An Efficient Deep Neural Architecture for Detecting Different Stages of Age-Related Macular Degeneration

by

Iftekher Uddin Rafi
22301549
Antara Shamiha
24341148
Md. Ibrahim Akanda
21201233
Richa Saha
20201081

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 2026

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Iftekher Uddin Rafi 22301549	Antara Shamiha 24341148
Md. Ibrahim Akanda 21201233	Richa Saha 20201081

Approval

The thesis/project titled "**An Efficient Deep Neural Architecture for Detecting Different Stages of Age-Related Macular Degeneration**" submitted by

1. Iftekher Uddin Rafi(22301549)
2. Antara Shamiha(24341148)
3. Md. Ibrahim Akanda(21201233)
4. Richa Saha(20201081)

Of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 31, 2026.

Examining Committee:

Supervisor:
(Member)

Md Sabbir Ahmed
Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Dr. Md. Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Age-related macular degeneration (AMD) is a leading cause of vision loss in the elderly population, characterized by progressive retinal degeneration that advances through distinct stages: early, intermediate, and late (advanced). Clinical management, early intervention, and prevention of irreversible visual impairment require proper and timely diagnosis of these stages. Nevertheless, the manual interpretation of retinal images is still laborious and prone to interobserver error, which demonstrates the necessity of the use of high-quality automated diagnostic systems. Our work in this paper suggests an effective deep neural network to identify and classify the stages of AMD with the use of retinal imaging, namely, optical coherence tomography (OCT). The research first performs a binary detection to make the distinction between AMD and Non-AMD cases. We then categorise the cases of detected AMD on the basis of early, intermediate and late stages. The framework proposed makes use of transfer learning using various state-of-the-art convolutional neural network (CNN) backbones, such as ResNet50, EfficientNetB4, ConvNeXt Small, ConvNeXt Tiny, and EfficientNetV2_L that have been selected due to their effective representational ability and efficient computing. These are fine-tuned models that are used to observe subtle pathological changes related to drusen accumulation, pigmentary defects, geographic atrophy, and neovascularization throughout the disease range. The architecture assumes the combination of the standardized preprocessing, efficient training methods, and the relative assessment of the chosen networks to find an effective and resource-efficient model of the multi-class AMD staging. Experimental findings demonstrate good performance. ConvNeXt Small reaches 97.9% accuracy in detection and 83.5% in staging. It outperforms ResNet50 (97.5% detection, 77.4% staging), EfficientNetB4 (97.5% detection, 69.9% staging), ConvNeXt Tiny (96.7% detection, 78.2% staging), and EfficientNetV2_L (96.7% detection, 80.4% staging). All models improve early-stage sensitivity, with recalls above 86%. Experimental results demonstrate that the proposed approach achieves robust performance across all disease stages, with improved sensitivity in early-stage detection. This work highlights the potential of modern deep learning models as scalable decision-support tools for large-scale screening and clinical diagnosis, contributing to improved outcomes in the management of age-related macular degeneration.

Keywords: Age-related macular degeneration (AMD), optical coherence tomography (OCT), deep learning, convolutional neural networks (CNN), transfer learning, disease staging, hierarchical classification, explainable AI (XAI)

Table of Contents

Chapter 1

Introduction

1.1 Introduction

Age-related macular degeneration (AMD) ranks as a top cause of irreversible vision loss in older adults worldwide, especially in developed countries[5]. Doctors classify AMD into three stages: early, intermediate, and late (advanced). Late AMD splits into dry (geographic atrophy) and wet (neovascular) forms[2]. Small drusen and slight pigment changes mark early AMD with little effect on sight. Intermediate AMD shows larger drusen or pigment issues that raise progression risk. Late AMD causes major central vision loss from geographic atrophy in dry cases or vessel growth and leakage in wet cases[18].

Global AMD cases reached about 196 million in 2020 and will climb to 288 million by 2040 due to population aging[5]. Recent estimates put worldwide cases near 200 million today. In the US alone, nearly 20 million adults aged 40 and older live with some form of AMD, including 1.49 million with late-stage vision-threatening AMD. We need early detection and precise staging. Timely steps slow the disease. Anti-VEGF injections treat wet AMD and preserve vision in many patients. For example, studies show 63% of eyes lose 15 letters or fewer over 10 years with ongoing treatment. Nutritional supplements help intermediate cases reduce progression risk.

Manual review of retinal images takes time. Experts often disagree. Shortages of ophthalmologists limit access. Color fundus photos and OCT scans serve as main tools.

Deep learning solves these problems. Convolutional neural networks (CNNs) analyze images automatically. They learn detailed features from data. This delivers accurate, consistent results for mass screening. Transfer learning boosts results on small medical datasets. Pre-trained models cut training needs. New architectures deliver strong accuracy with low computing demands. Clinics in limited-resource areas adopt them easily.

This study, "An Efficient Deep Neural Architecture for Detecting Different Stages of Age-Related Macular Degeneration," presents a system. It applies transfer learning to ResNet50, EfficientNetB4, ConvNeXt Small, ConvNeXt Tiny and EfficientNetV2_L. These backbones classify AMD stages from OCT images. The comparison finds the best efficient model for multi-class detection. We gain a practical tool for quick staging. Early action improves our outcomes in AMD care.

1.2 Research Problem

Existing approaches fall short in several critical areas. Many models limit themselves to binary tasks like AMD presence versus absence. They overlook subtle differences that separate early, intermediate, and late stages[19][23]. Early-stage changes remain faint on OCT scans. These faint features prove hard to detect reliably. Models often show lower accuracy for early detection as a result[19]. Datasets used in training frequently lack balance across stages. This imbalance biases models toward more common classes and weakens performance on underrepresented early or intermediate cases[29][26].

Models struggle with generalization when applied to varied clinical scans from different devices or populations. Noise, artifacts, and inter-class visual similarities in OCT images contribute to this issue[30][26]. Resource-heavy architectures require substantial computational power. This demand makes deployment difficult in low-resource clinics or remote settings[32][27].

Few studies perform direct head-to-head comparisons of lightweight, high-performance backbones such as ResNet50, EfficientNetB4, ConvNeXt Small, and ConvNeXt Tiny. These comparisons focus specifically on multi-class AMD staging (early, intermediate, late dry, late wet) using OCT images[23][32].

Age-related macular degeneration (AMD) is clinically classified into three main stages **early**, **intermediate**, and **late** (also called advanced) primarily based on the Age-Related Eye Disease Study (AREDS) criteria and subsequent refinements [2] [1]. These stages reflect progressive macular changes visible on fundus photography and optical coherence tomography (OCT), with increasing risk of vision loss.

Early AMD is characterized by the presence of small drusen (<63 μm, often called "hard" drusen) or a few intermediate drusen (63–124 μm), with no pigmentary abnormalities or significant visual symptoms. These changes are typically asymptomatic and detected incidentally during routine eye exams. On OCT, early AMD may show subtle sub-RPE deposits or minor drusen without disruption of outer retinal layers [4][9].

Intermediate AMD involves more extensive drusen either numerous intermediate drusen or at least one large drusen (125 μm, often "soft" or confluent) and/or mild pigmentary abnormalities (e.g., retinal pigment epithelium [RPE] hyper- or hypopigmentation) without advanced features. Patients may experience mild visual symptoms, such as difficulty with low-light reading or delayed dark adaptation, but central vision is usually preserved. OCT frequently reveals larger drusen with associated RPE elevation, early hyperreflective foci, or subtle outer retinal changes [2][6].

Late (advanced) AMDis subdivided into two forms:

- **Dry (atrophic or geographic atrophy [GA])** — characterized by well-defined areas of RPE and photoreceptor loss (geographic atrophy), leading to progressive central vision loss. On OCT, GA appears as hypertransmission of signal into the choroid with loss of outer retinal bands and RPE [9][7].

- **Wet (neovascular or exudative)**— defined by abnormal choroidal neovascularization (CNV), resulting in leakage, hemorrhage, or subretinal fluid. OCT shows subretinal/intraretinal fluid, fibrovascular pigment epithelial detachment (PED), or hemorrhagic PED, often with rapid vision decline if un-

treated [2][10].

CNV stands for choroidal neovascularization. It describes abnormal blood vessels that grow from the choroid through breaks in Bruch's membrane into the sub-retinal pigment epithelium or subretinal space. **MNV** stands for macular neovascularization. Experts now use this term for neovascular changes in the macula in age-related macular degeneration [10]
[16] [13] . MNV includes all forms of neovascularization in the macula, such as type 1, type 2, and type 3. CNV refers only to vessels that originate from the choroid. MNV covers broader neovascular processes, including those that start in the retina (type 3 MNV). The 2020 consensus nomenclature recommends MNV over CNV for AMD cases. This change reflects that neovascularization in AMD often involves the macula and includes retinal-origin vessels. Non-exudative MNV appears in intermediate AMD and carries risk of progression to exudative forms. Exudative MNV defines late-stage AMD and requires treatment.
This study focuses on the multi-class classification of age-related macular degeneration (AMD) into three clinically relevant stages: early, intermediate, and late (advanced). These classifications, derived from the AREDS simplified severity scale and clinical refinements, guide risk stratification and treatment decisions (e.g., AREDS2 supplements for intermediate AMD, anti-VEGF for wet AMD). This thesis addresses the problem: Develop and evaluate an efficient deep neural architecture. Transfer learning on the chosen CNN backbones to classify AMD stages versus OCT images. Attain excellent overall accuracy, great sensitivity to early-stage detection, and minimal computational cost. Provide a practical resource that can be used in clinical practice in varied international contexts

1.3 Research Objectives

Age-related macular degeneration (AMD) has a silent developmental process in the initial and middle stage. This silent evolution slows down the diagnosis and increases the chances of forever blindness among rapidly aging populations all over the world [5]. Nutritional supplements help intermediate cases reduce progression risk[28]. Color fundus photos and OCT scans serve as main tools[3].
Existing models limit themselves to binary tasks. They overlook subtle differences between early, intermediate, and late stages [1][2]. Early-stage changes appear faint on OCT scans. These features prove hard to detect reliably [19]. Datasets lack balance across stages. This imbalance weakens performance on early or intermediate cases [23][26]. Models struggle with generalization to varied clinical scans from different devices or populations [27][29]. Resource-heavy architectures slow deployment in clinics [17]. Few studies compare lightweight backbones like ResNet50, EfficientNetB4, ConvNeXt Small, ConvNeXt Tiny and EfficientNetV2_L for multi-class AMD staging on OCT images [20].
This thesis pursues these objectives.

- To develop an efficient deep neural architecture for multi-class classification of age-related macular degeneration (AMD) stages (early, intermediate, and late/advanced) using optical coherence tomography (OCT) images.

- To apply transfer learning using state-of-the-art convolutional neural network backbones specifically ResNet50, EfficientNetB4, ConvNeXt Small, and ConvNeXt Tiny and compare their effectiveness for AMD stage classification.
- To evaluate and compare the performance of the developed models in terms of overall accuracy, precision, recall, and F1-score across the different AMD stages.
- To achieve high sensitivity (recall), particularly for the early-stage detection of AMD, enabling timely clinical intervention and prevention of irreversible vision loss.
- To minimize computational cost, model size, and inference time, thereby creating a practical and deployable solution suitable for resource-limited clinical settings and low-resource environments.
- To assess the generalization capability and robustness of the proposed models by evaluating their performance on diverse, independent OCT datasets to ensure reliability in real-world clinical scenarios.
- To provide a resource-efficient, accurate, and clinically applicable deep learning tool that enhances global AMD screening, early diagnosis, and accurate staging processes.

Chapter 2

Literature Review

2.1 Summary

Leingang et al. (2023) present an empirical deep learning research identifying the clinically crucial problem of automated diagnosis and staging of age-related macular degeneration (AMD) from real-world 3D OCT data, a setting mainly overlooked by previous researches that depends on curated medical datasets or single 2D B-scans. AMD demonstrates diverse structural retinal transitions across stages, making consistent staging challenging. Automated OCT-based AI systems can reduce inter-grader variability, improve early detection, and support scalable AMD screening in realworld clinical environments [17]. This study portrays that deep learning models can consistently detect and stage age-related macular degeneration (AMD) from large, real-world OCT datasets with performance comparable to that of retinal specialists. The PINNACLE framework shows strong generalizability across heterogeneous clinical data, highlighting the feasibility of deploying AI tools in routine ophthalmic practice. Using retrospective OCT volumes collected between 2007 and 2019 from two major UK eye hospitals under the PINNACLE consortium, the authors assembled a dataset of 3,995 Topcon OCT volumes and introduced a two-stage deep learning pipeline in which a 2D ImageNet-pretrained ResNet50 first classifies individual B-scans and extracts latent features, which are then aggregated and analyzed by lightweight ResNet models to predict the existance of four clinically meaningful biomarkers DRUSEN (iAMD), macular neovascularization (nAMD), macular atrophy (GA), and NORMAL at the volume level, while also providing B-scan-level localization and uncertainty estimation via Monte-Carlo dropout. The approach was evaluated against fully 3D ResNet baselines using robust metrics, including ROCAUC, balanced accuracy, and Matthews correlation coefficient, achieving strong performance on a retinal-expert-graded hold-out test set with an average ROC-AUC of approximately 0.94, and demonstrating competitive or superior findings compared to 3D CNNs despite lower computational complexity. External validation on the Duke dataset (ROC-AUC = 0.994) and population-scale UK Biobank data further showed good generalization across scanners and the ability to detect generic retinal fluid and atrophy beyond AMD-specific cases. However, there were limitations including scanner dependence, challenges in drusen detection, and lack of public data or code, this work makes a significant contribution by demonstrating a scalable, uncertainty-aware framework for staging AMD in heterogeneous real-world OCT datasets, enabling efficient retrospective analysis and large-scale

clinical research.

The study by Neri et al. (2025) is an empirical deep learning–based study that addresses the clinically significant challenge of automatically classifying macular neovascularization (MNV) with three subtypes type 1, type 2, and type 3 in patients with age-related macular degeneration (AMD) using constructive optical coherence tomography (OCT) images, a task that is important because MNV subtype identification directly influences prognosis and therapeutic decisions yet currently relies on expert evaluation that is time-consuming and subject to inter-observer variability[31]. Different MNV subtypes in AMD are associated with distinct projections and treatment responses. Reliable automated subtype classification can help clinicians in personalized treatment planning and reduce dependency on invasive multimodal imaging. The authors developed a complete automated pipeline using a retrospective dataset of 193 eyes collected from a single tertiary referral center in Italy, acquired with Heidelberg SPECTRALIS OCT systems under standardized clinical protocols and ethical approval. A major contribution of the work is the introduction of an anatomically informed "homogenization" preprocessing strategy, in which retinal layers are segmented and OCT volumes are spatially aligned and rescaled to a fixed three-dimensional representation, thereby reducing anatomical variability and ensuring that the deep learning models focus on clinically applicable retinal regions. After preprocessing, the study analyzed multiple 3D convolutional neural network architectures, including DenseNet201, ResNet50, ResNeXt-101, and a custom architecture named MnvNet, enabling systematic comparison of different model architectures within the same experimental setup. Model training and evaluation were performed using 5-fold stratified cross-validation with internal validation for early stopping, a robust approach that reduces overfitting given the relatively limited dataset size.The results consistently illlustrated that homogenized inputs significantly performed better than non-homogenized OCT volumes across all architectures when performance was evaluated using clinically important metrics like sensitivity, specificity, and receiver operating characteristic area under the curve (ROC-AUC). With ROC-AUC values of approximately 0.95 for type 1, 0.97 for type 2, and 0.91 for type 3, the top-performing configuration, MnvNet with homogenization, achieved strong discriminative performance for all MNV subtypes along with high sensitivity and specificity, especially for the more common subtypes. To improve interpretability, the researchers used occlusion sensitivity analysis, which highlighted retinal regions most dominant for classification and showed alignment with known pathological features of MNV, thereby boosting medical trust in the model's predictions. The technique is novel as it combines 3D OCT-based deep learning with anatomically guided preprocessing for fine-grained MNV subtype classification, going beyond earlier research that normally depends on multimodal imaging or binary AMD detection. However, the authors also acknowledge key constraints, including the model sample size, single-center data source, lack of external validation, and absence of publicly accessible code or datasets, which may restrict generalizability and reproducibility. Despite these limitations, the research proposed compelling evidence that deep learning models, when paired with anatomically informed preprocessing, can accurately and reliably classify MNV subtypes from structural OCT alone, emphasizing strong potential for future clinical decision-support systems in AMD management while illustrating the need for larger, multi-center validation

studies.

Arrigo et al. (2021) supervised a potential interventional clinical research exploring macular neovascularization (MNV) in age-related macular degeneration, central serous chorioretinopathy (CSC), and best vitelliform macular dystrophy (BVMD) using quantitative optical coherence tomography angiography to identify disease-specific morpho-functional attributes. AMD-associated MNV is structurally more complex and severe than MNV in other retinal diseases. Quantitative OCTA biomarkers can assist in differentiating AMD-related neovascularization and enhance disease-specific diagnosis and supervision[12]. Although MNV is present in a variety of macular conditions, few studies have directly evaluated its quantitative OCTA features across various diseases, which limits the accuracy of diagnosis and treatment planning. This study fills a key gap in retinal research as it incorporated 98 eyes from 98 consecutive patients with naïve MNV: 66 AMD, 18 CSC, and 14 BVMD, all imaged with multimodal retinal imaging and followed for one year post-treatment with anti-VEGF therapy or photodynamic therapy. Using image skeletonization and Fijibased workflows, the authors assessed quantitative OCTA features such as best-corrected visual acuity (BCVA), central macular thickness (CMT), vessel density of superficial, deep, and choriocapillaris plexa, as well as MNV vessel tortuosity (VT) and vessel dispersion (Vdisp). These measurements were critical in vascular networks description of MNV as well as cross-disease group comparisons. Even though the formal machine learning models were not used, the methodology involved high-resolution OCT/OCTA imaging, automated binarization and skeletonization, and statistical analysis, including ANOVA and Bonferonni-corrected post-hoc testing; the prior research was used as a baseline. Main results depicted that AMD-related MNV disclosed greater VT and larger lesion areas compared to CSC or BVMD, Vdisp, and speckled vascular patterns further distinguished disease-specific characteristics. Nonetheless, it has certain drawbacks, such as small sample sizes in CSC and BVMD, the lack of code or external validation of replication, and the possibility of single-center selection bias. . Despite these barriers, the work provides insightful quantitative remarks of disease-specific MNV behavior, facilitating more accurate diagnosis and customized treatment approaches in clinical ophthalmology.

Clemens, Eter, and Alten (2024) demonstrated a comprehensive review of type 3 macular neovascularization (MNV3) in age-related macular degeneration, identifying the crucial necessity to better acknowledge this distinct neovascular subtype, which varies from types 1 and 2 in clinical advancement and its emergence[22] A distinct AMD phenotype, type 3 MNV is strongly connected with bilateral disease and reticular pseudodrusen. Since the course of the disease and the response to treatment are different from those of other neovascular types, early detection with OCT and OCTA is key. In order to specify MNV3, the review article combines factual observations from histopathology,clinical studies and multimodal imaging. It focuses on how the virus first emerged in the deep vascular plexus without being associated to the retinal pigment epithelium and how it evolved across OCT-classified stages 1-3. The review points out the epidemiological importance of MNV3, identifies that it is more common than previously thought, arterial hypertension, has substantial associations with bilaterality, advanced age, closely associated with reticular pseudodrusen, a possible precursor lesion. In order to establish a coherent staging

framework and identify the morphological and functional attributes of MNV3, such as hyperreflective foci in the outer retina, the authors methodologically combine data from optical coherence tomography (OCT), fluorescein angiography, OCT angiography and histologic examinations. Although the paper does not present unique experimental data, it rationally evaluates existing therapeutic approaches, diagnostic imaging protocols, and their outcomes, indicating that anti-VEGF therapy remains strong, though treatment reaction may vary by stage and lesion difficulty. The review highlights insufficient understanding of early lesion pathogenesis, possible biases in retrospective imaging studies, and a lack of longitudinal datasets to assess long-term outcomes and recurrence rates are among the deficiencies in current knowledge. It also interpretes how existing staging and treatment standard need to be authinticated by uniform imaging criteria and longterm multicenter studies. By combining previous histopathologic observations with advanced imaging modalities, Clemens et al. (2024) provide a complete architecture for identifying, classifying, and managing MNV3, offering practical medical perceptions and establishing a foundation for future work into automated detection using deep learning, risk stratification, and tailored therapeutic plans. The review is extremely essential to researchers and physicians to create the much-needed gap in the task of managing AMD: identification of MNV3 among other neovascular subtypes should be done early, and individualized treatment plans should be developed. Overall, this fusion guides future research and technology advancements in ophthalmology while highlighting the prevalence, intricacy and clinical significance of MNV3.

Vaghefi et al. (2020) implemented this research in an experimental clinical trial that executed a convolutional neural network (CNN) to diagnose moderate dry age-related macular degeneration (AMD)[11]. In order to amplify diagnosis accuracy with deep learning, it evaluates the combination of multiple imaging mechanisms, such as Optical Coherence Tomography (OCT), OCT Angiography (OCT-A) and color fundus photography (CFP). To enhance and improve diagnostic accuracy, the convolutional neural network (CNN) underwent training on several visual modalities at once. The most common reason behind irreversible vision loss in people with age over 60 is age-related macular degeneration which is in short known as AMD. The specialists spent a significant amount of time in the diagnosis and monitoring of AMD by using manually interpreted images. Although there are multiple image processing models available.The key objective of this particular research surrounds is to figure out whether combining these modalities into a CNN could accelerate the diagnostic procedure, have higher accuracy, and reduce the error of intermediate dry AMD detection. Three cohorts of participants, including 75 individuals from Auckland, New Zealand, were established: the categories of young healthy individuals, aged healthy individuals, and those with intermediate dry AMD. Participants' OCT, OCT-A, and CFP scans were gathered. After then, the dataset was separated into training (60 percent) and testing (20 percent) sets. In order to prevent data leaking and guarantee patient-level validation, this split was made based on the participants rather than photos.Any posterior eye conditions that could affect the choroidal or retinal vasculature, which includes hypertensive retinopathy, DR, polypoidal choroidal vasculopathy, glaucoma, as well as extreme myopia were not included in the study(6D).

Schranz et al. (2022) present an empirical clinical study that investigates the relationships between vascular and fluid-related criteria in neovascular age-related macular degeneration (nAMD) by leveraging deep learning–based optical coherence tomography (OCT) analysis, addressing a key gap in understanding how automated fluid quantification correlates with established clinical and imaging biomarkers of disease function and treatment response in nAMD.Retinal fluid accumulation is a fundamental indicator of neovascular AMD activity[20]. Automated fluid quantification using AI can improve objective assessment of disease severity and guide personalized anti-VEGF treatment approaches. nAMD is a prominent factor of vision loss among older adults and is characterized by macular neovascularization and associated fluid accumulation that can result in irreversible retinal damage if inadequately treated; while previous research has used OCT and OCT angiography to assess fluid and vascular changes qualitatively, few studies have systematically linked quantitative fluid biomarkers with vascular features and clinical outcomes using artificial intelligence. The study retrospectively analyzed volumetric OCT scans from patients with nAMD treated in routine clinical practice, using a previously developed deep learning model to automatically quantify retinal fluid volumes and derive vascular parameters from the same scans. Deep learning allowed precise segmentation of intraretinal, subretinal, and sub-retinal pigment epithelial fluid compartments, enabling detailed volumetric characterization of fluid burden that was then statistically related to vascular indicators such as vessel density and flow patterns detected on OCT angiography. The authors assessed against fluid and vascular metrics to visual acuity and treatment history to explore whether specific quantitative features could reflect disease severity and therapeutic response. Instead of other artificial intelligence baselines, model performance and parameter correlations were evaluated using standard statistical analyses, such as regression models and correlation coefficients; deep learning automation itself acted as the methodological innovation by enabling reliable and consistent quantification across sizable OCT datasets. Core findings suggested that fluid volume may be a more direct indicator of treatment demand than vascular morphology only. Automated fluid metrics portrayed stronger correlations with regular measures of disease activity, such as central retinal thickness and clinical grading of exudation, while vascular parameters illustrated weaker associations. These findings backs earlier findings that subretinal fluid volume can impact treatment frequency and visual outcomes, but the use of deep learning quantification boosts the accuracy and objectivity of these correlations. The promise and complexity of AI-based biomarkers in controlling nAMD are demonstrated by the study's consistent analysis of fluid and vascular data using an automated deep learning pipeline applied to regular clinical OCTs. Its retrospective nature, possible selection bias from single-center data, and lack of prospective validation or external datasets that could reduce generalizability are some of its drawbacks. However, by showing how quantitative OCT biomarkers gained from deep learning might improve our understanding of disease causes and perhaps direct individualized treatment techniques in neovascular AMD, this study enhance the field.

