, F., Arslan, O., Toslak, D., Bulut, M., & Yaman, A. (2024). Automated diagnosis of alzheimer's disease using oct and octa: A systematic review. *IEEE Access*, *12*, 104031–104051. https://doi.org/10.1109/ACCESS.2024.3434670

UK Biobank. (2022). UK Biobank - UK Biobank [[Online; accessed 2022-08-13]]. https://www.ukbiobank.ac.uk/

van der Heide, F. C. T., Khawaja, A., & et al. (2024). Associations of inner retinal layers with risk of incident dementia: An individual participant data analysis of four prospective cohort studies [Query date: 2025-05-06 08:36:53]. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, *20*(1), 211–220.

Wagner, S. K., Fu, D. J., Faes, L., Liu, X., Huemer, J., Khalid, H., Ferraz, D., Korot, E., Kelly, C., Balaskas, K., Denniston, A. K., & Keane, P. A. (2020). Insights into systemic disease through retinal imaging-based oculomics. *Translational Vision Science and Technology*, *9*(2). https://doi.org/10. 1167/tvst.9.2.6

Walters, K., Hardoon, S., Petersen, I., Iliffe, S., Omar, R. Z., Nazareth, I., & Rait, G. (2016). Predicting dementia risk in primary care: Development and validation of the dementia risk score using routinely collected data. *BMC Medicine*, *14*(1), 6. https://doi.org/10.1186/s12916-016-0549-y

Wang, X., Jiao, B., Liu, H., Wang, Y., Hao, X., Zhu, Y., Xu, B., Xu, H., Zhang, S., Jia, X., Xu, Q., Liao, X., Zhou, Y., Jiang, H., Wang, J., Guo, J., Yan, X., Tang, B., Zhao, R., & Shen, L. (2022). Machine learning based on Optical Coherence Tomography images as a diagnostic tool for Alzheimer ' s disease. *CNS Neuroscience and Therapeutics*, (May), 1–12. https://doi. org/10.1111/cns.13963

Wang, X., Li, H., Xiao, Z., Fu, H., Zhao, Y., Jin, R., Zhang, S., Kwapong, W. R., Zhang, Z., Miao, H., & Liu, J. (2022). *Screening of Dementia on OCTA Images via Multi-projection Consistency and Complementarity* (Vol. 13432 LNCS). Springer Nature Switzerland. https:// doi. org/ 10 . 1007/978-3-031-16434-7_66

WHO. (2014). *International statistical classification of diseases and related health problems 10th Revision.* World Health Organization. https://icd.who.int/ browse10/2014/en%5C#/F00%5C-F09

WHO. (2022). World Health Organization: Dementia Fact Sheet. https://www. who.int/news-room/fact-sheets/detail/dementia

Wikipedia. (2025). Diagram of the eye and placement of the retinal implants. https://commons.wikimedia.org

Wisely, C. E., Wang, D., Henao, R., Grewal, D. S., Thompson, A. C., Robbins, C. B., Yoon, S. P., Soundararajan, S., Polascik, B. W., Burke, J. R., Liu, A., Carin, L., & Fekrat, S. (2022). Convolutional neural network to iden-tify symptomatic Alzheimer's disease using multimodal retinal imaging. *British Journal of Ophthalmology*, *106*(3), 388–395. https://doi.org/10. 1136/bjophthalmol-2020-317659

Wisely, C. E., Wang, D., Henao, R., Grewal, D. S., Thompson, A. C., Robbins, C. B., Yoon, S. P., Soundararajan, S., Polascik, B. W., Burke, J. R., Liu, A., Carin, L., & Fekrat, S. (2024). A convolutional neural network using multimodal retinal imaging for differentiation of mild cognitive impair-

ment from normal cognition. *Ophthalmology Science*, *4*. https://doi.org/10.1016/j.xops.2023.100355

Xu, C. (2025). Octa image-based machine learning models for discriminating alzheimer's disease from neurodegenerative and ocular conditions, 324–331. https://doi.org/10.5220/0013141300003911

Xu, J. J., Zhou, Y., Wei, Q. J., Li, K., Li, Z. P., Yu, T., Zhao, J. C., Ding, D. Y., Li, X. R., Wang, G. Z., & Dai, H. (2022). Three-dimensional diabetic macular edema thickness maps based on fluid segmentation and fovea detection using deep learning. *International Journal of Ophthalmology*, *15*(3), 495–501. https://doi.org/10.18240/ijo.2022.03.19

Yanagihara, R. T., Lee, C. S., Ting, D. S. W., & Lee, A. Y. (2020). Methodological challenges of deep learning in optical coherence tomography for retinal diseases: A review. *Translational Vision Science and Technology*, *9*(2), 17–19. https://doi.org/10.1167/tvst.9.2.11

Yi, F., Yang, H., Chen, D., Qin, Y., Han, H., Cui, J., Bai, W., Ma, Y., Zhang, R., & Yu, H. (2023). Xgboost-shap-based interpretable diagnostic framework for alzheimer's disease. *BMC medical informatics and decision making*, *23*(1), 137.

Yoon, J. M., Lim, C. Y., Noh, H., Nam, S. W., Jun, S. Y., Kim, M. J., Song, M. Y., Jang, H., Kim, H. J., Seo, S. W., Na, D. L., Chung, M. J., Ham, D. I., & Kim, K. (2024). Enhancing foveal avascular zone analysis for alzheimer's diagnosis with ai segmentation and machine learning using multiple radiomic features. *Scientific Reports*, *14*. https : / / doi . org/ 10 . 1038/s41598-024-51612-8

You, J., Zhang, Y.-R., Wang, H.-F., Yang, M., Feng, J.-F., Yu, J.-T., & Cheng, W. (2022). Development of a novel dementia risk prediction model in the general population: A large, longitudinal, population-based machine-learning study. *eClinicalMedicine*, *53*, 101665. https://doi.org/10.1016/j.eclinm.2022.101665

Yousefzadeh, N., Tran, C., Ramirez-Zamora, A., Chen, J., Fang, R., & Thai, M. T. (2024). Neuron-level explainable ai for alzheimer's disease assess-

ment from fundus images [Query date: 2025-05-06 08:36:53]. *Scientific reports*, *14*(1), 7710. https://doi.org/10.1038/s41598-024-58121-8

Zhang, Y., Shen, S., Li, X., Wang, S., Xiao, Z., Cheng, J., & Li, R. (2024). A multiclass extreme gradient boosting model for evaluation of transcriptomic biomarkers in alzheimer's disease prediction. *Neuroscience letters*, *821*, 137609.

Zhong, Y., Chalise, P., & He, J. (2023). Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data. *Communications in Statistics: Simulation and Computation*, *52*, 110–125. https://doi.org/10.1080/03610918.2020.1850790

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., Liu, T., Xu, M., Lozano, M. G., Woodward-Court, P., Kihara, Y., Allen, N., Gallacher, J. E., Littlejohns, T., Aslam, T., Bishop, P., Black, G., Sergouniotis, P., Atan, D., . . . Keane, P. A. (2023). A foundation model for generalizable disease detection from retinal images. *Nature*, *622*, 156–163. https://doi.org/10.1038/s41586-023-06555-x

CURRICULUM VITAE