Wu and Liu (2022) conducted a review paper that summarizes existing oculomics studies using DL models to analyze retinal images [15]. The work focuses on the potential of retinal imaging as a predictive tool for systemic diseases by using machine

learning models. The issue addressed by this review is how deep learning-based retinal image analysis can be used. We can use it in order to Diagnose and assess the risk for conditions like cardiovascular diseases, neurodegenerative diseases, chronic kidney disease and Alzheimer's disease The review discusses the potential of DL to improve noninterfering diagnosis but also touches on the challenges in clinical application and generalizability. It can also be used to Predict systemic health biomarkers (e.g., age, sex, cardiovascular risk). The datasets used here are UK Biobank and EyePACS . The primary methodology discussed here involves the application of deep learning models, such as CNN for analyzing retinal images focusing on OCT and Fundus photography. The review mentions various performance metrics from the studies, such as: cardiovascular disease prediction which has an area under the curve between 0.7 and 0.88 across different studies, Blood pressure and hypertension prediction which has Mean absolute error (MAE) for systolic and diastolic BP predictions ranging from 6 to 15 mmHg. Besides, it includes age and sex prediction. The review paper has some new findings and comparisons. Some DL models showed spectacular "super-human" capabilities in predicting systemic markers like age, sex, and CVD risk, outperforming classical methods. The review highlights the need for more studies incorporating multimodal data (e.g., combining OCT with fundus images) and clinical metadata (e.g., age, ethnicity) to improve prediction accuracy, as shown in some studies.OCT models were generally less robust than fundus-based models, especially when predicting systemic biomarkers. The review paper also found lackings and challenges. Many of the DL models struggled with external validation, often failing to generalize across different populations or ethnicities.While DL models performed well in predicting some variables. However, they struggled with others, like thyroid function, blood glucose levels, and lipid profiles. Most of the reviewed studies only used a single imaging modality (either fundus photography or OCT), and more research is needed to explore multimodal approaches and use them combinedly.

Leng et al. (2023) state that deep learning-based diagnostic systems, especially convolutional neural networks (CNNs), have exceptionally high accuracy in detecting age-related macular degeneration (AMD), as proven by a systematic review and meta-analysis of 18 research investigations involving 56 models and more than 778,000 retinal images gathered through OCT, fundus photography, OPTOS, and multimodal sources[18]. AMD is a leading cause of irreversible vision loss, and early detection is important to prevent progression to advanced stages. AI-based analysis of fundus and OCT images can enable scalable, objective, and cost-effective AMD screening, more specifically in settings with limited access to retinal specialists. Compared to previous AMD analyses that relied on traditional machine learning or weak architectures with handcrafted features and limited generalizability, this paper reports significantly improved diagnostic performance, with a pooled sensitivity of 94%, specificity of 97%, and an AUC of 0.9925.ResNet-based models routinely outperformed CNN variations such as VGG and AlexNet, especially to their residual connections, which successfully handle the vanishing slope problem and enable deeper, better stable training. The authors also identified the kind of AMD (dry versus wet) and network depth as the key factors to performance variance, noting that additional depth does not necessarily guarantee improved accuracy and may lead to overfitting a problem that has been highlighted in earlier AMD-focused deep

learning research.In contrast to Transformer-based methods, which typically require significantly larger datasets and are still understudied in AMD-specific diagnostic tasks, CNNs continue to be more dependable and widely used for AMD detection, primarily because of their strong inductive bias for local spatial feature extraction and their capacity to perform well on relatively small medical imaging datasets. In comparison to other deep learning architectures, this study presents CNN-based models, especially ResNets, trained on OCT data as a more developed and clinically feasible option. However, it also highlights the necessity of further research incorporating larger, heterogeneous datasets and sophisticated architectures to further enhance robustness and practical applicability.

Alenezi et al. (2024) discusses that Age-related macular degeneration (AMD) damages the central retina and ranks as a top cause of vision loss worldwide[21]. It affects about 8.7% of people now, and experts project 288 million cases by 2040. Doctors use optical coherence tomography (OCT) as the main tool for detailed retinal scans. OCT supports AI applications well because it shows clear cross-sections of the retina. Early AMD detection helps prevent severe vision loss. Deep learning models spot drusen and retinal pigment epithelium changes with high accuracy. These models match or beat retinal specialists in spotting early signs. Studies by He et al. and Pead et al. show deep learning and local outlier factor methods detect AMD lesions automatically in OCT and color fundus images. Intermediate AMD requires close watch for progression to wet forms. Models predict exudation risk in non-exudative eyes. Banerjee et al. built a hybrid model with imaging features plus patient demographics for personalized risk assessment. Some models add genetic data to fundus images for better progression forecasts. Transcriptome studies confirm AMD involves the whole body, not only the retina. Late AMD often features choroidal neovascularization (CNV), known as wet AMD. Ensemble models classify CNV from normal eyes with precision of 96.59% and recall of 98.27%. Hybrid attention networks like MHANet reach classification accuracy of 99.76% by focusing on lesion areas and cutting background noise. Models sometimes mistake active wet AMD for inactive cases. Experts stress ongoing refinement with clinician input to fix these errors. Modern AI moves to ensemble setups for stronger results. Teams combine ResNet, EfficientNet, and attention layers. Trainable weights adjust during training to blend model strengths. Class activation maps show clinicians the exact features, like fluid or new vessels, behind AI decisions. This builds trust and aids real-world use. You gain faster, more reliable AMD screening and staging from these OCT-based AI tools.

Alsaih et al. (2020) analyze and contrast four deep learning architectures FCN, U-Net, SegNet, and Deeplabv3+ for computerized classification of age-related macular degeneration (AMD) which associated retinal fluids from spectroscopic optical coherence tomography (SD-OCT) volumes, focusing on the limitations of preceding machine learning based and single architecture CNN approaches. Unlike previous studies that focused on single fluid types or limited datasets, this study systematically assesses three major retinal fluids (IRF, SRF, and PED) using large scale open comparisons from the RETOUCH and OPTIMA challenges, with an emphasis on both internal-dataset performance and cross-dataset generalization[8]. This study's primary contribution is to show that patch driven training outperforms

whole image training across all architectures by enhancing regional feature learning and class imbalance management. Patched Deeplabv3+ and extended U-Net models outperform typical U-Net-centric segmentation pipelines in terms of Dice similarity coefficients, demonstrating the efficacy of multi-scale contextual learning and various spatial pyramid pooling for OCT-based AMD segmentation. Furthermore, the study demonstrates that fine-tuning pre-trained networks across scanner vendors enhances generalization while shortening training time, emphasizing the relevance of transfer learning and architecture comparison in developing effective clinical-grade AMD segmentation systems.

Sotoudeh-Paima et al. (2022) present a multi-scale convolutional neural network based on a feature pyramid network (FPN) framework to get beyond a basic drawback of traditional OCT-based AMD classifiers, which is their limited receptive field and inadequate sensitivity to lesion-scale heterogeneity across retinal layers[14]. Drusen—extracellular deposits under the retinal pigment epithelium—are hallmark biomarkers of early and intermediate AMD stages and strongly impact disease progression risk. The network's multi-scale design boosts detection of drusen and associated retinal irregularities, enabling reliable AMD classification and providing a basis for threat assessment and early intervention. Larger-scale choroidal neovascularization (CNV) structures and fine-grained drusen patterns were frequently not concurrently captured by earlier CNN-based methods, which mostly relied on single-resolution feature maps taken from deep backbones like VGG or ResNet. Instead of using computationally costly multi-model ensembles or image pyramid pipelines, this suggested FPN-CNN explicitly integrates hierarchical feature combinations across various scales of space within a single end-to-end architecture, allowing the model to learn locally prevalent micro-structural abnormalities and broader contextual signals. The authors show consistent enhancements in performance over transfer-learning baselines through thorough assessment on two diverse OCT datasets (NEH and UCSD). Accuracy increases of up to 3.3% indicate enhanced generalization across scanners and acquisition techniques. This study demonstrates that carefully constructed multi-scale convolutional representations can accomplish comparable robustness with significantly lower complexity of structure and training cost, while later attention-based and transformer-driven models attempt to simulate long-range dependencies via global self-attention. Though effective, FPN-based CNNs might gain benefit from integration with attention or transformer modules due to the lack of explicit global context modeling and limited interpretability mechanisms. This could lead to the development of hybrid CNN–Transformer frameworks for clinically reliable AMD diagnosis.

Hamid et al. (2024) propose a hybrid deep learning framework that addresses the weaknesses of purely CNN-based classifiers that primarily depend on spatial feature learning alone by integrating MobileNetV1 as a convolutional feature extraction algorithm with a long short-term memory (LSTM) network for multi-class classification of age-related macular degeneration (AMD) using OCT images[24]. The paper explains that age-related macular degeneration (AMD) is clinically divided into dry (atrophic) and wet (neovascular) forms based on the existence of neovascularization, with drusen formation as a key hallmark of dry AMD. Drusen are yellowish extracellular deposits beneath the retina that accumulate due to AMD-

related retinal pigment epithelium dysfunction, and their number, size, and distribution influence disease severity and progression to advanced stages. The study uses a deep learning framework to classify OCT images into normal, dry (drusen-related), and wet AMD (with choroidal neovascularization), highlighting the importance of detecting drusen as a structural indicator for early/intermediate AMD. In order to differentiate between dry and wet AMD, previous AMD studies made extensive use of standalone CNN architectures like ResNet, VGG, DenseNet, and multi-scale CNN variants. However, these methods mainly disregarded the sequential dependency and contextual relationships embedded within complicated feature representations. The suggested CNN–LSTM architecture, on the other hand, makes use of MobileNet's lightweight depthwise separable convolutions to extract discriminative spatial features. The LSTM then models these features to capture long-range relationships across high-level feature vectors. This study achieves better performance on the Kermany OCT dataset by introducing temporal modeling at the feature level without improving convolutional depth, in contrast to multi-scale CNN and FPN-based techniques that improve receptive fields through architectural complexity. The model achieves 98.85% accuracy, 99.09% sensitivity, and 99.1% specificity for three-class AMD classification, greatly outperforming DenseNet201-, InceptionV3-, and CNN-only baselines. This work shows that CNN–LSTM hybrids may approach contextual reasoning with significantly reduced computing cost and data needs, despite transformer-based architectures' goal of capturing global context through self-attention processes. However, future hybrid CNN–Transformer models could boost robustness and interpretability due to their reliance on manually created image preprocessing and lack of explicit global attention mechanisms, making this work a crucial step in the transition between traditional CNN classifiers and newly developed attention-driven AMD diagnosis frameworks.

2.2 Research gaps

Early and intermediate AMD show drusen as the main sign. Deep learning models detect drusen well in many cases. Some models struggle with consistent detection because drusen vary in structure across stages. Traditional CNNs have small receptive fields. These small fields miss the fine differences in drusen patterns across retinal layers. We need models with larger context to catch these details.
Prediction of progression to wet AMD remains weak in intermediate dry eyes. Imaging and patient data are synthesized in hybrid models to generate risk scores. Strong instruments for customized risk assessment in day-to-day operations are still insufficient. Models still have high error rates in the event of the detection of intermediate dry AMD, and the stage is critical in attentive monitoring and prompt intervention. These are errors that we must practice on in order to improve effective diagnosis. Late AMD is characterized by macular neovascularization (MNV) and models tend to categorize AMD as binary (present or absent). They usually fail to highlight MNV subtypes (Type 1, Type 2, Type 3). The identification of subtypes is essential in treatment strategy. Current subtype models do not have the large multi-center tests and to have confidence on these classifications, we need more validation data. There is a lot of confusion between the active wet AMD and the inactive cases in the late stages of the disease. This error can result in late treatment. Since the early

lesion alterations in Type 3 MNV do not have any clear explanation, cooperation with doctors may help to analyze the errors and perfect the models. Longitudinal data are not available to trace long-term outcomes or repetition. In order to bridge this gap, we have to collect additional follow-up information.

AI is used in few research to connect fluid volume to vascular characteristics like vessel density or flow. These connections aid in evaluating therapy reaction and disease activity. When we quantify fluid and vessels together, we get better findings. Many high-end models rely on particular scanners. They fail when tested on different machines, populations, or ethnic groups. Public code and datasets stay rare for outside checks.

CNNs ignore sequence and global context in retinal images. Move to hybrid CNNTransformer or CNN-LSTM setups. These handle long-range links and full-image attention better. Most work uses one image type (OCT or fundus). Combine OCT with fundus photos and patient data. Multimodal models raise accuracy and link eye health to whole-body risks.

Direct comparison of microvascular features in MNV across diseases remains missing. We need to compare AMD with other conditions to sharpen management plans. We need to address these gaps to build AI tools we trust in clinics.

Chapter 3

Dataset Description and Analysis

3.1 Data Source and Clinical Context

The dataset used in this study is the OCTDL dataset (Optical Coherence Tomography Dataset to Image-Based Deep Learning Methods), which can be found in Mendeley Data and is characterized in the data descriptor that comes along with it [25]. OCTDL is a publicly accessible collection of retinal OCT B-scans images recorded under real clinical environments and categorized into various retinal disease groups.
OCT is a non-invasive imaging technique that generates cross-sectional images of the retinal layers. In the case of age-related macular degeneration (AMD), OCT is usually employed since disease-specific structural alterations may be seen as patterns ,including drusen-related RPE elevation, outer retinal bands disruption, atrophic alterations, and fluid-related abnormalities in the advanced stages of the disease. The inputs of the deep learning models in this thesis are OCT B-scans applied to screen AMD and classify stages.

3.2 Dataset Representation

OCTDL contains 1,618 OCT B-scan images with labels provided in a metadata CSV file. The dataset includes both normal scans and several retinal pathologies, which allows the experiments to be designed in a way that resembles screening and follow-up workflows.
This thesis focuses on AMD for two reasons. First, AMD is the largest category in OCTDL, which supports training and comparison of multiple deep learning backbones using the same experimental pipeline. Second, the AMD subset contains stage-related labels (via the "subcategory" field), which makes it possible to define a staging task in addition to a binary detection task.
According to these properties, there are two tasks:

- Task A (Detection): distinguish between AMD and Non-AMD, with the latter comprising Normal and all other diseases.
- Task B (Staging): categorize Early vs Intermediate vs Late, with the use of AMD images only.

	file_name	disease	subcategory	condition
1	amd_e_1.jpg	AMD	early	drusen
2	amd_f_15.jpg	AMD	late	MNV
3	amd_s_121.jpg	AMD	intermediate	MNV_suspected

Table 3.1: Example of OCTDL dataset

3.3 Dataset Organization and Image Acquisition

OCTDL is structured in a folder format with each type of disease being represented in a separate folder (e.g., AMD, NORMAL, and other conditions). The annotations in the CSV are associated with each image file and its label record.
All samples are 2D OCT B-scans. The images depict common clinical variability, such as contrast and brightness differences and OCT-specific speckle noise. This applies to the current work since the models should acquire characteristics that are stable under this variation.
In the experiments that are presented in this thesis, the number of images is 1,618. Out of them, 885 images are AMD images, 733 images are non-AMD images. Table 3.2 shows the per-class distribution.

Disease	Images
AMD	885
DME	143
ERM	133
NORMAL	284
RVO	93
VID	58
RAO	22
Total	1618

Table 3.2: OCTDL Dataset Distribution by Disease Category (After Filtering)

3.3.1 Image Characteristics

The retinal OCT dataset used in this study has the following properties:

- Imaging modality: Optical Coherence Tomography (OCT)
- Scan type: Retinal B-scan
- Color format: Grayscale (converted to three-channel RGB format for compatibility with pretrained convolutional neural networks)
- Spatial resolution: Variable across samples
- Field of view: Macular and parafoveal regions

The images exhibit common OCT-specific artifacts such as speckle noise, intensity inhomogeneity, and contrast variation, thereby representing realistic diagnostic conditions.

3.4 Annotation File and Metadata Description

The dataset contains an annotation file, OCTDL_labels.csv, which has a single record per image. The disciplines applicable in this thesis are:

- file_name: image filename,
- disease: main disease class,
- subcategory: severity/stage-related label (to derive AMD stage),
- condition: further characterization of prevailing observations.

The name of the stage of AMD is the label in the staging task, which is founded on the subcategory field. In the process of implementation, the string is normalized (removing non-white-spaces and changing it to lower case), which reduces the number of inconsistencies and then the string is transformed into three categories: early, intermediate, and late. Images whose stage values are missing or not matching are not included in the staging subset in order to have the final formulation that is single-label and consistent.

3.5 Dataset Utilization Strategy

To find proper classification on of AMD stages we use deep learning framework in two steps: This study focuses on the following objectives:

- Binary AMD Detection – distinguishing AMD from non-AMD retinal conditions
- Multi-Class AMD Stage Classification – categorizing AMD images into progressive disease stages

This plan is inspired by real world clinical evaluations where we first check if the disease is present before detecting the severity of it.

3.6 AMD Detection Dataset

For the first which is to detect if AMD is present the whole dataset is taken and if it is present it is assigned as a positive class and if not then it is assigned as negative class.

Class Index	Label Description
0	Non-AMD
1	AMD

Table 3.3: Binary class labels

This formulation evaluates the model's ability to discriminate AMD from other retinal pathologies based solely on OCT structural features.

3.7 AMD Staging Dataset Selection

After the distiction of AMD presence only those datas are used to classify the stage the AMD is in. With exsisting progression models labels are mapped and normalised.

3.8 AMD Stage Categorization

For the AMD staging task (Task B), images labeled as AMD were further categorized into three clinically meaningful stages: Early, Intermediate, and Late. Stage assignment is derived directly from the dataset metadata (the "subcategory" field in OCTDL_labels.csv) after normalization (string trimming and lowercasing) to reduce inconsistencies (e.g., casing differences). Only samples with valid and consistent stage labels were retained so that each OCT image corresponds to exactly one stage label (single-label classification).
The reason behind the three stage grouping is that it is based on progressive severity and it serves to facilitate a viable grading output in automatic decision support. Since intermediate AMD is a transitional phenomenon and has visual similarities to both early and late stages, inter-stage boundaries are not necessarily sharp visually in single B-scans; this difficulty will be overcome later with imbalance-aware training and evaluation metrics which focus on class-wise performance.

3.8.1 Early-Stage AMD

The cases of AMD at the initial stages of the disease in OCTDL are typified by comparatively mild structural alterations. Common patterns encompass small to medium size drusen-relevant detachments underneath the retinal pigment epithelium (RPE) and negligible disturbance of the retinal general layer structure. Early changes in most scans are localized and slight and this may lead to challenges in automated differentiation at the early stage of the scan.
Number of images (Early): 192

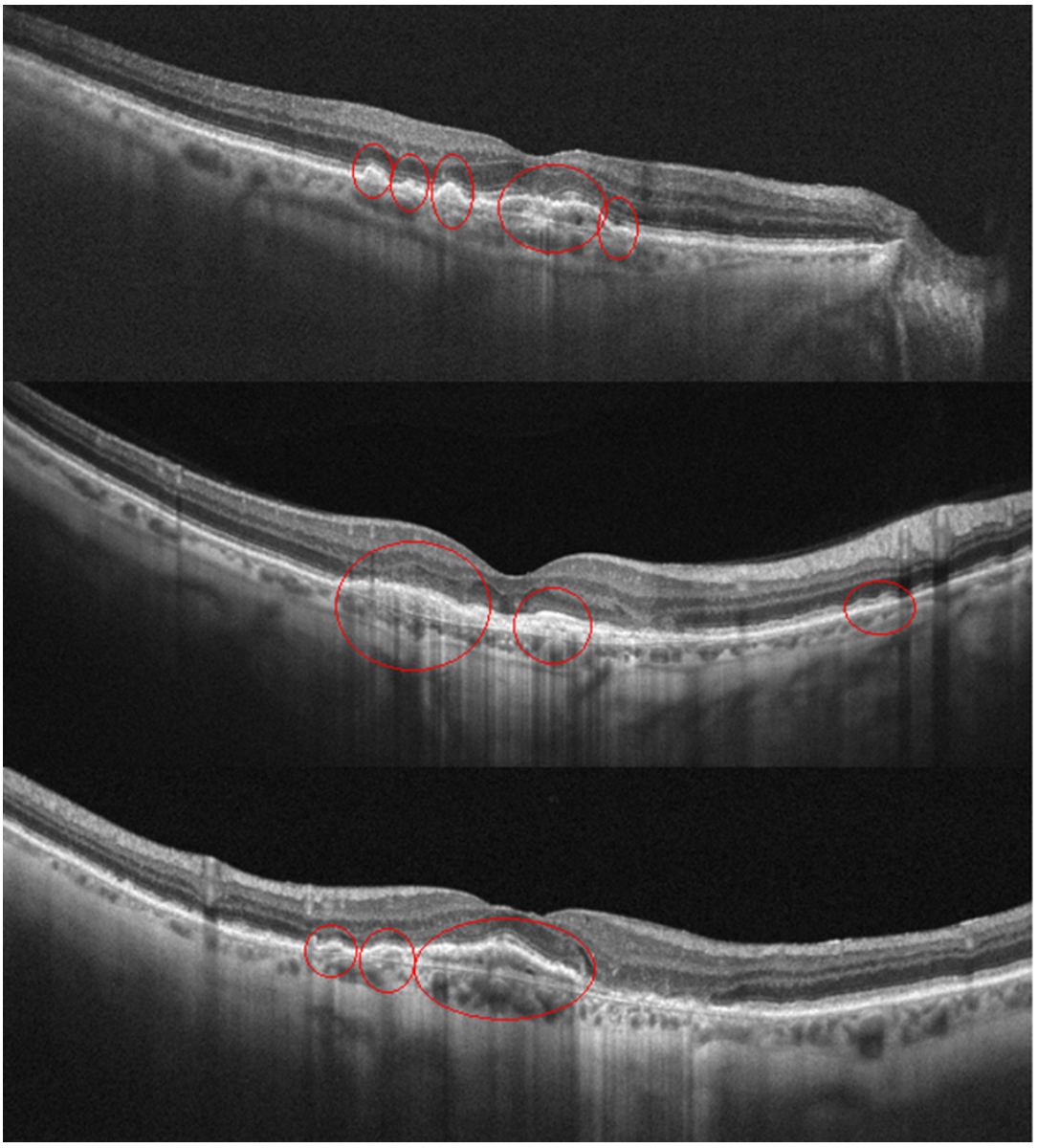

Figure 3.1: A representative Early AMD OCT image (Drusen marked)

3.8.2 Intermediate-Stage AMD

AMD in the intermediate stage is a terminal of earlier alterations and usually comprises larger and/or increased drusen and more apparent irregularities concerning the RPE-photoreceptor interface. Intermediate AMD, in comparison with early AMD, has higher structural variation although visually it overlaps with the adjacent stages particularly with only one B-scan.
Number of images (Intermediate): 175.

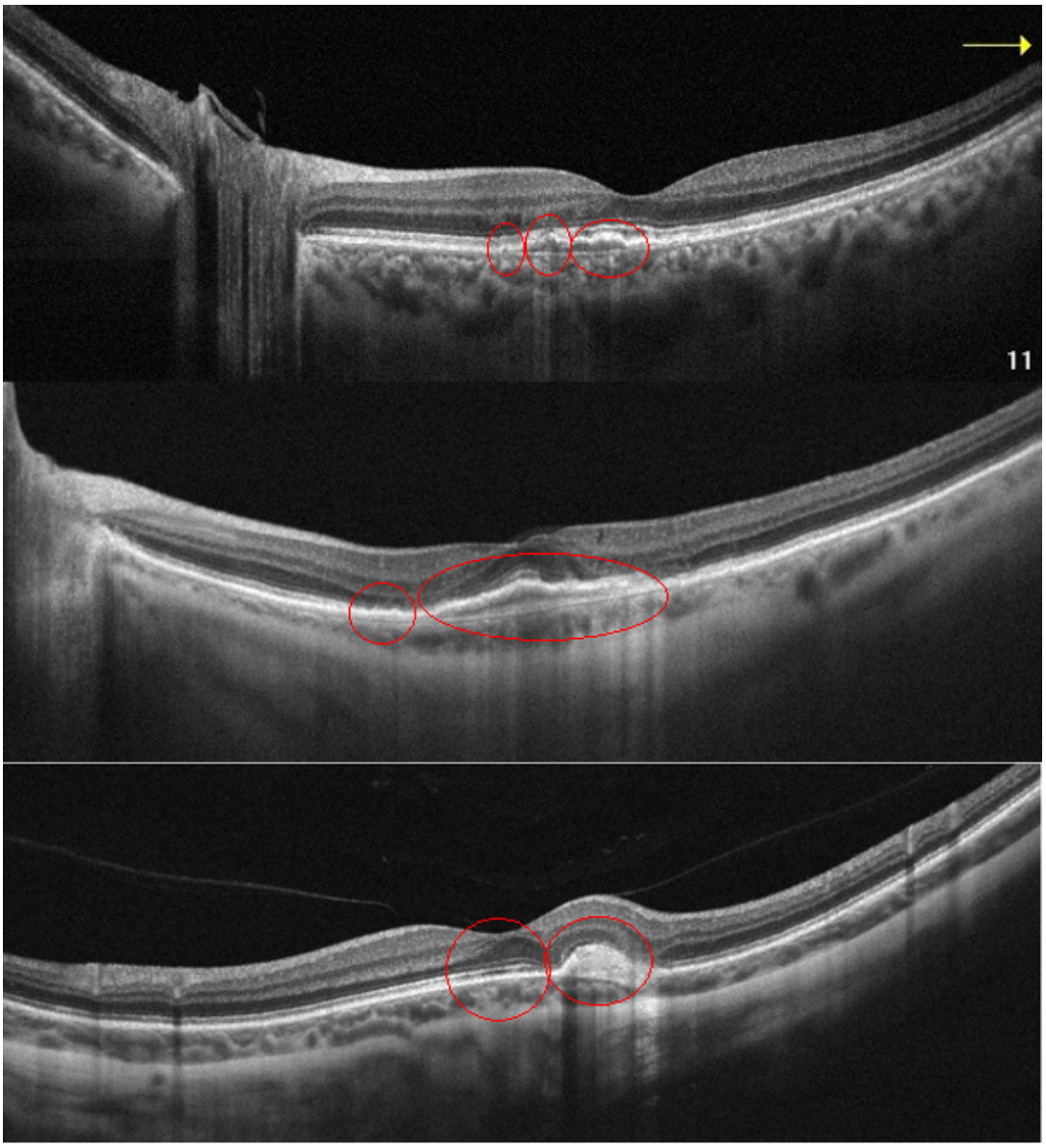

Figure 3.2: A representative Intermediate AMD OCT image (MNV suspect marked)

3.8.3 Late-Stage AMD

Advanced structural abnormalities and severe retinal degeneration are associated with late-stage AMD. In OCT, the late cases can show great impairment of the normal macular architecture, either atrophic patterns and/or neovascular related alteration, according to the case. Such scans are usually more intense and spatially extensive pathological signals than early or intermediate AMD.
Number of images (Late): 518

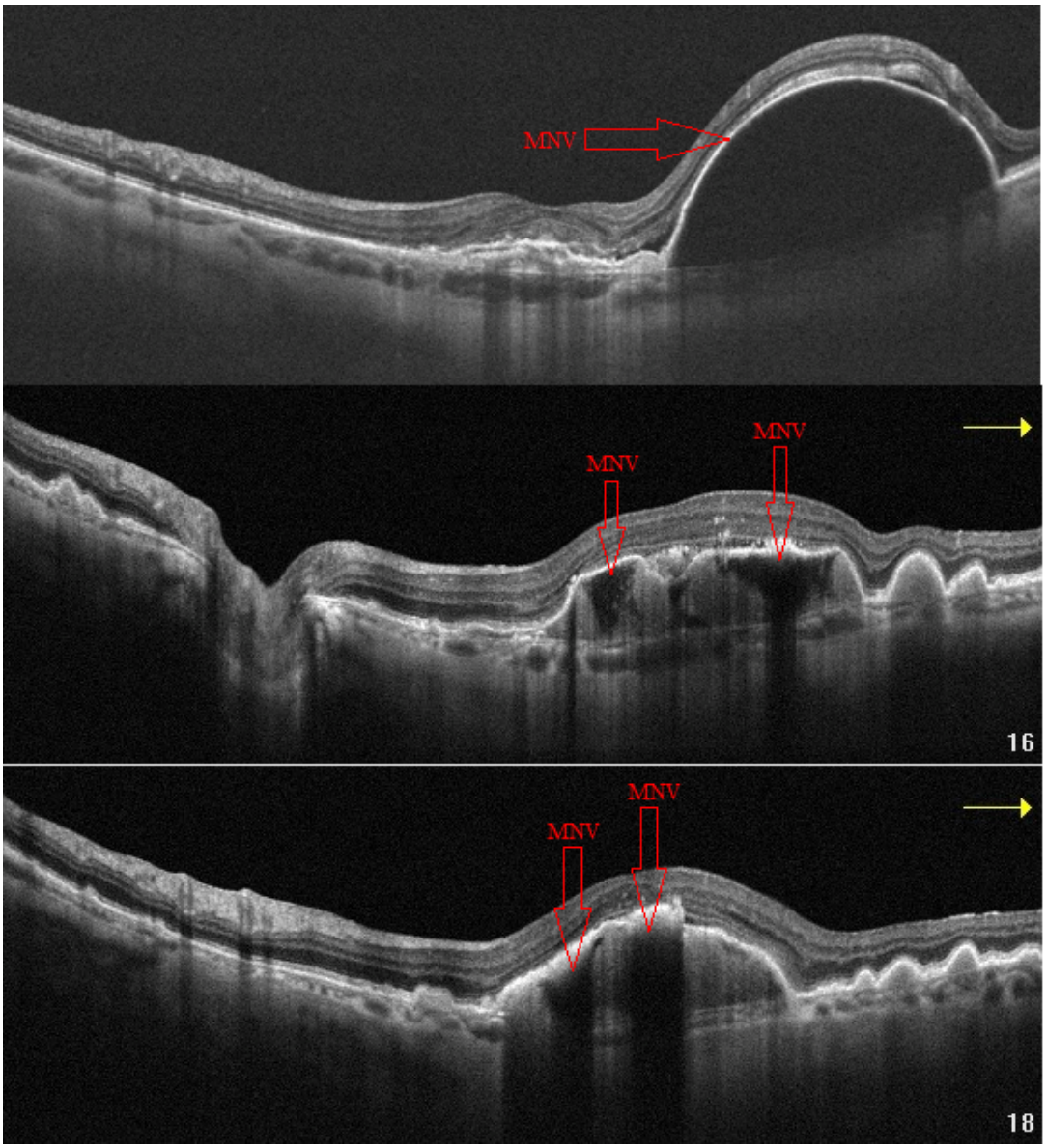

Figure 3.3: A representative Late AMD OCT image (MNV marked)

3.9 Target Classes for AMD Stage Classification

The AMD staging task is formulated as a single-label, multi-class classification problem with three mutually exclusive target classes. One label is assigned to each OCT B-scan in the staging subset:

Class Index	AMD Stage
0	Early
1	Intermediate
2	Late

Table 3.4: Multi class labels

This definition is used consistently across dataset splitting, model training, and

evaluation.

3.10 Dataset Distribution and Class Imbalance

After filtering to AMD-only samples with valid stage labels, the final staging subset contains 885 OCT images.

AMD Stage	Number of Images	Percentage (%)
Early	192	21.7
Intermediate	175	19.8
Late	518	58.5
Total	885	100

Table 3.5: AMD multi stage image distribution

This distribution is imbalanced, with Late-stage AMD representing the majority class. This imbalance is significant in the process of developing a model as a classifier may appear to have a high overall accuracy by giving weight to the majority class and under performing on the minority stages. Due to this reason, subsequent chapters will highlight imbalance-considerate training methods and report class-conditioned measures (e.g., per-class recall, macro-averaged F1, and balanced accuracy) in order to represent a clinically significant assessment.

3.11 Visual Characteristics Across AMD Stages

From an OCT imaging perspective, AMD stages exhibit progressively distinct structural patterns:

- Early AMD: the condition is characterized by localized mildly elevated drusen with minimum retinal layers distortion.
- Intermediate AMD: is more likely to have bigger/greater number of drusen, irregularity at the RPE border, and premature disorganization in the peripheral retina.
- Late AMD: commonly shows widespread architectural disruption, including patterns consistent with advanced degeneration (e.g., atrophic regions and/or severe morphological changes).

Because adjacent stages can share overlapping features–particularly Early vs Intermediate accurate staging requires robust feature extraction capable of capturing subtle differences in morphology and texture.

3.12 Dataset Challenges and Implications for Model Design

Although OCTDL provides clinically relevant real-world OCT scans, several dataset properties increase classification difficulty and directly influence model design choices:

1. Minor inter-class differences (in particular Early vs Intermediate): B-scans do not always separate into distinct stages and this makes it harder to distinguish between neighboring classes

2. OCT-specific noise and acquisition variability: Speckle noise, contrast differences, and imaging artifacts may obscure weak biomarkers, particularly in early-stage scans.

3. Another issue is the aspect of class imbalance, with AMD taking the lead in the staging subset. Unless correct mitigation measures are taken, this imbalance may bias the model training process on the majority class.

These issues encourage the application of the transfer learning to contemporary convolutional backbones and training approaches that can enhance the generalization and sensitivity to the minority classes.

3.13 Data Complexity and Clinical Variability

The OCT images in OCTDL reflect real clinical variability arising from multiple factors, including:

- speckle noise which is a characteristic of OCT acquisition,
- anatomical variation between patients,
- synonymous morphological patterns in AMD stages,
- scan quality, contrast and signal-strength variation.

Such factors may lead to borderline cases when the stage is not highly emphasized in one B-scan. Consequently, staging task is a more difficult task compared to binary detection task and should be evaluated through class-wise performance and not by general accuracy.

3.14 Ethical and Licensing Considerations

OCTDL is publicly distributed through Mendeley Data and accompanied by a reuse license specified on the dataset page. In this thesis, the dataset is used strictly for academic research purposes. The dataset source is cited appropriately, and all usage follows the terms stated in the dataset license, including attribution requirements. The OCTDL dataset contained 733 Non-AMD disease and Normal real-world retinal OCT B-scans.

Disease	No. of images
Diabetic Macular Edema	143
Epiretinal Membrane	133
Retinal Vein Occlusion	93
Normal	284
Vitreomacular Interface Disease	58
Retinal Artery Occlusion	22

Table 3.6: Non-AMD images unused

3.15 Summary

In summary, the OCTDL dataset provides a clinically meaningful and methodologically sound foundation for automated AMD detection and staging. Its real-world OCT imaging characteristics, structured annotations, and progressive disease representation make it an appropriate benchmark for evaluating efficient deep learning architectures aimed at improving early diagnosis and disease management in age-related macular degeneration.

Chapter 4

Model Description

4.1 Methodology and Implementation

This chapter presents the complete methodology and implementation details of the proposed deep learning framework for (i) binary detection of Age-Related Macular Degeneration (AMD) and (ii) multi-class staging of AMD (Early, Intermediate, Late) from retinal Optical Coherence Tomography (OCT) B-scan images.
The experimental design follows a clinically meaningful workflow consisting of two sequential components:

- Binary Screening/Detection: Find out if there is AMD from the OCT image .
- Conditional Staging: If AMD is detected then we find out what stage the AMD is in..

This two step system is inspired by real world disease detection by doctores where they first find out if the patient have disease and then assess the severity of the said disease.
Framing the system as two complementary modules supports practical deployment as a decision-support pipeline for screening, referral prioritization, and longitudinal monitoring.

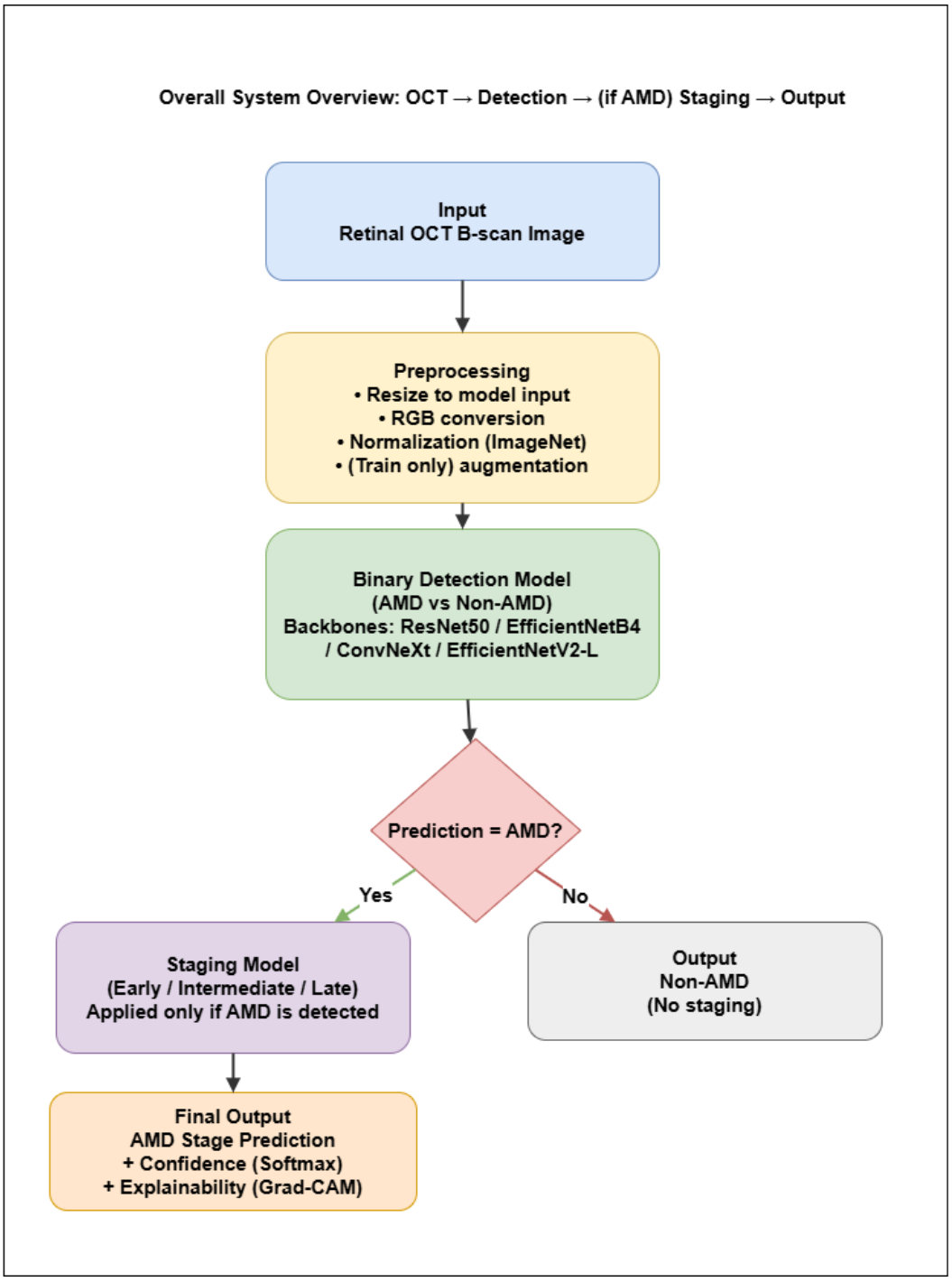

Figure 4.1: Overall methodology overview

4.2 AMD and non-AMD

4.2.1 AMD

The models were trained on the **OCTDL dataset** from Mendeley Data, containing **1,618 real-world retinal OCT B-scans**. The staging subset focused specifically on 885 AMD images:

- **Early (192 images):** Characterized by small to medium drusen.

- **Intermediate (175 images):** Increased drusen size and retinal layer irregularity.
- **Late (518 images):** Severe degeneration, often featuring geographic atrophy or neovascularization.

4.2.2 non-AMD

The OCTDL dataset contained 733 non-AMD disease real-world retinal OCT B-scans.

Disease	no. of images
Diabetic Macular Edema	143
Epiretinal Membrane	133
Retinal Artery Occlusion	93
Vitreomacular Interface Disease	58
Normal	284

Table 4.1: Non-AMD images unused

4.3 Task Definition and Label Engineering

4.3.1 Task A — Binary AMD Detection (AMD vs Non-AMD)

To reflect a clinically relevant screening scenario, the first task was formulated as a binary classification problem that distinguishes AMD from all other retinal conditions present in the dataset. All images labeled with disease = AMD were assigned to the positive class, whereas images belonging to any other disease category were grouped into a single negative class. This design evaluates whether a model can detect AMD in the presence of heterogeneous non-AMD pathologies (i.e., a more challenging and realistic setting than "AMD vs Normal" only).
Formally, a binary target label y_det was defined for each image as:

- y_det = 1 (AMD) if the standardized disease folder equals "AMD"
- y_det = 0 (Non-AMD) otherwise

This is implemented as:

- df_det["y_det"] = (df_det["disease_folder"] == "AMD").astype(int)

Accordingly, the class names are:

- Class 0: Non-AMD
- Class 1: AMD

4.3.2 Task B — AMD Staging (Early / Intermediate / Late)

The second task focuses on disease progression and is formulated as a three-class classification problem restricted to AMD cases only. First, all AMD-labeled images were selected:

- df_stage = df[df["disease_folder"] == "AMD"].copy()

Next, the stage label was derived from the subcategory field. To avoid inconsistencies caused by casing or trailing spaces, the subcategory string was normalized:

- subcategory_norm = lower(strip(subcategory))

A stage mapping was then defined:

- Early → 0
- Intermediate → 1
- Late → 2

Implemented as:

- stage_map = {"early": 0, "intermediate": 1, "late": 2}
- df_stage = df_stage[df_stage["subcategory_norm"].isin(stage_map)]
- df_stage["y_stage"] = df_stage["subcategory_norm"].map(stage_map).astype(int)

After filtering to valid AMD stages, the final staging subset contained 885 AMD images, distributed as:

- Early: 192
- Intermediate: 175
- Late: 518

4.3.3 Clinical Pipeline Interpretation (Detection → Staging)

Although the detection and staging models are trained and evaluated as separate experiments, they are intended to operate sequentially in a clinical decision-support pipeline. In deployment, it's first seen if AMD is present in the patient and then what stage the AMD is in is dtetcted. It reflicts the workflow of clinical setting where first disease presence is confirmed before getting to evaluate the gravity of the disease so that the next step to how to cure is decided.

4.4 Dataset Splitting Strategy (Train / Validation / Test)

Stratified sampling is utilized for partitioning the tasked dataset into training, test and validation as to make sure that unseen samples have good generalization and proportions are preserved across splits. For both detection and staging experiments, the dataset was split into:

- Training set: 70%
- Validation set: 15%
- Test set: 15%

The splitting was implemented in two sequential steps using train_test_split with stratify based on the task label:

1. Training vs. temporary split:

   70% training and 30% temporary, stratified by label train_test_split(df, test_size=0.30, stratify=labels, random_state=SEED)

2. Validation vs. test split:

   The 30% temporary set was split equally into validation and test (15% each overall), stratified by label train_test_split(temp, test_size=0.50, stratify=temp_labels, random_state=SEED)

This procedure was applied independently for:

- the detection dataset (df_det, label y_det), and
- the staging dataset (df_stage, label y_stage).

4.5 Preprocessing Pipeline (Normalization and Input Preparation)

4.5.1 Grayscale-to-RGB Conversion for Transfer Learning

OCT images are typically acquired as single-channel grayscale scans. However, the CNN backbones used in this study are initialized with ImageNet-pretrained weights and therefore expect three-channel RGB input. To maintain compatibility with these pretrained models without altering the first convolutional layer, each OCT image is converted to RGB at load time using:
Image.open(path).convert("RGB")
This operation replicates the original grayscale intensity across three channels. Although it does not introduce new color information, it allows direct reuse of pretrained weights and is a common practice in medical imaging transfer learning.

4.5.2 Spatial Standardization and Resizing

CNN backbones require a fixed input tensor size. Therefore, all images are resized to img_size × img_size, where img_size is chosen per backbone to balance expected accuracy and GPU memory constraints. The implementation uses the following input sizes:
Detection task:

- ResNet50: 224 × 224
- EfficientNetB4: 380 × 380
- ConvNeXt Tiny: 224 × 224
- ConvNeXt Small: 224 × 224
- EfficientNetV2-L: 320 × 320

Staging task:

- ResNet50: 224 × 224
- EfficientNetB4: 380 × 380
- ConvNeXt Tiny: 224 × 224
- ConvNeXt Small: 224 × 224
- EfficientNetV2-L: 384 × 384

These resolutions reflect common recommended input sizes for each architecture family and were selected to maintain stable training under available GPU resources.

4.5.3 Intensity Normalization (ImageNet Statistics)

After conversion to tensor, images are normalized using ImageNet channel statistics:

- Mean: [0.485, 0.456, 0.406]
- Standard deviation: [0.229, 0.224, 0.225]

Implemented as:
transforms.Normalize(mean, std)
This normalization aligns the OCT input distribution with the scale expected by ImageNet-pretrained filters, improving optimization stability and transfer learning effectiveness during fine-tuning.

4.6 Data Augmentation (Training-Time Robustness Techniques)

4.6.1 Rationale for Augmentation in OCT

Retinal OCT images acquired in clinical practice exhibit substantial variability arising from differences in scanner manufacturers, acquisition protocols, operator settings, and patient-specific factors (e.g., fixation stability and ocular media clarity). In addition, OCT inherently contains speckle noise and may include motion artifacts or intensity inhomogeneity. Such variability can reduce model generalization if the training distribution is narrow. Therefore, data augmentation was applied to simulate realistic acquisition variations, increase effective training diversity, and reduce overfitting particularly important given the limited size of the AMD staging subset.

4.6.2 Training-Time Augmentation Pipeline

Augmentation was applied only to the training split. Validation and test sets were processed deterministically to ensure unbiased evaluation. The training transform pipeline implemented in this work consists of the following operations (in order):

1. Resize to target input resolution

   All images are resized to the model-specific input size (e.g., 224×224 for ResNet50/ConvNeXt; 380×380 for EfficientNetB4), ensuring consistent tensor dimensions.

2. RandomResizedCrop (scale = 0.80–1.00)

   A random crop is sampled and resized back to the target resolution. This introduces mild translation and scale variability, which improves robustness to differences in scan framing and retinal centering.

3. RandomHorizontalFlip (p = 0.5)

   Horizontal flipping simulates left/right orientation variability (e.g., acquisition from different eyes or mirrored scan orientation). This encourages the model to focus on disease morphology rather than absolute lateral position.

4. RandomRotation (±10°)

   Small rotations simulate mild scan tilt and acquisition misalignment. Limiting the rotation range preserves anatomical plausibility while improving robustness.

5. ColorJitter (brightness ±0.08, contrast ±0.08)

   Although OCT is fundamentally grayscale, intensity and contrast distributions vary across devices and acquisition settings. Light brightness/contrast perturbations help the model generalize to such variations.

6. ToTensor()

   Converts the PIL image into a normalized float tensor in the range [0, 1] prior to noise injection and normalization.

7. Gaussian noise injection (custom augmentation)

   Adds controlled noise to mimic speckle-like artifacts present in OCT.

8. ImageNet normalization

   Finally, inputs are normalized using ImageNet mean and standard deviation to match the distribution expected by ImageNet-pretrained backbones.

4.6.3 Gaussian Noise Injection(OCT-Specific Regularization)

To increase robustness to OCT noise characteristics, a custom augmentation module (AddGaussianNoise) was applied after tensor conversion. With probability p = 0.20, Gaussian noise sampled from $N(0, std^2)$ with std = 0.015 is added to the image tensor. Values are clipped to the valid intensity range [0,1] to prevent saturation artifacts. This strategy encourages the model to learn stable structural representations (e.g., layer boundaries, drusen-related elevations) rather than overfitting to scanner-specific noise patterns.

4.6.4 Evaluation Transforms (Validation/Test)

For validation and testing a predictive preprocessing pipeline was used to maintain the inetgrity. Every singles image was resized to a certain input resolution, converted to a tensor and then normalized using the ImageNet statistics. No other transformation is used so that true generalization to unknown image is observed.

4.7 Handling Class Imbalance

4.7.1 Imbalance in AMD Staging

The AMD staging subset is imbalanced, with late-stage AMD (N = 518) representing the majority of samples, while early (N = 192) and intermediate (N = 175) are underrepresented. Without mitigation, this imbalance can lead to models that achieve high overall accuracy by over-predicting the majority class, while failing to achieve clinically important sensitivity for early-stage AMD.

4.7.2 Weighted Sampling (Batch Balancing)

To increase exposure to minority classes during training, the implementation includes an optional WeightedRandomSampler. Class weights are computed inversely proportional to the class frequencies in the training split:

$$w\_c = \frac{1}{count\_c} \tag{4.1}$$

Each training sample is assigned the weight of its class, and batches are drawn with replacement. This sampling strategy approximates class-balanced mini-batches and reduces training bias toward the majority class.

4.7.3 Weighted Cross-Entropy (Loss Balancing)

In addition (or alternatively), the staging task applies class-weighted cross-entropy loss, where the loss contribution of minority classes is increased through a weight vector derived from training class counts. In the implementation, weights are computed using an inverse-frequency formulation and normalized to have mean 1, then passed to:

$$criterion = nn.CrossEntropyLoss(weight = class\_weights\_tensor) \tag{4.2}$$

Important reporting note: Using both weighted sampling and weighted loss simultaneously can over-compensate for imbalance. For publication clarity and probability calibration, the final reported configuration should explicitly state whether imbalance was handled by (i) sampling, (ii) weighted loss, or (iii) both, and should ideally include an ablation study justifying the choice.

4.8 MixUp and CutMix (Sample-Mixing Regularization for Staging)

4.8.1 Motivation

For AMD staging, the dataset is relatively small and exhibits subtle inter-class differences (particularly between early and intermediate stages). To reduce overfitting and improve robustness, the training pipeline incorporates sample-mixing regularization using MixUp and CutMix. These techniques encourage smoother decision boundaries, reduce sensitivity to individual samples, and can improve generalization in imbalanced or limited-data settings. In this work, MixUp/CutMix were applied only during training and primarily for the multi-class staging task.

4.8.2 MixUp Implementation: Linear Interpolation in Input Space

MixUp generates a synthetic training example by linearly combining two samples from the same batch:

- $\lambda \sim \beta(\alpha, \alpha)$
- $x\_\text{mix} = \lambda x + (1 - \lambda)x'$

The targets are mixed accordingly through a convex combination of losses:

- $L = \lambda \cdot \text{CE}(f(x\_\text{mix}, y) + (1 - \lambda) \cdot \text{CE}(f(x\_\text{mix}), y')$

In the implementation, $\alpha = 0.4$, controlling the strength of interpolation.

4.8.3 CutMix Implementation: Regional Replacement in Input Space

CutMix replaces a randomly sampled rectangular region of an image with a patch from another image. The mixing ratio λ is sampled from $\lambda \sim \text{Beta}(\alpha, \alpha)$, the patch

area is derived from $\sqrt{(1-\lambda)}$, and λ is then corrected based on the actual replaced area. Targets are mixed using the same convex combination principle as MixUp. In this work, CutMix uses $\alpha = 0.4$.

4.8.4 Combined Strategy (MixUp+CutMix)

The staging task uses a combined augmentation mode in which, for each mini-batch:

- CutMix is applied with probability p_cutmix = 0.5
- otherwise MixUp is applied

4.9 Model Backbones and Architecture Descriptions

4.9.1 Shared Transfer Learning Procedure (Common to All Backbones)

All experiments were conducted under a unified transfer learning strategy to enable a fair comparison among architectures. Each backbone network was initialized with ImageNet pretrained weights (via torchvision.models), leveraging the strong generic visual representations learned from large-scale natural image data. Although OCT scans differ substantially from ImageNet images, transfer learning remains effective because early and mid-level CNN features (e.g., edge, texture, and shape detectors) provide useful initialization and improve convergence on small medical datasets.
For each backbone, the original ImageNet classification head (1000 output classes) was replaced with a task-specific classifier head matching the required number of outputs:

- Binary AMD detection: 2 output neurons (Non-AMD, AMD)
- AMD staging: 3 output neurons (Early, Intermediate, Late)

To reduce overfitting, a dropout layer was inserted immediately before the final linear layer:
Dropout probability: p = 0.4
Concretely, the adapted head follows the form:
Dropout(0.4) → Linear(in_features, num_classes)
This design preserves the pretrained feature extractor while enabling task-specific discrimination. The same adaptation pattern was applied across models, ensuring that performance differences primarily reflect backbone feature-learning capacity rather than differences in classifier design.

4.9.2 ResNet50 — Architecture Description

ResNet50 (Residual Network with 50 layers) is a widely adopted convolutional neural network backbone that addresses the degradation problem encountered in very deep neural networks through the introduction of residual (skip) connections. Rather

than learning a direct mapping, each residual block learns a residual function that is added to the block input. The formula considers aspects like the gradient flow is stable and the depth of network increased for optimization. This formulation facilitates stable gradient flow and improves optimization as network depth increases. Even if This kind of setup works well when applying transfer learning, especially since adjusting complex models with small medical image sets often causes shaky training results. so it makes sense to move forward like this.

Key architectural characteristics.
Each residual block produces an output defined as $y = F(x) + x$
Here, x stands for the input feature map, while F(x) shows the change made by the block. That is why it feels just right
Bottleneck residual blocks:
A bottleneck design helps ResNet50 use fewer parameters without losing expressive power. Inside each block, one finds stacked convolutional layers arranged in a specific sequence. Each bottleneck block consists of:

- a 1×11 times 11×1 convolution for channel reduction,
- a 3×33 times 33×3 convolution for spatial feature extraction, and
- a 1×11 times 11×1 convolution for channel expansion.

Deep levels of feature detection happen here, though processing demands stay manageable. Still, the structure supports complex patterns without slowing down too much.

Relevance to OCT-based AMD analysis.
ResNet50 serves as a strong baseline for transfer learning in medical imaging applications for several reasons. The presence of residual connections supports stable fine-tuning on limited datasets, while the depth of the network enables hierarchical feature representations capable of capturing subtle morphological patterns in OCT images. These include drusen-related retinal pigment epithelium elevation, localized hyperreflective deposits, and gradual disruption of the outer retinal layers. In addition, the architecture is well studied in the literature, providing a robust and interpretable point of comparison against newer and more computationally efficient backbones such as EfficientNet and ConvNeXt.

Pretrained initialization:
The ResNet50 backbone was initialized using ImageNet-pretrained weights (`ResNet50_Weights.IMAGENET1K_V2`).

Classifier adaptation:
The original fully connected classification layer was replaced with a task-specific head composed of a Dropout layer with a rate of 0.4 followed by a Linear layer with an output dimension equal to the number of target classes. This modification preserves the pretrained convolutional feature extractor (from conv1 through layer4) while adapting the final decision layer to support either binary disease detection or three-class disease staging.

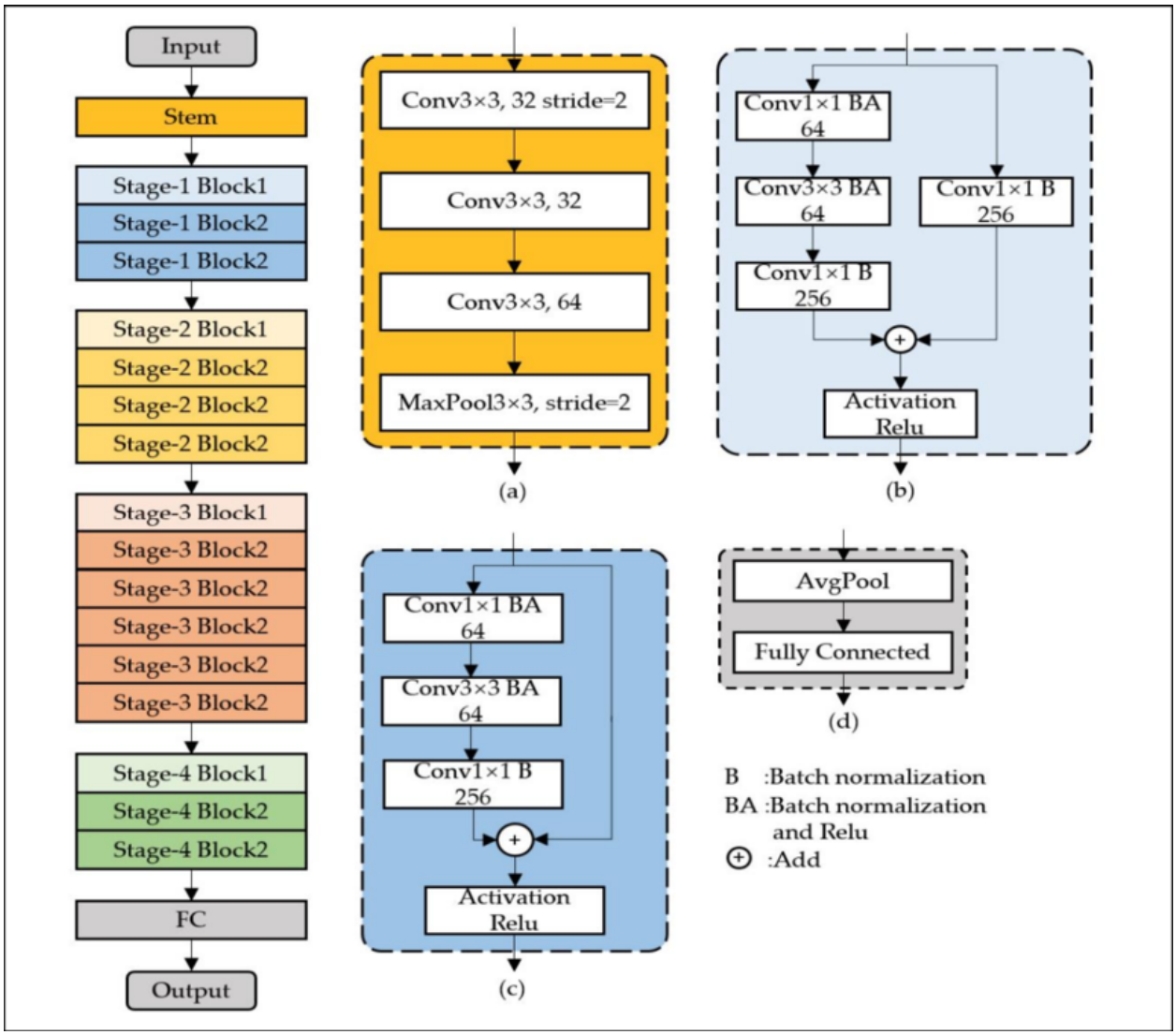

Figure 4.2: ResNet50 Model Architecture and Layers

4.9.3 EfficientNetB4 — Architecture Description

EfficientNet is a family of convolutional neural networks designed to achieve high classification accuracy while maintaining strong computational efficiency. Unlike earlier scaling strategies that increase only a single network dimension (such as depth or width), EfficientNet introduces a compound scaling approach that jointly scales network depth, width, and input resolution using a principled scaling coefficient. This coordinated scaling strategy enables more efficient performance gains than unbalanced scaling methods.
EfficientNetB4 Takes more space but packs stronger performance across its design lines. Running on bigger image sizes sets it apart from older models like ResNet50 right away. Optimization leans hard into doing more math without wasting steps along the way. High precision goals fit neatly when hardware limits start closing in. Built smart so heavy lifting happens quietly behind the scenes. Speed and sharpness link up under tight budgets by default.

Core architectural components.

Starting off, EfficientNet models rely heavily on stacked MBConv blocks, these pieces save parameters while staying effective across newer compact convolution networks. A key trait is that they're built to do more with less hardware fuss, turning up often where speed and size matter most.

The key architectural components include:
MBConv (inverted residual) blocks: Inside every MBConv block, channels grow first. Then comes a special kind of filtering that works slice by slice across space. After that, things shrink again into a tighter form. All this cuts down on math-heavy steps without losing detail sharpness

Squeeze-and-Excitation (SE) attention: Squeeze-and-excitation blocks adjust channel importance using overall context. Instead of treating all features equally, they boost useful signals while quieting weaker ones. Diagnostically meaningful patterns gain strength this way. Global awareness guides these shifts - no guesswork involved .

Nonlinear activation: Sometimes EfficientNet uses Swish or SiLU type activations, based on how it's built. When swapped in, these tend to help training flow better than regular ReLU does under some conditions. Smooth patterns in data get captured more naturally, which changes how gradients move through layers. Not every setup sees gains, yet many notice subtle improvements in stability during learning phases.

Relevance to OCT-Based AMD Detection and Staging
EfficientNetB4 is particularly well suited to OCT-based AMD analysis for several reasons. Its high representational efficiency enables the modeling of subtle morphological variations across different disease stages, including small drusen elevations and early irregularities in the outer retinal layers. The combination of squeeze-and-excitation attention and multi-scale feature extraction further supports sensitivity to fine retinal structures. In addition, EfficientNetB4 offers a favorable accuracy–efficiency trade-off, making it a strong candidate for potential deployment in resource-limited clinical environments compared to heavier architectures.

Implementation Details in This Study
Pretrained initialization:
The EfficientNetB4 backbone was initialized using ImageNet-pretrained weights (`EfficientNet_B4_Weights.IMAGENET1K_V1`).
Classifier adaptation:
The original EfficientNet classifier module was modified to match the target tasks. The dropout probability was set to 0.4, and the final linear layer was replaced with a task-specific output layer. The adapted classifier supports two output classes for binary disease detection (Non-AMD vs. AMD) and three output classes for disease staging (Early vs. Intermediate vs. Late AMD). This modification preserves the pretrained EfficientNet feature extractor while adapting the final classification head to OCT-based AMD detection and staging tasks.

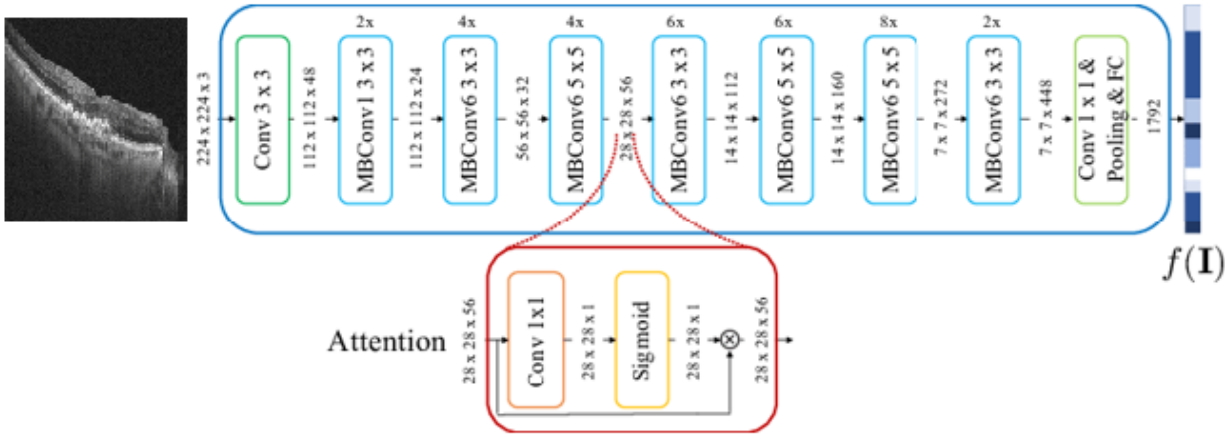

Figure 4.3: EfficientNetB4 Model Architecture and Layers

4.9.4 ConvNeXt Tiny and ConvNeXt Small — Architecture Description and Rationale

ConvNeXt is a modern convolutional network family designed to achieve Transformer-level performance while remaining entirely convolutional. Instead of using self-attention, ConvNeXt improves classical CNN design by adopting architectural and training principles that became common with Vision Transformers, resulting in stronger feature extraction and improved optimization behavior for transfer learning.

Key architectural characteristics.

ConvNeXt incorporates several deliberate CNN design updates:

- Large-kernel depthwise convolutions: ConvNeXt increases the effective receptive field using large-kernel depthwise convolution, enabling the model to capture broader contextual information. This is beneficial for OCT analysis because clinically relevant retinal changes (e.g., extended RPE elevation, diffuse reflectivity shifts, and layer continuity disruptions) often span a substantial portion of the scan.
- Simplified stage-based hierarchy: The architecture uses a clean stage design with progressive downsampling across multiple stages, analogous to ResNet-style feature pyramids. This supports multi-scale representation learning, which is important for detecting both localized biomarkers (e.g., small drusen) and more global changes (e.g., widespread atrophy).
- Updated normalization and layer ordering: ConvNeXt adopts revised normalization and block arrangements (inspired by Transformer-era "pre-norm" conventions), which generally improves training stability and fine-tuning behavior particularly important when adapting ImageNet-pretrained models to medical domains.
- GELU activations and modern training recipe assumptions: GELU nonlinearities and modernized training assumptions contribute to improved representational smoothness and strong performance on fine-grained visual discrimination problems.

Use of ConvNeXt Tiny and ConvNeXt Small in this study.
Both ConvNeXt Tiny and ConvNeXt Small were included as separate backbone candidates within the same experimental framework. Using two capacity levels from the ConvNeXt family enables evaluation of how increasing backbone capacity affects (i) binary AMD detection performance and (ii) fine-grained staging sensitivity, particularly for minority or borderline classes. Rather than positioning one as a competitor against the other, they are treated as complementary variants representing different points on the accuracy–compute trade-off curve.
Relevance to OCT-based AMD detection and staging.
OCT-based AMD recognition requires sensitivity to subtle textural and structural patterns (e.g., small drusen morphology, early irregularities of retinal layers, and progressive disruptions). ConvNeXt's receptive field design and stable optimization characteristics make it suitable for such fine-grained discrimination while avoiding the overhead and interpretability complexity often associated with attention-based models.
Implementation details in this study.

- Pretrained initialization: IMAGENET1K_V1 weights were used for both ConvNeXt variants via torchvision.models.
- Classifier adaptation: the final classifier head was replaced with a task-specific head of the form:

  Dropout(0.4) → Linear(in_features, num_classes)

  where num_classes = 2 for binary detection and num_classes = 3 for AMD staging.

This consistent head replacement ensures that differences in performance across backbones can be attributed primarily to feature extraction capacity and architectural design rather than differences in the classification head.

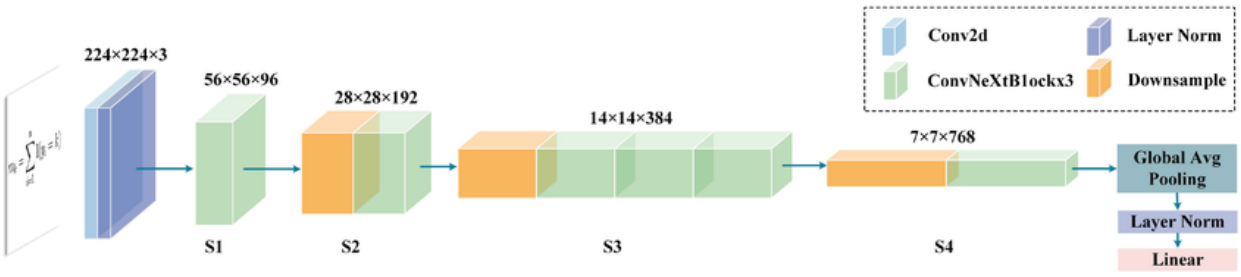

Figure 4.4: ConvNeXt Tiny and ConvNeXt Small Model Architecture and Layers

4.9.5 EfficientNetV2-L — Architecture Description and Rationale

What sets EfficientNetV2 apart begins with its role as an upgraded version of earlier EfficientNet models, built to train faster and scale smarter without losing precision - especially as models grow bigger. Instead of just pushing more power into the system, it tweaks how parts connect inside, smoothing out hiccups during learning phases. When stacked against older versions like EfficientNetB4, you notice less

waiting around thanks to clever changes under the hood. These upgrades help avoid shaky progress when increasing size, letting performance climb on steadier ground.
Key architectural characteristics. EfficientNetV2 improves upon the original EfficientNet design primarily through the use of:

- Fused-MBConv blocks (early stages):

  EfficientNetV2 replaces some MBConv blocks with fused-MBConv, which combines expansion and depthwise convolution into a single regular convolution operation. This reduces memory access overhead and accelerates training on modern hardware while preserving representational capacity.

- Improved scaling behavior:

  EfficientNetV2 uses updated scaling rules and training recipes to better balance accuracy and computational cost across model sizes.

- High-capacity variants for complex feature learning:

  Larger EfficientNetV2 models (such as V2-L) provide increased depth/width and can capture more complex visual patterns, which may be beneficial for fine-grained medical classification tasks.

Relevance to OCT-based AMD detection and staging. EfficientNetV2-L was included to evaluate whether a higher-capacity, modern EfficientNet variant provides measurable benefits for:

- subtle stage discrimination (e.g., early vs intermediate AMD),
- complex morphological patterns such as drusen aggregation, irregular RPE elevation, and outer retinal disruption,
- improved probability separation (AUC) in multi-class settings.

At the same time, the increased parameter count and compute requirements allow assessment of the performance–efficiency trade-off relative to more lightweight backbones.
Implementation details in this study.

- Pretrained initialization: EfficientNet_V2_L_Weights.IMAGENET1K_V1
- Classifier adaptation: the original classifier head is modified as:
  - Dropout(p = 0.4)
  - Linear(in_features, num_classes) where num_classes = 2 for detection and num_classes = 3 for staging.
- Mixed precision note: In the implementation, AMP (automatic mixed precision) was disabled for EfficientNetV2-L due to stability considerations observed with this large-capacity model under FP16 training on the target hardware configuration. Training was therefore performed using standard FP32 computation for this backbone.

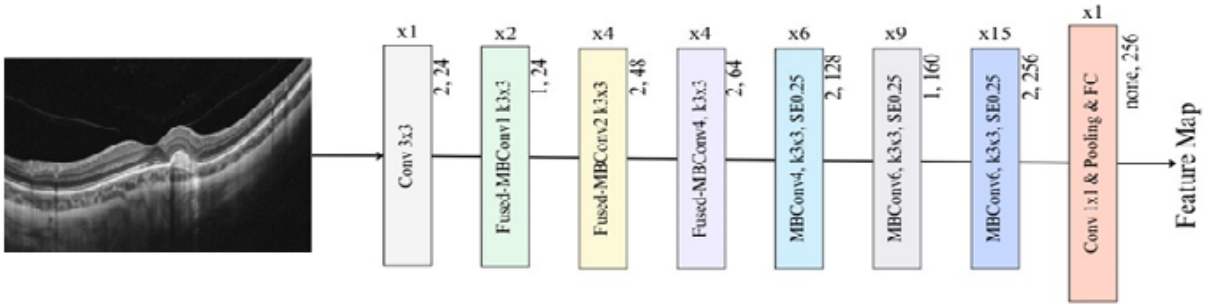

Figure 4.5: EfficientNetV2-L Model Architecture and Layers

4.10 Training Strategy and Hyperparameter Configuration

4.10.1 Unified Training Routine (run_task)

To ensure methodological consistency and a fair comparison across backbone networks, all experiments were executed using a single standardized training routine implemented in the function run_task(...). This routine encapsulates the complete experimental workflow, including: (i) selection of task-specific preprocessing pipelines (training vs. evaluation transforms), (ii) stratified splitting into train/validation/test subsets, (iii) construction of PyTorch datasets and data loaders, (iv) model instantiation and classifier-head replacement for the target task, (v) two-phase transfer learning with controlled freezing/unfreezing of parameters, (vi) optimizer, learning-rate scheduling, and mixed-precision configuration, (vii) early stopping and best-checkpoint selection based on validation loss, and (viii) final evaluation and visualization (loss curves and confusion matrices). By enforcing a shared training and evaluation pipeline, differences in performance between backbones can be attributed primarily to architectural differences rather than inconsistent experimental settings.

4.10.2 Two-Phase Transfer Learning: Freeze → Fine-Tune

Training followed a two-phase fine-tuning strategy to stabilize optimization when adapting ImageNet-pretrained networks to OCT images.

- Phase 1: Classifier-head training (frozen backbone).

At initialization, all model parameters are frozen (requires_grad = False). Only the newly defined classifier head is unfrozen and trained: for ResNet50, this corresponds to model.fc, whereas for EfficientNet and ConvNeXt variants it corresponds to the model.classifier module. This phase adapts the final decision layer to the OCT domain while preserving generic pretrained features, reducing the risk of catastrophic forgetting.

- Phase 2: Full fine-tuning (unfreeze all layers).

After freeze_epochs + 1, all layers are unfrozen (requires_grad = True) and the entire network is fine-tuned using a smaller learning rate. This phase enables deeper feature adaptation to OCT-specific biomarkers associated with AMD detection and staging, while maintaining training stability.

4.10.3 Loss Functions

Task-specific loss functions were used as follows:

- Binary AMD detection: Standard cross-entropy loss, CrossEntropyLoss(), was applied because the binary task is less sensitive to class imbalance than staging (depending on class distribution after grouping all non-AMD diseases).
- AMD staging: Class-weighted cross-entropy loss, Class-weighted cross-entropy: CrossEntropyLoss($weight =$ `class_weights`), was used to address the strong class imbalance across stages (late-stage cases being dominant). Class weights were computed from the training split and normalized before being applied to the loss.

4.10.4 Optimizer and Regularization (AdamW)

All models were optimized using AdamW, which decouples weight decay regularization from the adaptive gradient updates. This optimizer is commonly preferred for transfer learning as it improves generalization and stabilizes fine-tuning.

- Optimizer: torch.optim.AdamW
- Weight decay:
  - Detection: 1e-4
  - Staging: 3e-4

The larger weight decay for staging reflects the increased risk of overfitting on the smaller AMD-only subset.

4.10.5 Learning-Rate Scheduling

To adapt learning rates dynamically and reduce overfitting when validation performance plateaus, ReduceLROnPlateau was used:

- mode: ”min” (monitors validation loss)
- factor: 0.5 (halve the LR when plateau detected)
- patience: 2 epochs
- minimum learning rate: 1e-6

This scheduler reduces the learning rate only when necessary, supporting stable convergence in both head-training and full fine-tuning phases.

4.10.6 Model-Specific Learning Rates (Head vs Fine-Tune)

Because backbone architectures differ in capacity, training stability, and sensitivity to fine-tuning, model-specific learning rates were adopted. Each model used:

- a larger learning rate during Phase 1 (training the classifier head), and
- a smaller learning rate during Phase 2 (fine-tuning the full network).

The learning rates used in the implementation were:

- ResNet50:
  - head_lr = 1e-3
  - ft_lr = 2e-4 (detection) / 5e-5 (staging)
- EfficientNetB4:
  - head_lr = 3e-4
  - ft_lr = 1e-4 (detection) / 5e-5 (staging)
- ConvNeXt Tiny / Small:
  - head_lr = 5e-4
  - ft_lr = 2e-4 (detection) / 5e-5 (staging)
- EfficientNetV2-L:
  - head_lr = 1e-4
  - ft_lr = 5e-5 (detection) / 2e-5 (staging)

4.10.7 Epoch Budget, Early Stopping, and Best-Model Selection

Each experiment was trained for a maximum of 25 epochs. To prevent overfitting and to reduce unnecessary computation, early stopping was applied:

- maximum epochs: 25
- early stopping patience: 7 epochs without meaningful validation loss improvement
- best checkpoint: model state corresponding to the minimum validation loss

This is particularly important for medical datasets with limited sample size and high risk of overfitting.

4.10.8 Gradient Clipping

To improve numerical stability (especially during fine-tuning) and reduce the risk of exploding gradients, gradient norm clipping was applied:

- global max gradient norm: 1.0

4.11 Testing and Evaluation Protocol

4.11.1 Probability Generation and Numerical Sanitization

Midway through testing, logits straight from the model got turned into probabilities via softmax. Not every output stayed unchanged - some shifted sharply after scaling. One step changed everything: raw scores became likelihoods. Softmax stepped in right there, reshaping values piece by piece. Each number found its place within a distribution only then. Before that moment, nothing resembled chance. Only once reworked did results reflect possible outcomes clearly.
$p = \text{softmax}(\text{logits},\ \text{dim} = 1)$
The probability values come out of a softmax function applied along dimension one of the logits tensor.
Sometimes numbers like NaN or Inf show up by accident. A cleanup process steps in when they do. Wrong probability entries get swapped out. Afterward the list of chances gets adjusted so everything adds up right. That keeps measurements like ROC-AUC running without hiccups. Testing stays smooth even when odd cases appear.

4.11.2 Loss Curve Visualization (Training Diagnostics)

For each experiment, the training procedure recorded and visualized the following metrics per epoch:

- Training loss
- Validation loss

These loss curves provide insight into the convergence behavior of each model and can reveal overfitting, such as when the training loss continues to decrease while the validation loss increases. Loss curves are reported for all models and tasks, supporting both qualitative and quantitative assessment of training dynamics.

4.11.3 Classification Report

Model performance on the test set is summarized using a standard classification report, including:

- per-class precision, recall, and F1-score
- macro-averaged and weighted-averaged metrics (as applicable)

This is generated using classification_report(y_true, y_pred, ...) and supports detailed error characterization, especially for minority classes (e.g., early-stage AMD).

4.11.4 Confusion Matrix and Error Analysis

Fresh out of testing, confusion matrices take shape right there on the unseen data chunk
confusion_matrix(y_true, y_pred)

Heatmaps made with Seaborn showed the data clearly. What you see in confusion matrices reveals exactly where classes get mixed up - especially when telling apart close stages of AMD, like early versus intermediate, becomes tough.

4.11.5 ROC-AUC Computation

Area Under the Receiver Operating Characteristic Curve (ROC-AUC) is used to evaluate ranking quality and probability separation.

- Detection (binary): ROC-AUC computed using the probability of the AMD class (positive class).
- Staging (multi-class): macro-average one-vs-rest (OvR) ROC-AUC computed across the three classes, and additionally per-class OvR AUC values are computed when applicable.

4.11.6 Metrics for Imbalanced Multi-Class Staging

In multi-class staging tasks, class imbalance can render overall accuracy misleading. Therefore, evaluation emphasizes imbalance-aware metrics, defined as follows:

- Macro precision / Macro recall / Macro F1: Treat all classes equally, regardless of class frequency.
- Weighted F1: Reflects overall performance while accounting for class proportions.
- Balanced accuracy: Computes the average recall across classes.
- Per-class recall: Reports sensitivity for each stage individually (Early, Intermediate, Late).
- Per-class one-vs-rest (OvR) AUC: Measures the separability of each class based on predicted probabilities.

These metrics collectively provide a clinically meaningful assessment. Special attention is given to high recall for early-stage AMD, where timely intervention has the greatest clinical impact.

4.12 Interpretability: Grad-CAM Implementation and Clinical Utility (XAI)

Deep learning models can achieve high predictive performance on medical images while remaining difficult to interpret, which can limit clinical adoption. To address this concern, this study incorporates an explainable AI (XAI) component using Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM provides a class-discriminative localization map that highlights image regions contributing most strongly to a given prediction. In the context of OCT-based AMD detection and staging, Grad-CAM visualizations support qualitative verification that the

model is attending to clinically relevant retinal structures (e.g., regions corresponding to drusen-associated RPE elevation, outer retinal disruption, or areas of atrophy) rather than spurious artifacts or background patterns.

Target layer selection. Grad-CAM requires a convolutional feature representation from a deep layer that preserves semantic information while maintaining spatial structure. Accordingly, the implementation selects the final feature stage of each backbone as the Grad-CAM target layer: for ResNet50, the last residual block of the final convolutional stage (model.layer4[-1]) is used; for EfficientNetB4 and EfficientNetV2-L, the final block of the feature extractor (model.features[-1]) is selected; and for ConvNeXt Tiny/Small, the last feature stage (model.features[-1]) is used. These selections ensure that the activation maps are derived from high-level discriminative representations while remaining spatially meaningful for heatmap generation.

Grad-CAM computation. For a given input image, Grad-CAM is computed by first performing a forward pass and saving the activations of the selected target layer using a forward hook. A backward pass is then performed with respect to the score of the predicted class (or a user-specified class), and the gradients of this score with respect to the target-layer activations are captured using a backward hook. Channel-wise importance weights are obtained by global average pooling of the gradients over the spatial dimensions. The final class activation map is produced by computing a weighted sum of the stored activations using these importance weights, followed by a ReLU operation to retain only positive contributions. Finally, the heatmap is min–max normalized to the range [0,1] to allow consistent visualization across samples.

Visualization and clinical utility. To make explanations interpretable to clinicians, the Grad-CAM heatmap is resized to the original image resolution, mapped to a perceptually intuitive color scale (jet colormap), and then alpha-blended with the original OCT image to produce an overlay. For each explained case, three panels are reported: (1) the original OCT scan, (2) the Grad-CAM heatmap, and (3) the heatmap overlay. These qualitative explanations are used to (i) assess whether the model focuses on anatomically plausible retinal regions, (ii) identify common patterns in correct predictions for each AMD stage, and (iii) analyze failure cases (e.g., early vs intermediate confusion) by inspecting which structures the model emphasizes. Grad-CAM results are presented as representative examples in the Results section to complement quantitative metrics and strengthen the clinical credibility of the proposed approach.

4.13 Experimental Workflow and Work Plan

4.13.1 Phase I: Dataset Preparation

- Acquisition, filtering, and labeling of OCT images
- Verification of image-label consistency
- Train-validation-test splitting

4.13.2 Phase II: Preprocessing and Augmentation

- Image resizing, normalization, RGB conversion

- Data augmentation pipeline design and implementation
- Data loader configuration for efficient batch processing

4.13.3 Phase III: Model Development

- Backbone selection and classifier head design
- Transfer learning integration
- Initialization of hyperparameters

4.13.4 Phase IV: Model Training

- Two-stage training: detection $\rightarrow$ staging
- Loss weighting and optimizer configuration
- Early stopping and checkpointing

4.13.5 Phase V: Evaluation and Interpretability

- Quantitative metric computation
- Confusion matrix analysis
- Grad-CAM visualization and clinical interpretability assessment

4.13.6 Phase VI: Reporting

- Comparative performance evaluation across models
- Visualizations for both detection and staging tasks
- Documentation of methodology and experimental workflow

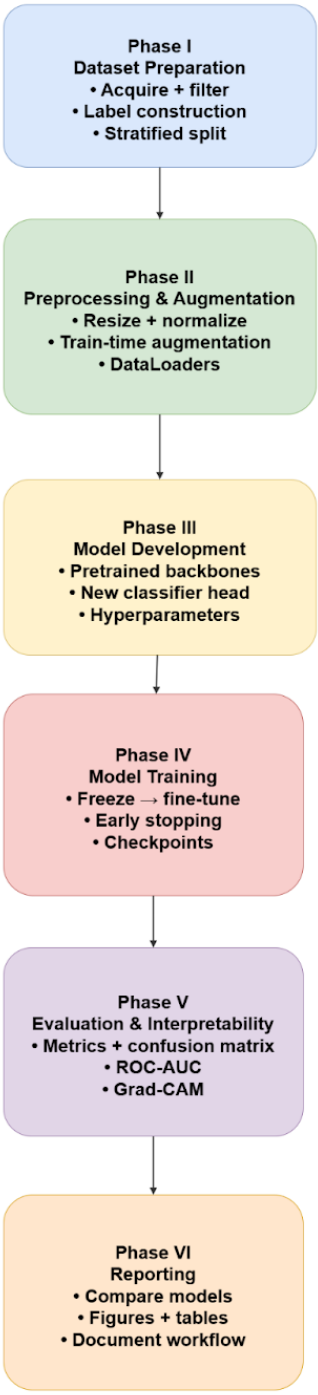

Figure 4.6: Experimental Workflow

Chapter 5

Results and Discussion

This section summarizes the implementation-level performance of the evaluated deep learning backbones under the unified pipeline described in Chapter 4. Results are reported for two tasks: (i) binary AMD detection (Not-AMD vs AMD) and (ii) AMD staging (Early vs Intermediate vs Late). Evaluation is conducted on the held-out test sets using classification reports, confusion matrices, and training/validation loss curves.

5.1 ResNet50 — Results

5.1.1 Binary AMD Detection (y_det)

On the binary screening task, ResNet50 achieved an overall accuracy of 0.9753 on 243 test images, indicating strong discrimination between AMD and non-AMD categories. Class-wise performance was high for both classes:

- Not-AMD: precision 0.9561, recall 0.9909, F1-score 0.9732 (support = 110)
- AMD: precision 0.9922, recall 0.9624, F1-score 0.9771 (support = 133)

The confusion matrix shows very few misclassifications: TN = 109, FP = 1, FN = 5, TP = 128, suggesting that false-positive screening alarms were rare, while the remaining errors were primarily false negatives (missed AMD). From a clinical screening perspective, reducing false negatives is particularly important because it corresponds to missed referrals or delayed follow-up.

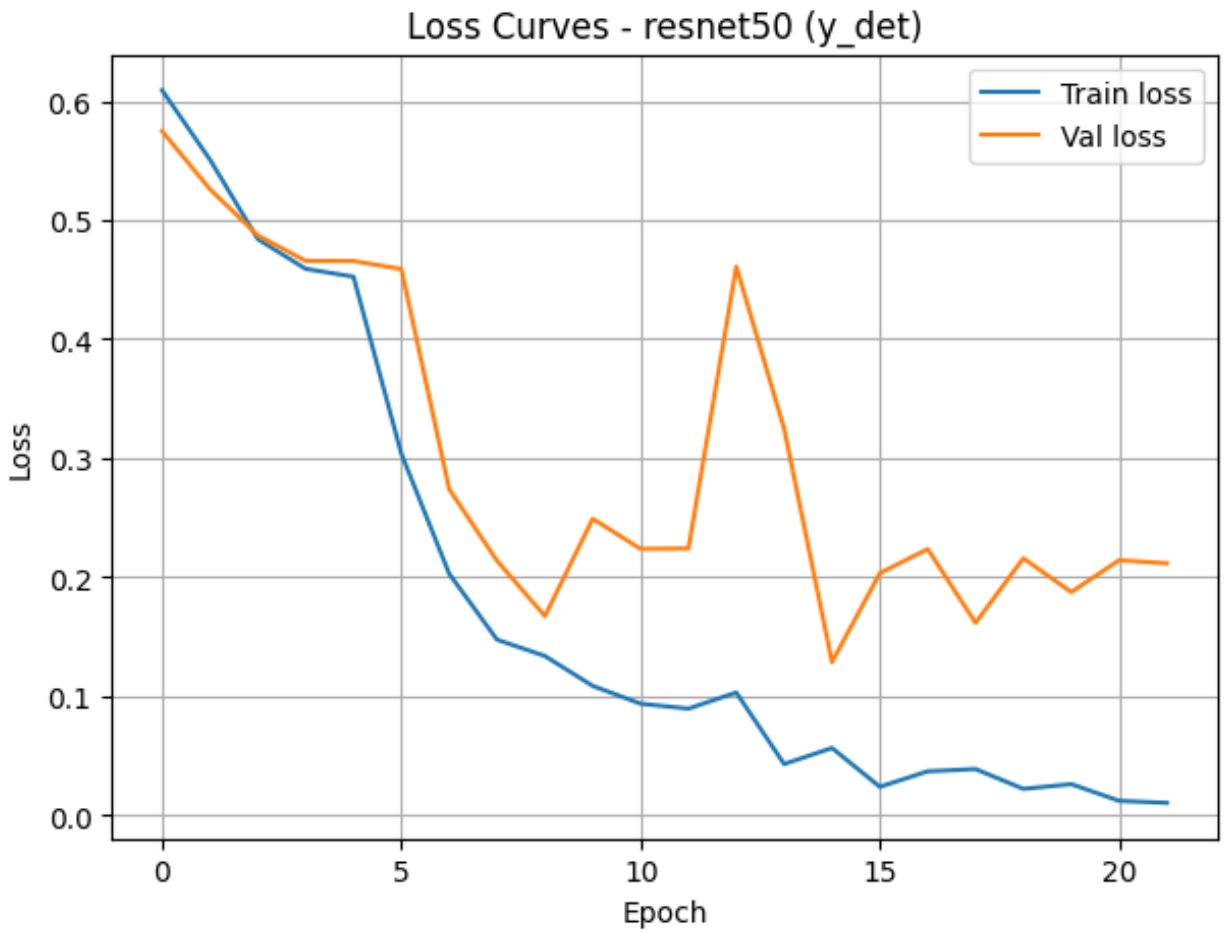

Figure 5.1: ResNet50 detection loss curves here

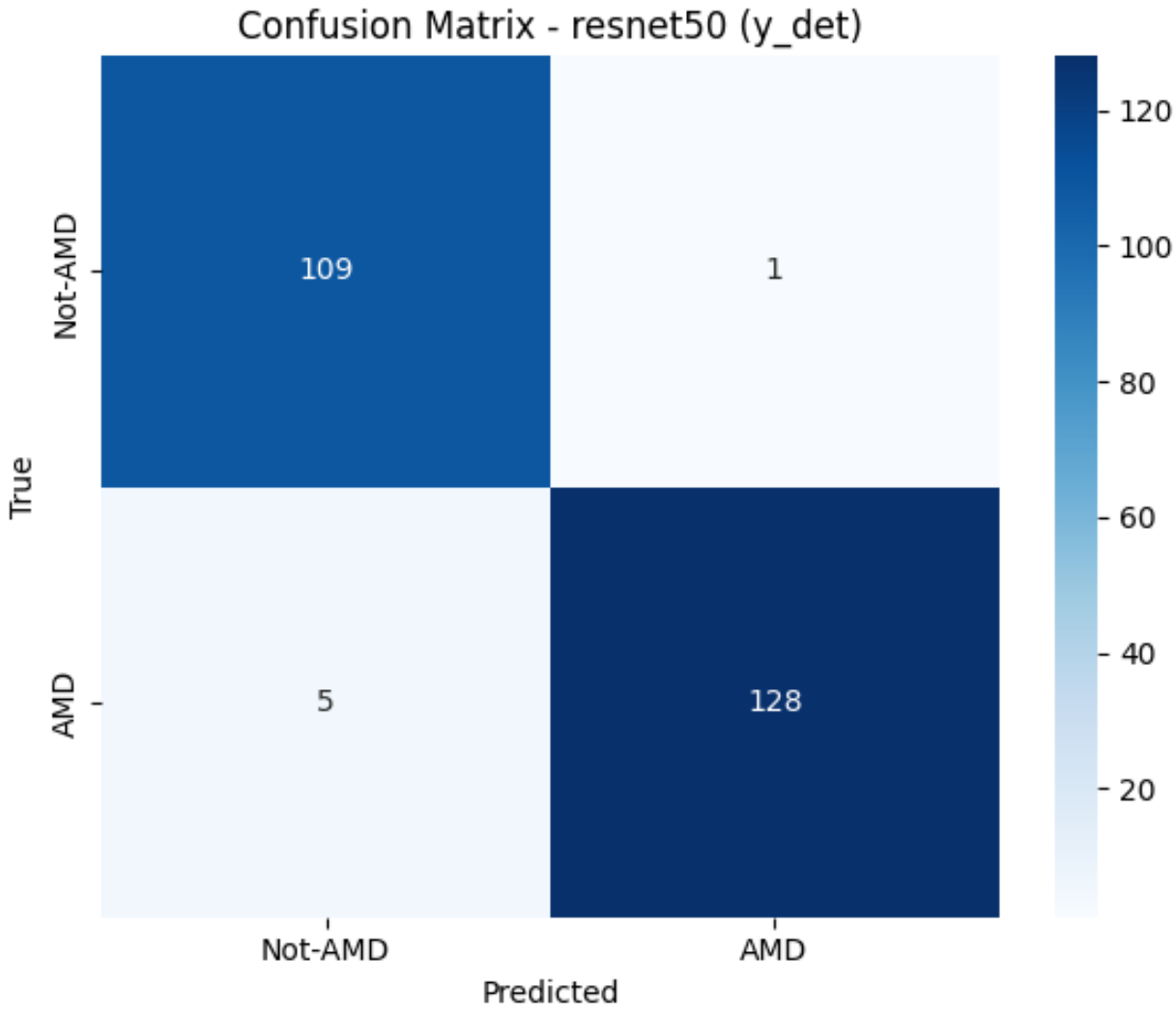

Figure 5.2: ResNet50 detection confusion matrix

5.1.2 AMD Staging (y_stage)

For three-class AMD staging, ResNet50 achieved an overall accuracy of 0.7744 on 133 test images. Given the class imbalance toward late AMD, the macro-averaged metrics provide a more balanced view of multi-class performance:

- Macro F1-score: 0.7136
- Weighted F1-score: 0.7799

Per-class performance was:

- Early: precision 0.6279, recall 0.9310, F1-score 0.7500 (support = 29)
- Intermediate: precision 0.5200, recall 0.5000, F1-score 0.5098 (support = 26)
- Late: precision 0.9692, recall 0.8077, F1-score 0.8811 (support = 78)

A key observation is the very high recall for Early-stage AMD (0.9310), indicating that the model is sensitive to early disease indicators. However, the Intermediate class remains the most challenging, with notable confusion toward both Early and Late stages. This pattern is consistent with the transitional nature of intermediate AMD and the overlap of structural manifestations across adjacent stages.

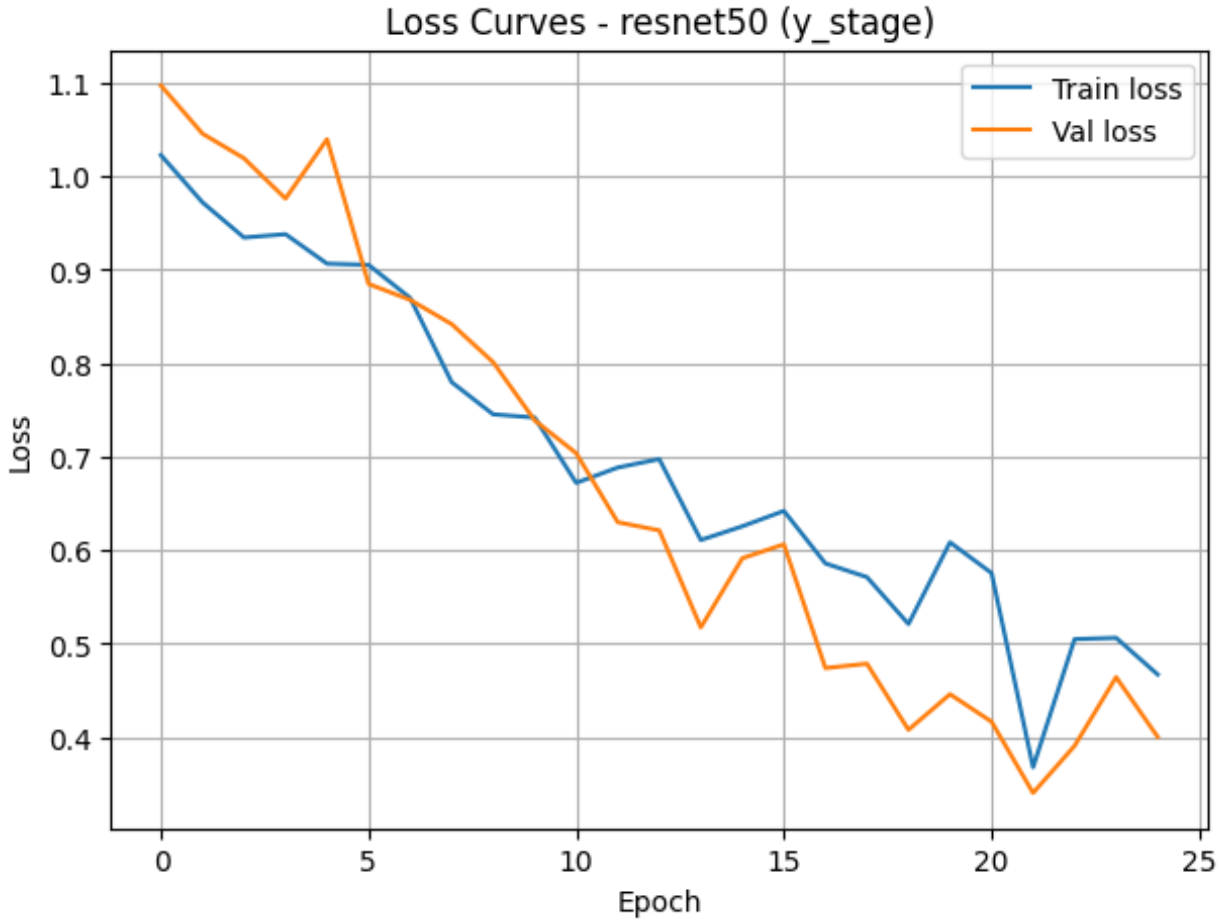

Figure 5.3: ResNet50 staging loss curves

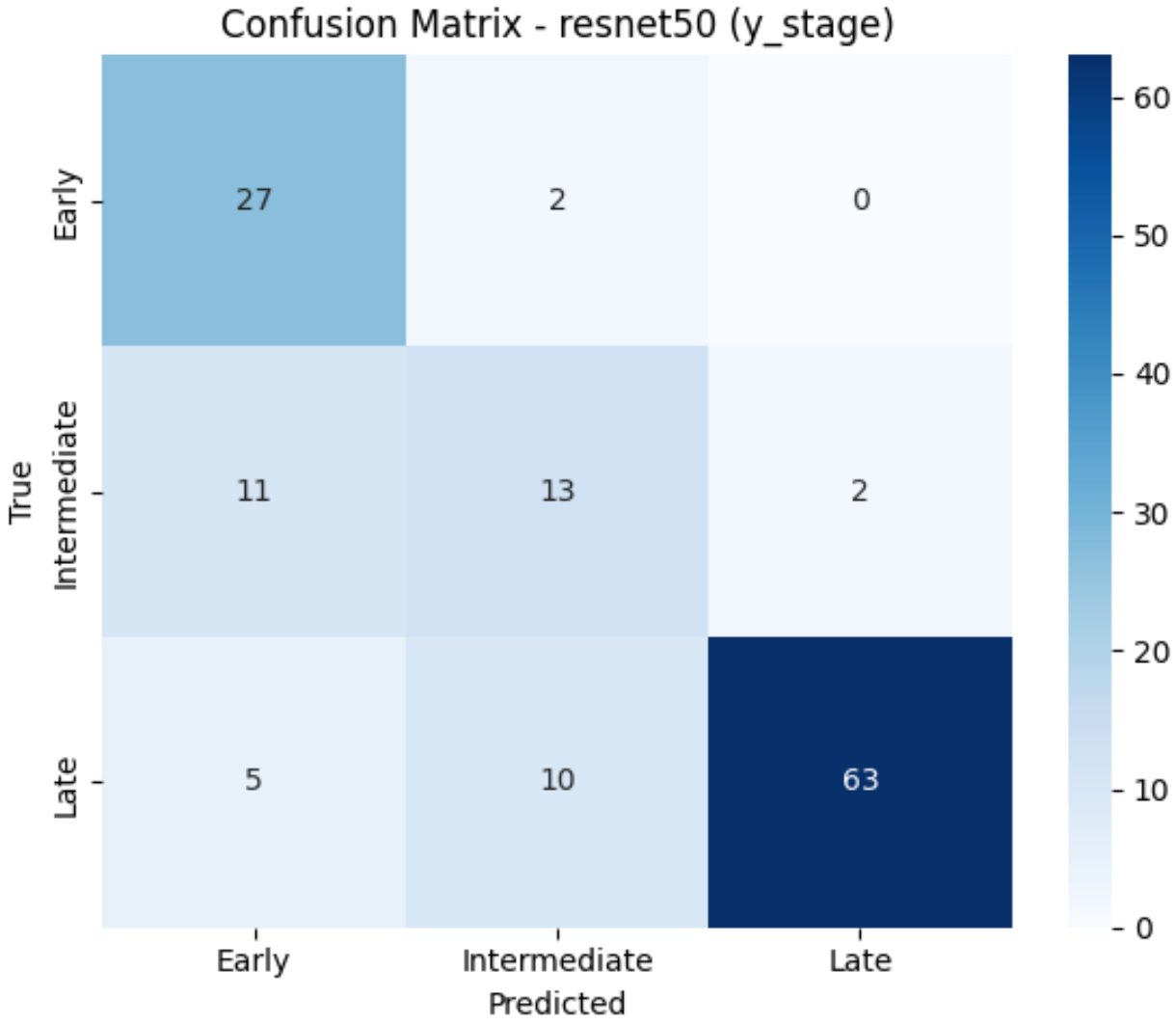

Figure 5.4: ResNet50 staging confusion matrix here

5.1.3 Summary (ResNet50)

Overall, ResNet50 provides excellent performance for binary AMD detection and moderate performance for AMD staging. It is particularly effective as the first stage in the proposed hierarchical pipeline. Staging performance indicates that additional representational capacity, enhanced imbalance strategies, or more specialized architectures may be required to improve discrimination of the Intermediate stage.

5.2 EfficientNetB4 — Results

5.2.1 Binary AMD Detection (y_det)

EfficientNetB4 achieved strong performance on the binary AMD detection task, with an overall accuracy of 0.9753 on 243 test images. Class-wise performance indicates that the model maintains high precision and recall for both classes:

- Not-AMD: precision 0.9561, recall 0.9909, F1-score 0.9732 (support = 110)

- AMD: precision 0.9922, recall 0.9624, F1-score 0.9771 (support = 133)

The confusion matrix shows the same error distribution observed in this run:
TN = 109, FP = 1, FN = 5, TP = 128
This implies very low false positives (only 1 Not-AMD image flagged as AMD) and a small number of false negatives (5 AMD images missed). The loss curves for detection indicate rapid convergence early in training, followed by a stable low training loss and comparatively small validation loss variation, suggesting effective optimization under the chosen transfer-learning schedule.

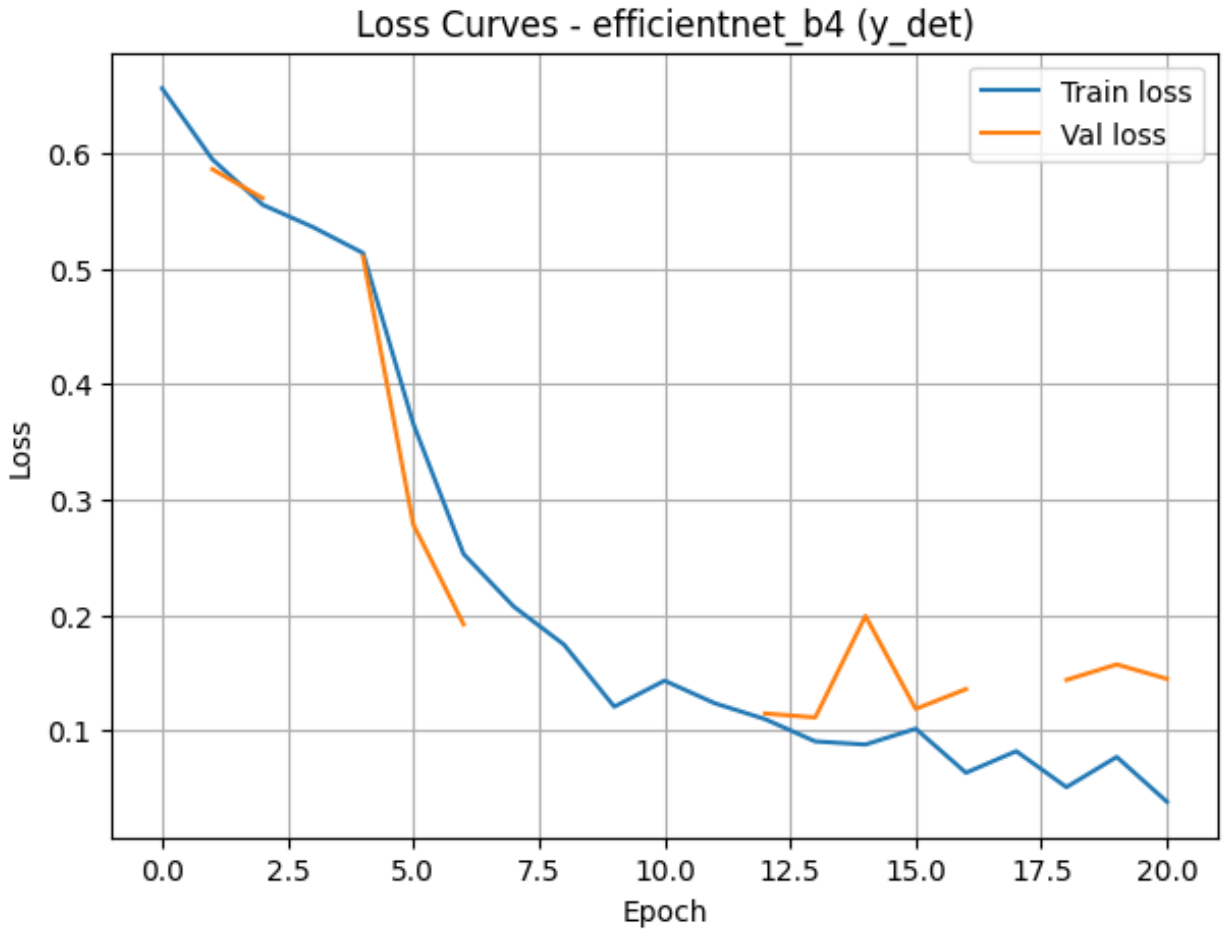

Figure 5.5: EfficientNetB4 detection loss curves

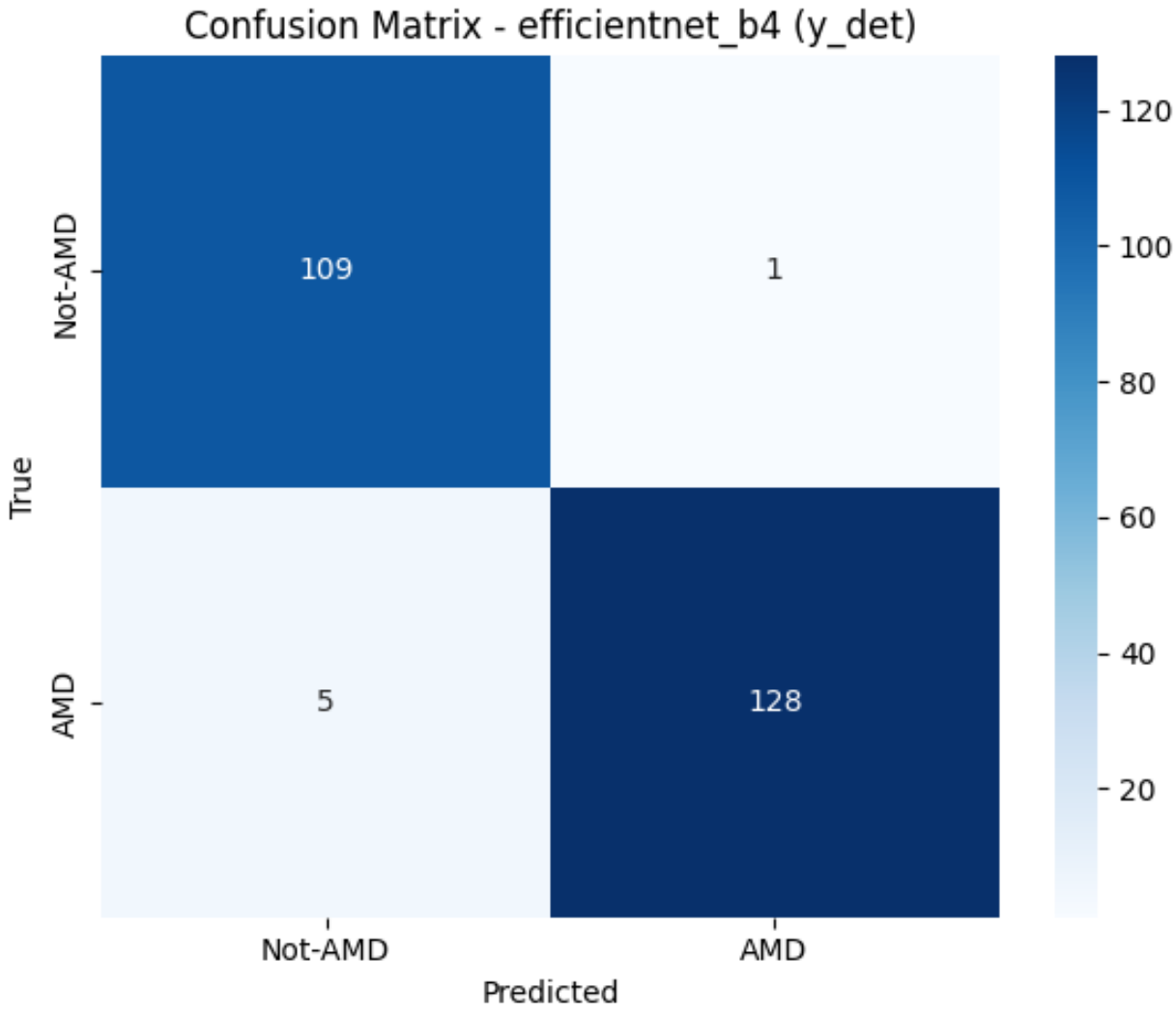

Figure 5.6: EfficientNetB4 detection confusion matrix

5.2.2 AMD Staging (y_stage)

On the three-class AMD staging task, EfficientNetB4 achieved an overall accuracy of 0.6992 on 133 test images. The macro-averaged metrics were:

- Macro F1-score: 0.6882
- Weighted F1-score: 0.7188

Per-class performance was:

- Early: precision 0.6944, recall 0.8621, F1-score 0.7692 (support = 29)
- Intermediate: precision 0.4082, recall 0.7692, F1-score 0.5333 (support = 26)
- Late: precision 1.0000, recall 0.6154, F1-score 0.7619 (support = 78)

These results indicate a characteristic imbalance-related trade-off. The model shows high recall which means the minority class is getting showcased better but for the

intermediate stage it shows low precision meaning over scaning happening. On the other hand late stage shows perfect precision but lesser recall meaning the model predicts late stage correctly but it fails to identify most of the time. The confusion matrix provides a clear explanation for these trends:

- Early: 25 correct, 4 predicted as Intermediate, 0 as Late
- Intermediate: 20 correct, 6 predicted as Early, 0 as Late
- Late: 48 correct, 25 predicted as Intermediate, 5 predicted as Early

The dominant error is Late → Intermediate (25 cases), which directly explains the reduced Late recall and the lowered Intermediate precision (Intermediate receives many Late samples). This suggests that EfficientNetB4, under the current training configuration, tends to interpret many Late-stage scans as Intermediate potentially reflecting sensitivity to intermediate-like drusen patterns while underweighting severe markers (e.g., advanced atrophy or late structural disruption) in some cases. Training behavior for staging shows steadily decreasing validation loss over epochs, indicating learning progress; however, the final classification behavior reveals that improved calibration or stronger separation for the Late stage may be required.

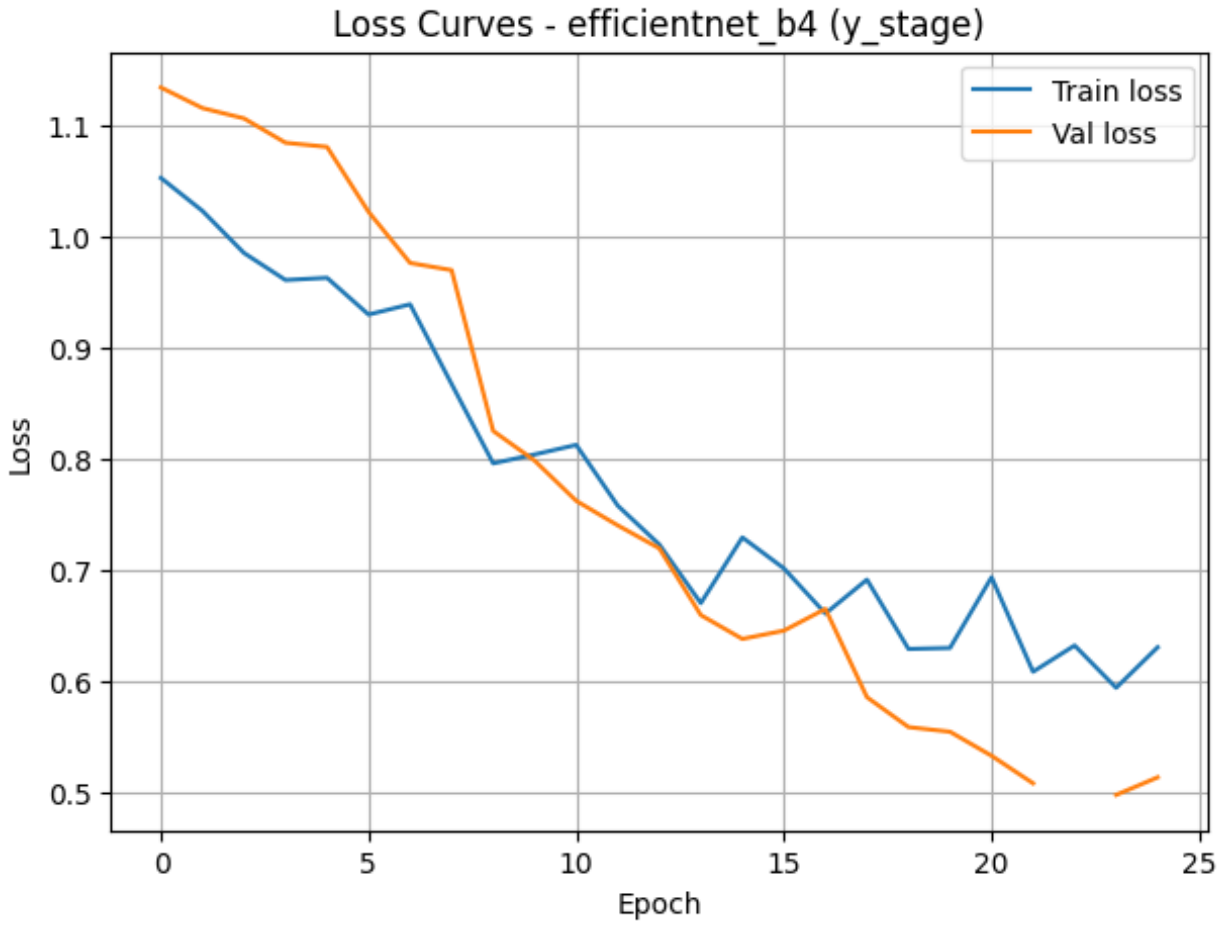

Figure 5.7: EfficientNetB4 staging loss curves

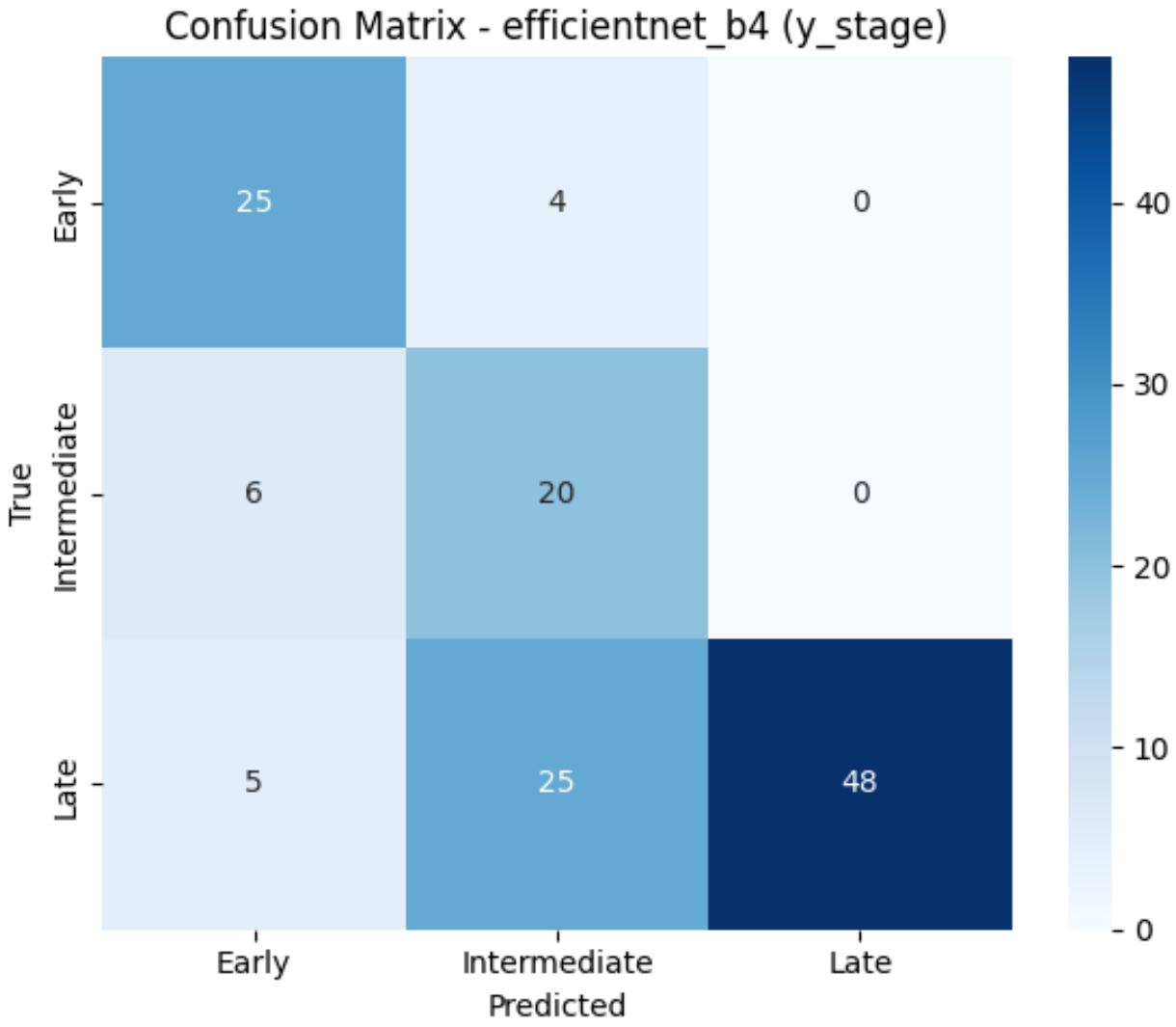

Figure 5.8: EfficientNetB4 staging confusion matrix

5.2.3 Summary (EfficientNetB4)

Overall, EfficientNetB4 performs very strongly for binary AMD detection (97.5% accuracy), comparable to the strongest backbones, and produces an extremely low false-positive rate. For AMD staging, performance is lower (69.9% accuracy) with a prominent confusion pattern where many Late-stage cases are predicted as Intermediate. This indicates that while the model learns features relevant for early detection and intermediate patterns, additional measures may be needed to improve Late-stage sensitivity and reduce Intermediate over-prediction.

5.3 ConvNeXt Tiny — Results

5.3.1 Binary AMD Detection (y_det)

ConvNeXt Tiny achieved strong performance on the binary AMD detection task, reaching an overall accuracy of 0.9671 on 243 test images. The class-wise results were:

- Not-AMD: precision 0.9554, recall 0.9727, F1-score 0.9640 (support = 110)
- AMD: precision 0.9771, recall 0.9624, F1-score 0.9697 (support = 133)

The confusion matrix indicates the following outcomes:
TN = 107, FP = 3, FN = 5, TP = 128
Relative to the strongest detection runs, ConvNeXt Tiny produces slightly more false positives (3 Not-AMD scans predicted as AMD), while maintaining the same number of false negatives (5). This suggests that the model remains sensitive to AMD, but is marginally more likely to flag non-AMD cases as AMDan acceptable trade-off in some screening settings where sensitivity is prioritized, but still relevant when considering referral burden.
The detection loss curves show rapid reduction in both training and validation loss in early epochs, followed by stabilization at low training loss. Validation loss decreases substantially after the early epochs and then fluctuates mildly, which is consistent with convergence under early stopping and with typical variance expected from limited validation data.

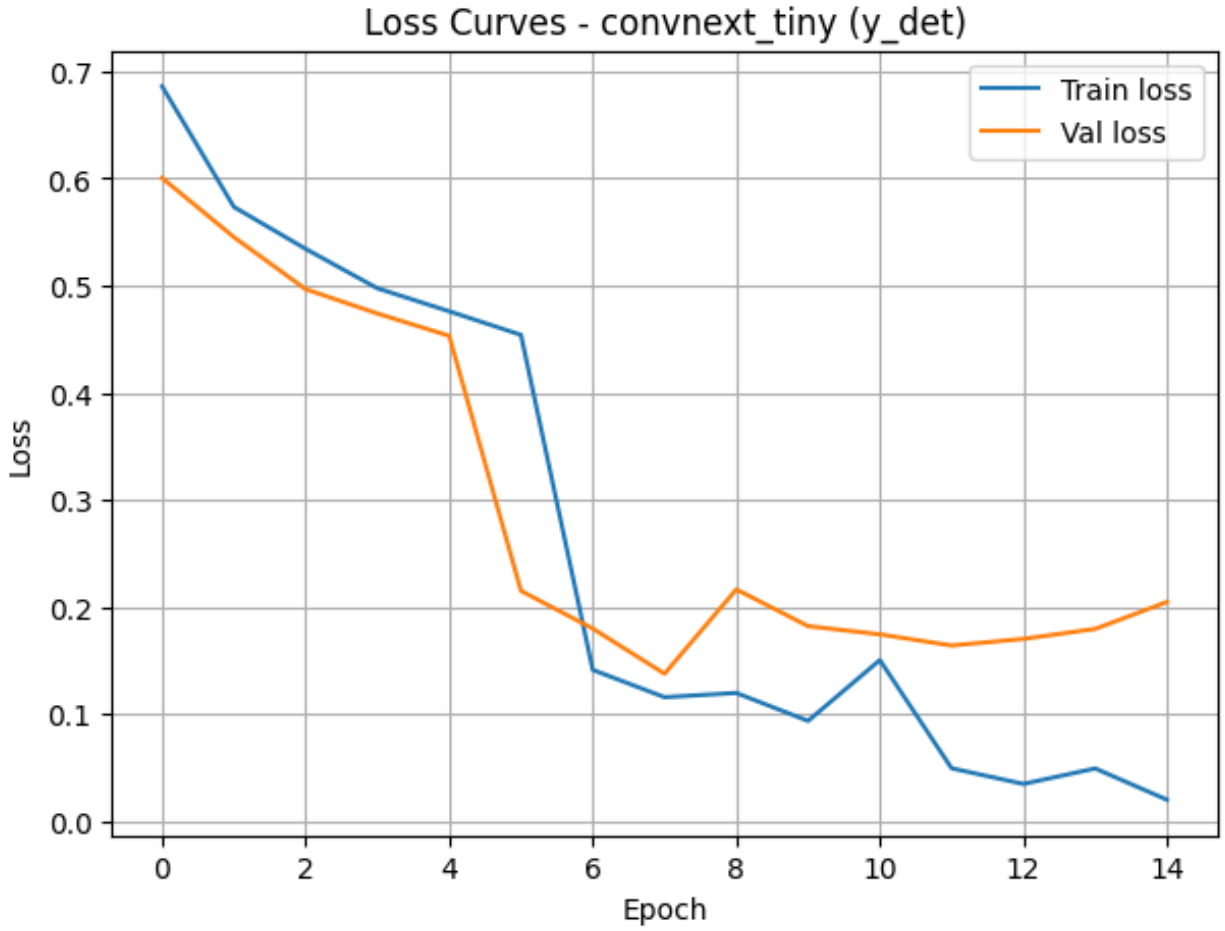

Figure 5.9: ConvNeXt Tiny detection loss curves

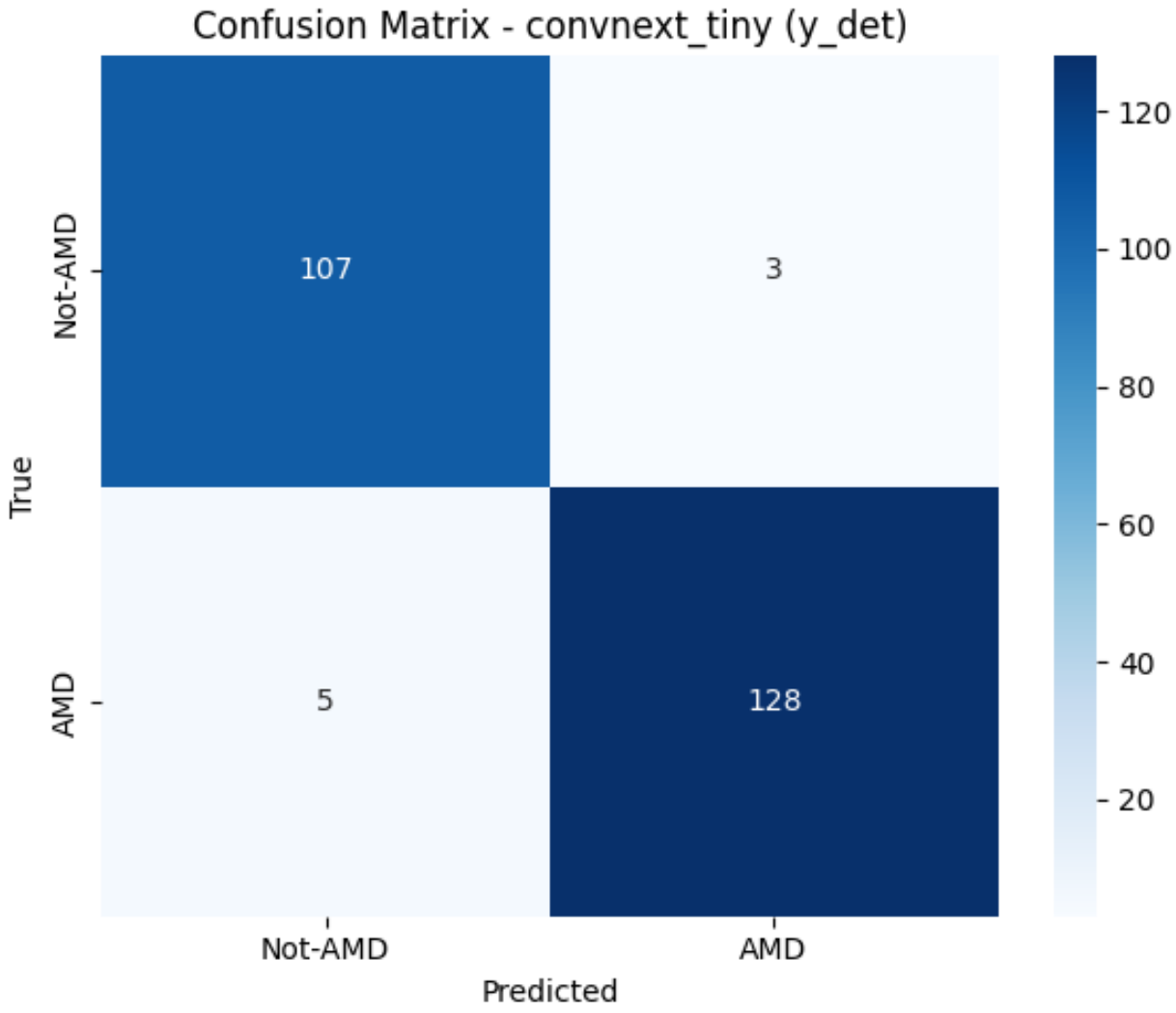

Figure 5.10: ConvNeXt Tiny detection confusion matrix

5.3.2 AMD Staging (y_stage)

For the three-class AMD staging task, ConvNeXt Tiny achieved an overall accuracy of 0.7820 on 133 test images. The macro and weighted metrics were:

- Macro F1-score: 0.7404
- Weighted F1-score: 0.7933

Per-class performance was:

- Early: precision 0.7143, recall 0.8621, F1-score 0.7812 (support = 29)
- Intermediate: precision 0.5000, recall 0.6538, F1-score 0.5667 (support = 26)
- Late: precision 0.9688, recall 0.7949, F1-score 0.8732 (support = 78)

These results indicate a balanced improvement in macro performance compared with some other backbones, suggesting that ConvNeXt Tiny better supports minority-class recognition than approaches that heavily favor the majority (Late) class. Early-stage recall remains strong (0.8621), while Intermediate performance is moderate but improved relative to a "majority-class" bias.
The staging confusion matrix shows the dominant error patterns:

- Early: 25 correct, 4 predicted as Intermediate, 0 as Late
- Intermediate: 17 correct, 7 predicted as Early, 2 predicted as Late
- Late: 62 correct, 13 predicted as Intermediate, 3 predicted as Early

As typically observed in AMD staging, the principal confusion occurs between Intermediate and the adjacent stages, especially Late → Intermediate (13 cases). Nevertheless, Late classification remains strong with high precision, and Early-stage predictions show relatively low leakage into Late.
Training curves for staging show a steady decrease of validation loss over training, with moderate oscillations in later epochs. The validation curve generally tracks the training curve, suggesting stable learning and controlled overfitting under the chosen augmentation and early-stopping settings.

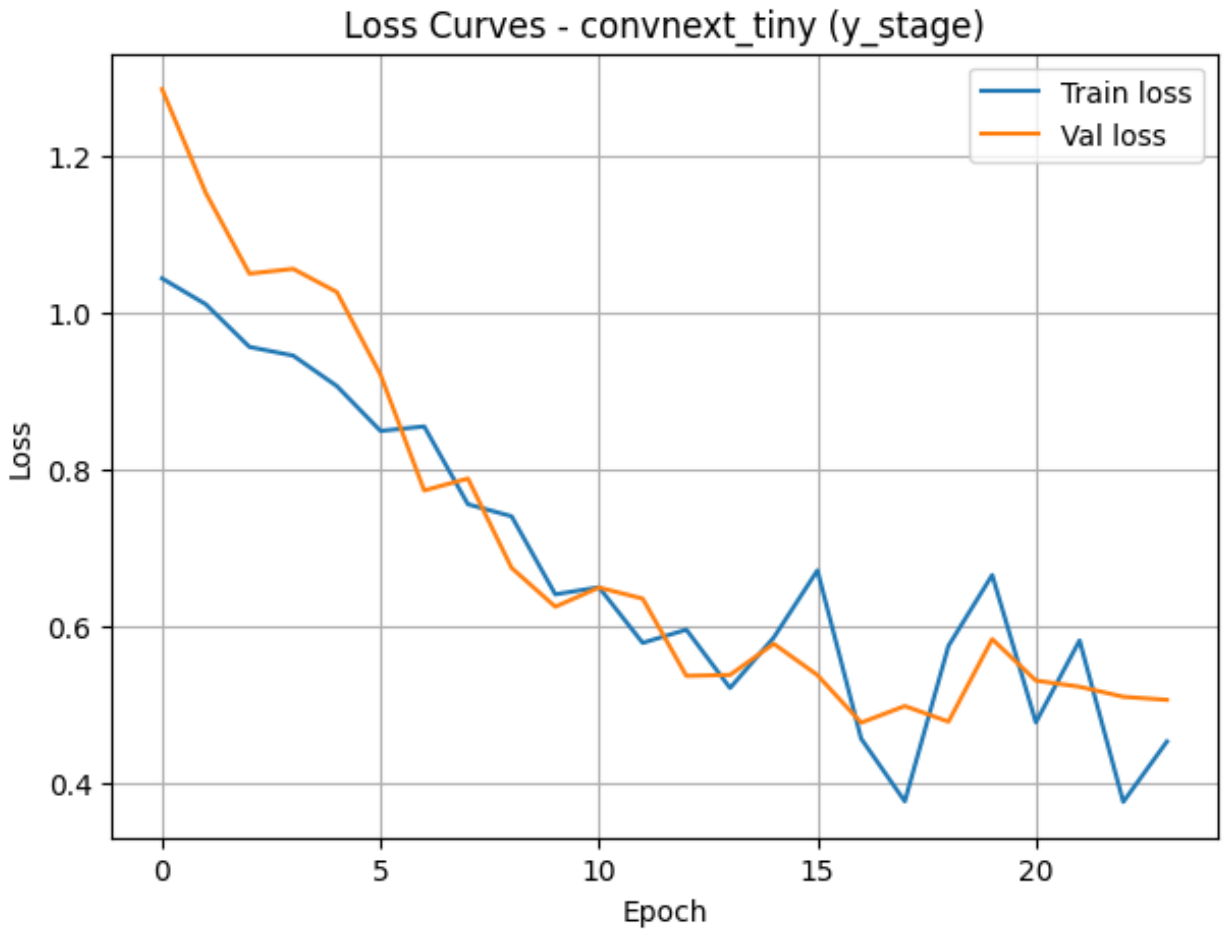

Figure 5.11: ConvNeXt Tiny staging loss curves

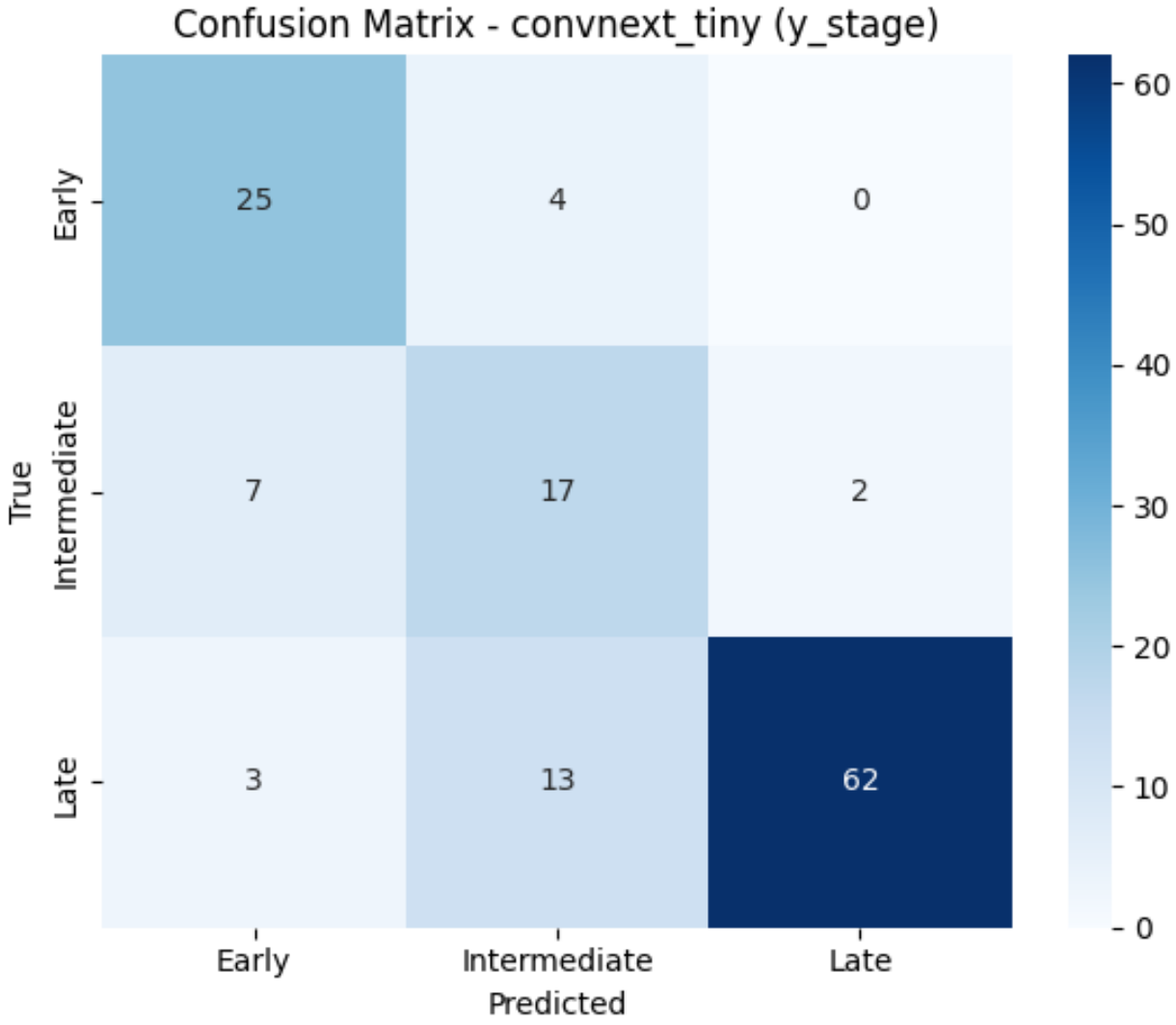

Figure 5.12: ConvNeXt Tiny staging confusion matrix

5.3.3 Summary (ConvNeXt Tiny)

In summary, ConvNeXt Tiny provides:

- High detection performance (96.7% accuracy) with a small increase in false positives compared with the best detection runs.
- Strong staging performance (78.2% accuracy) with improved macro-level balance, maintaining high recall for Early-stage AMD and strong Late-stage precision, while reducing (but not eliminating) Intermediate-stage ambiguity.

These results suggest ConvNeXt Tiny is a competitive backbone for the staging task, offering a strong trade-off between capacity and performance, and motivating evaluation of the larger ConvNeXt Small variant for potential further gains.

5.4 ConvNeXt Small — Results

ConvNeXt Small is selected as the best-performing backbone in this study based on its consistently strong performance across both tasks high reliability in binary AMD detection and the strongest overall staging performance among the evaluated models. In addition to quantitative metrics, this backbone is used for the interpretability analysis via Grad-CAM (reported separately), consistent with the study's decision to present XAI only for the final selected model.

5.4.1 Binary AMD Detection (y_det)

ConvNeXt Small achieved the highest detection performance in the provided results, reaching an overall accuracy of 0.9794 on 243 test images. Class-wise results were:

- Not-AMD: precision 0.9646, recall 0.9909, F1-score 0.9776 (support = 110)
- AMD: precision 0.9923, recall 0.9699, F1-score 0.9810 (support = 133)

The confusion matrix indicates:
TN = 109, FP = 1, FN = 4, TP = 129
This error profile is clinically favorable: false positives remain extremely low (1 case), while false negatives are reduced relative to several other backbones (4 missed AMD cases). The detection loss curves show rapid early convergence and low final training loss. Validation loss decreases substantially and remains relatively stable with mild fluctuations later in training, indicating strong optimization under the chosen fine-tuning schedule.

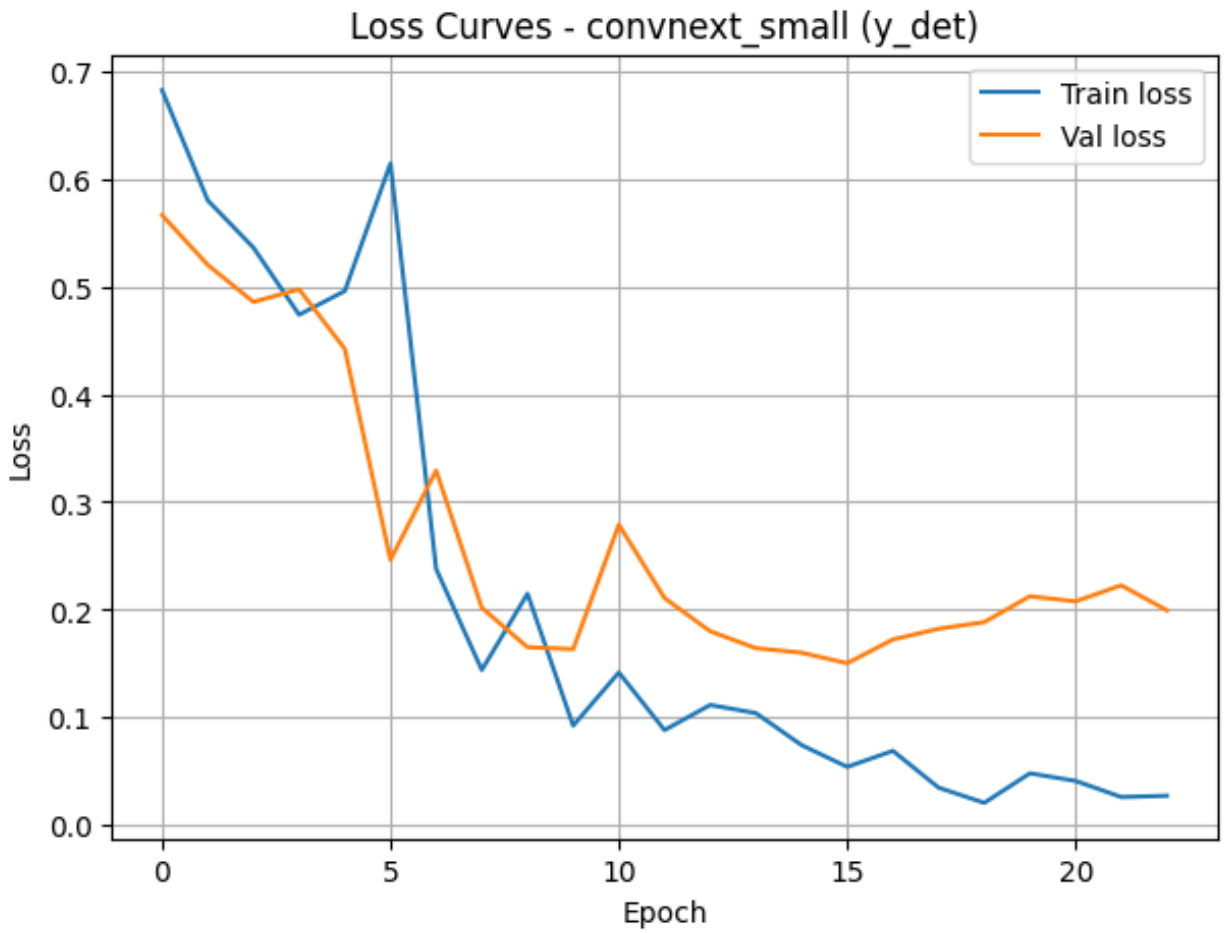

Figure 5.13: ConvNeXt Small detection loss curves

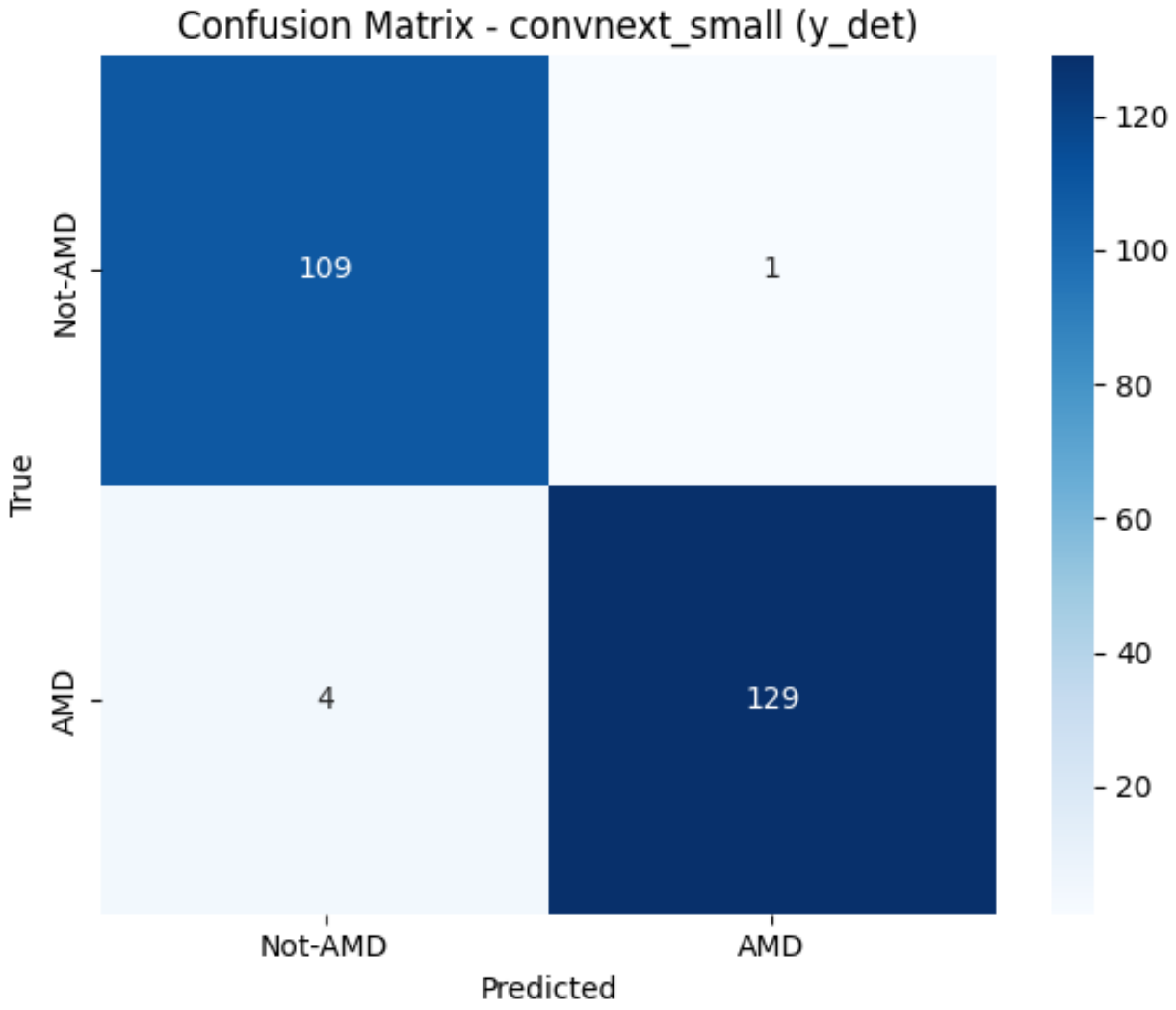

Figure 5.14: ConvNeXt Small detection confusion matrix

5.4.2 AMD Staging (y_stage)

ConvNeXt Small produced the strongest staging results among the models reported so far, with an overall accuracy of 0.8346 on 133 test images. Importantly, both macro and weighted metrics improve relative to prior backbones, suggesting better balanced performance across classes:

- Macro F1-score: 0.7636
- Weighted F1-score: 0.8306

Per-class results were:

- Early: precision 0.6667, recall 0.8966, F1-score 0.7647 (support = 29)
- Intermediate: precision 0.7222, recall 0.5000, F1-score 0.5909 (support = 26)
- Late: precision 0.9474, recall 0.9231, F1-score 0.9351 (support = 78)

These results indicate that ConvNeXt Small achieves very strong Late-stage recognition (high precision and high recall) while maintaining high sensitivity for Early-stage AMD (recall 0.8966). Intermediate remains the most challenging class, but its precision is improved compared to several backbones, suggesting fewer incorrect assignments into Intermediate.
The staging confusion matrix shows:

- Early: 26 correct, 3 predicted as Intermediate, 0 as Late
- Intermediate: 13 correct, 9 predicted as Early, 4 predicted as Late
- Late: 72 correct, 2 predicted as Intermediate, 4 predicted as Early

The dominant staging errors occur around the Intermediate class, which is consistent with the transitional nature of intermediate AMD and its overlap with adjacent stages. Notably, Late-stage performance is particularly strong (72/78 correct), with relatively few Late scans being downgraded to Intermediate (2 cases). This improved Late separability is a major advantage compared to EfficientNetB4, where Late → Intermediate confusion was substantial.
Training dynamics for staging show steadily decreasing validation loss and a stable gap between training and validation curves. The overall pattern suggests effective learning with controlled overfitting, supporting the selection of ConvNeXt Small as the final backbone for deployment in the hierarchical pipeline.

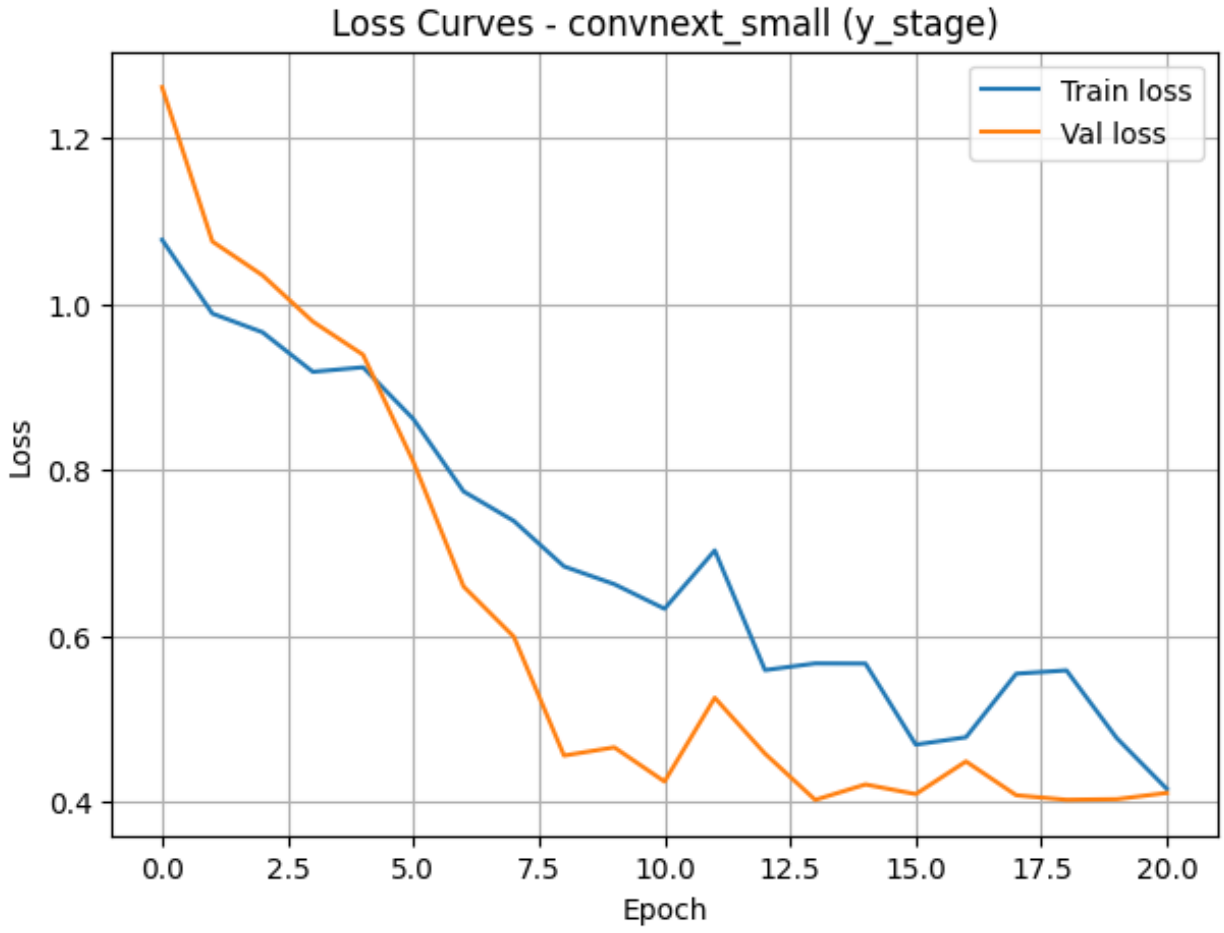

Figure 5.15: ConvNeXt Small staging loss curves

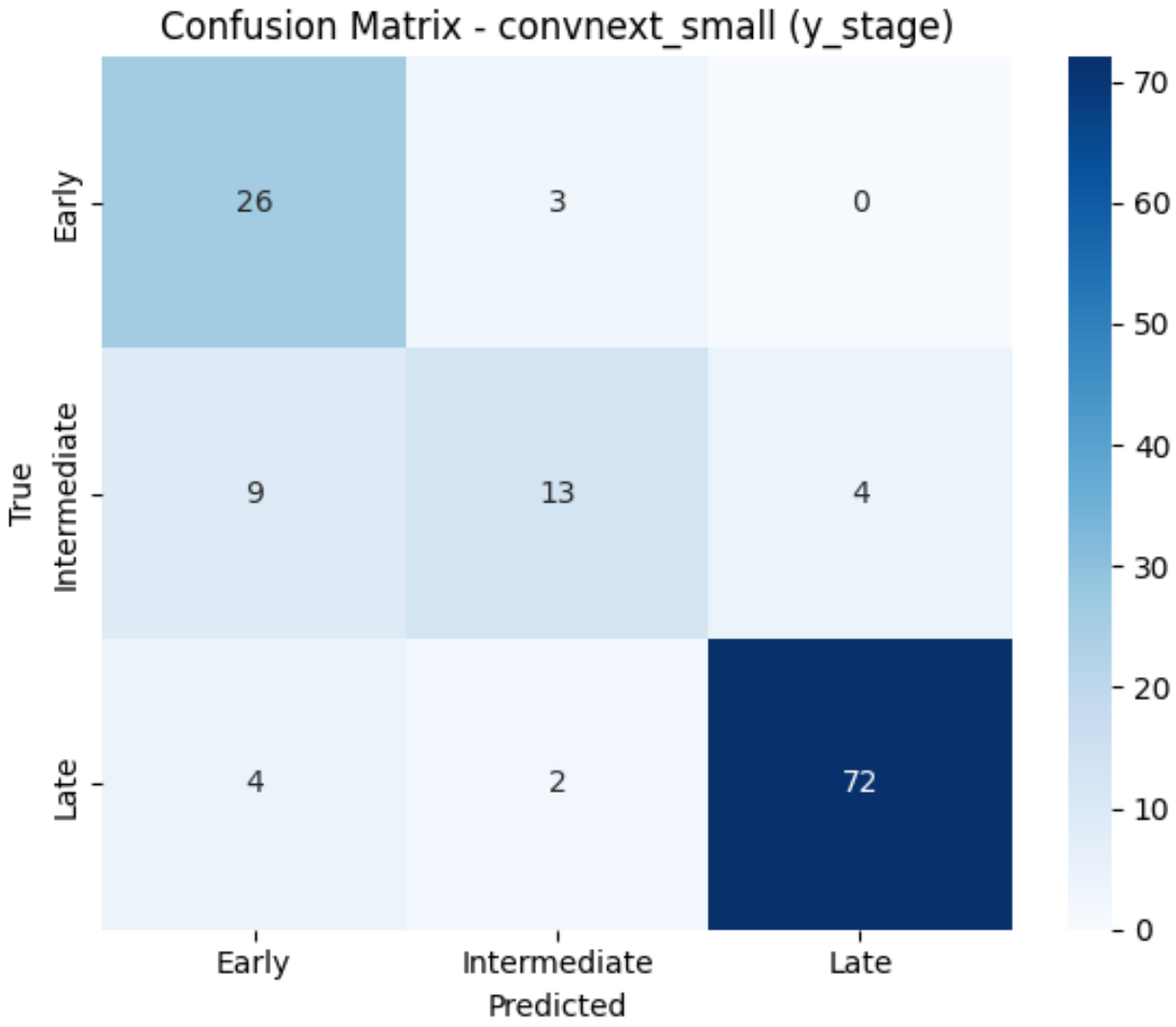

Figure 5.16: ConvNeXt Small staging confusion matrix

5.4.3 Summary (ConvNeXt Small)

ConvNeXt Small provides the strongest overall balance across tasks:

- Detection: highest accuracy (0.9794) with very low false positives and reduced false negatives.
- Staging: highest accuracy (0.8346) with improved macro/weighted F1 and notably strong Late-stage performance.

These results support ConvNeXt Small as the most suitable backbone for the proposed two-stage clinical pipeline (detection → conditional staging). Accordingly, the interpretability analysis using Grad-CAM is performed for this backbone to provide qualitative evidence that predictions are driven by clinically plausible retinal regions.

5.5 EfficientNetV2-L — Results

5.5.1 Binary AMD Detection (y_det)

EfficientNetV2-L achieved an overall accuracy of 0.9671 on 243 test images for binary AMD detection. The class-wise performance was:

- Not-AMD: precision 0.9474, recall 0.9818, F1-score 0.9643 (support = 110)
- AMD: precision 0.9845, recall 0.9549, F1-score 0.9695 (support = 133)

The confusion matrix indicates:
TN = 108, FP = 2, FN = 6, TP = 127
Compared with the strongest detection backbone (ConvNeXt Small), EfficientNetV2-L shows a slightly higher number of false negatives (6) and false positives (2). This suggests that the model is marginally less sensitive to AMD under the current configuration, which may be influenced by optimization stability constraints (AMP disabled) and the larger model capacity requiring careful fine-tuning.
The detection loss curves show training loss decreasing substantially after the early epochs, while validation loss decreases initially but fluctuates at a higher level thereafter. This pattern is consistent with effective fitting of the training data and potential limitations in generalization under the chosen epoch budget and regularization settings.

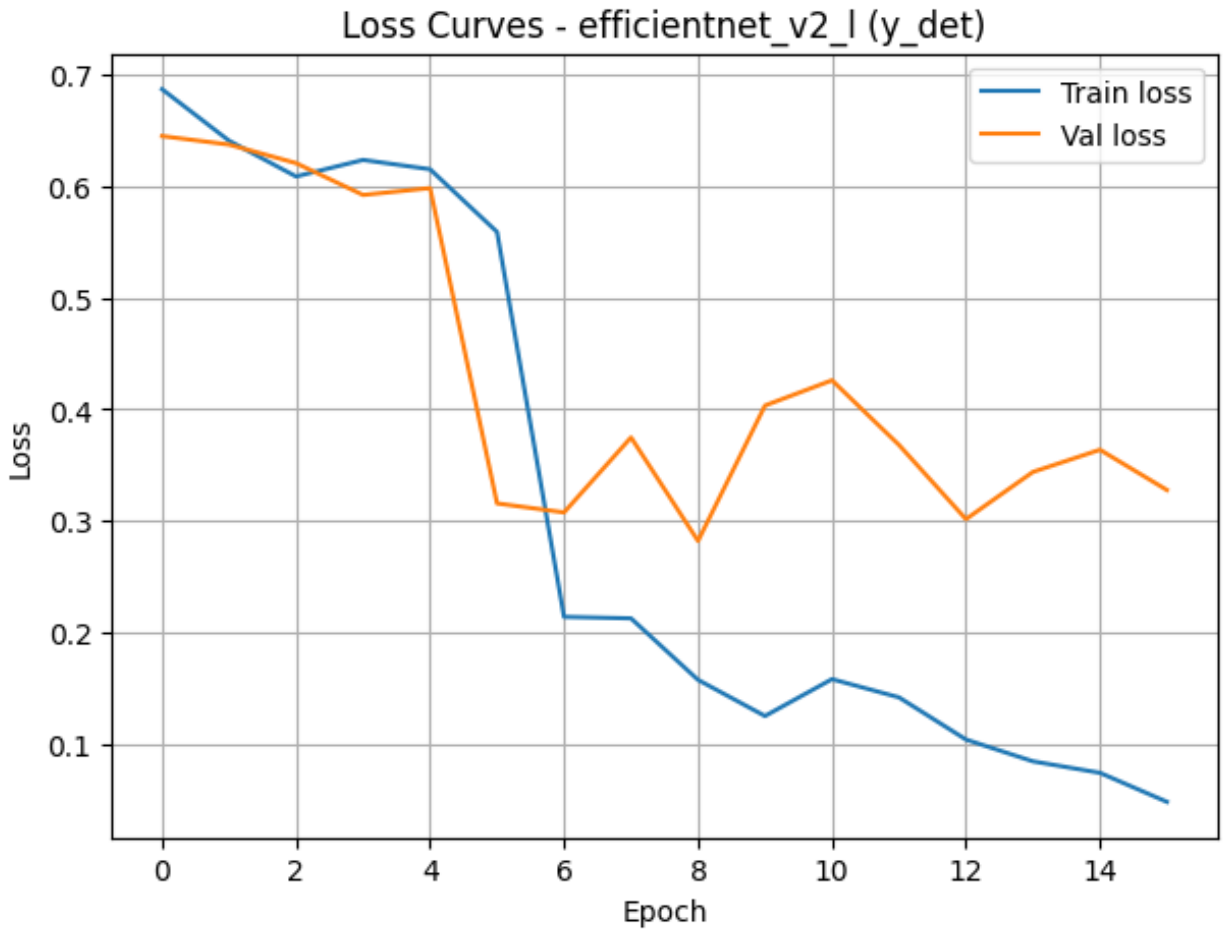

Figure 5.17: EfficientNetV2-L detection loss curve

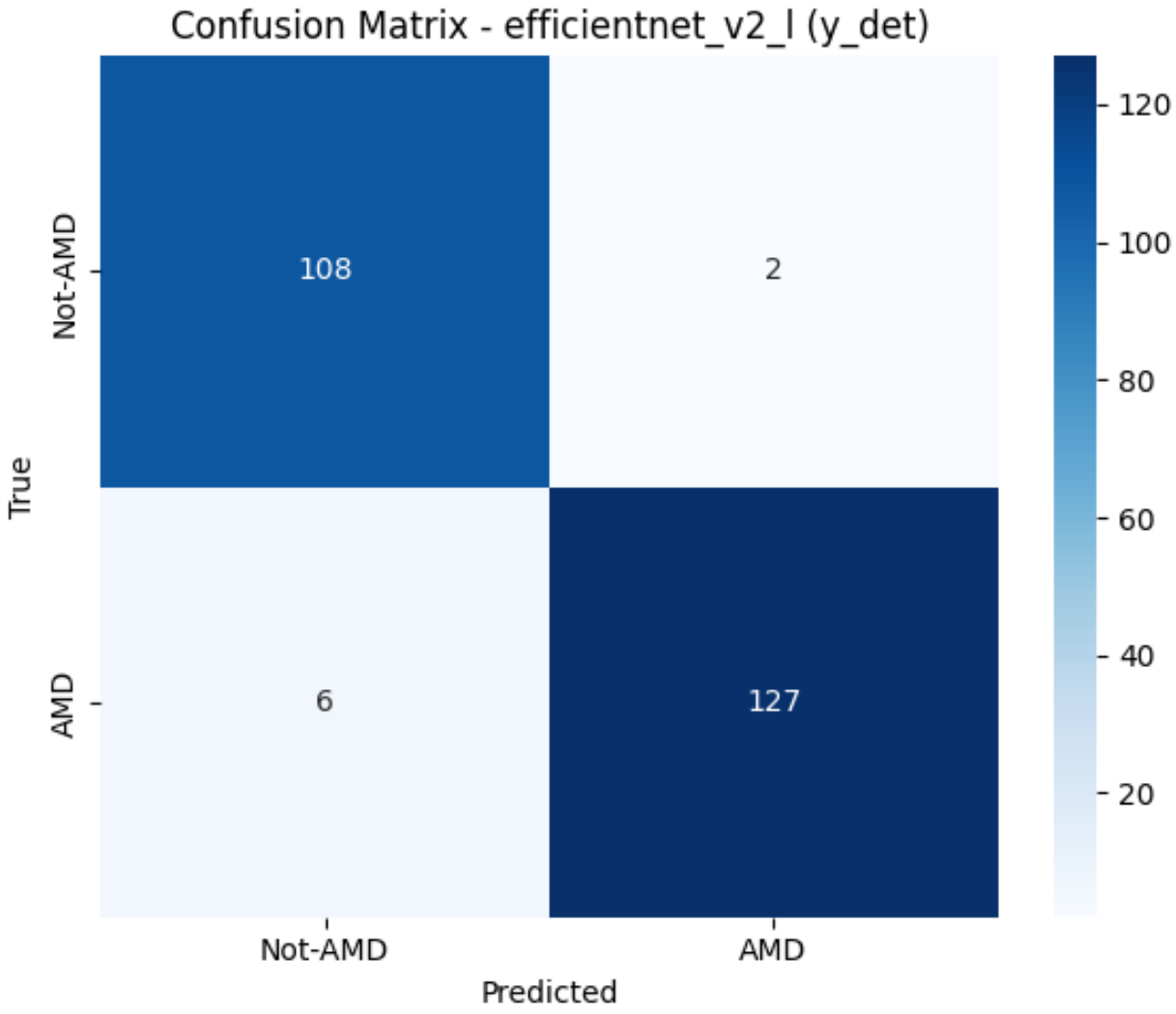

Figure 5.18: EfficientNetV2-L detection confusion matrix

5.5.2 AMD Staging (y_stage)

For AMD staging, EfficientNetV2-L achieved an overall accuracy of 0.8045 on 133 test images. The macro and weighted performance indicate improved balance across classes relative to some backbones:

- Macro F1-score: 0.7611
- Weighted F1-score: 0.8111

Per-class performance was:

- Early: precision 0.6667, recall 0.8966, F1-score 0.7647 (support = 29)
- Intermediate: precision 0.6071, recall 0.6538, F1-score 0.6296 (support = 26)
- Late: precision 0.9697, recall 0.8205, F1-score 0.8889 (support = 78)

These results show strong Late-stage discrimination and high Early-stage recall, while also improving Intermediate-stage precision/recall compared with backbones that tended to over-assign intermediate labels. Overall, the staging performance is competitive and relatively balanced at the macro level.
The confusion matrix shows:

- Early: 26 correct, 3 predicted as Intermediate, 0 as Late
- Intermediate: 17 correct, 7 predicted as Early, 2 predicted as Late
- Late: 64 correct, 8 predicted as Intermediate, 6 predicted as Early

The main errors remain concentrated around the Intermediate class boundaries and some Late → Early downgrading (6 cases), which is clinically plausible in borderline scans where advanced features may not be strongly expressed in a single B-scan.
Training curves for stagingshow a pronounced early reduction in validation loss and continued improvement until mid training, followed by mild oscillations. The validation curve generally remains below the training curve for a portion of training, which may reflect the regularization effect of augmentation during training and a deterministic validation pipeline.

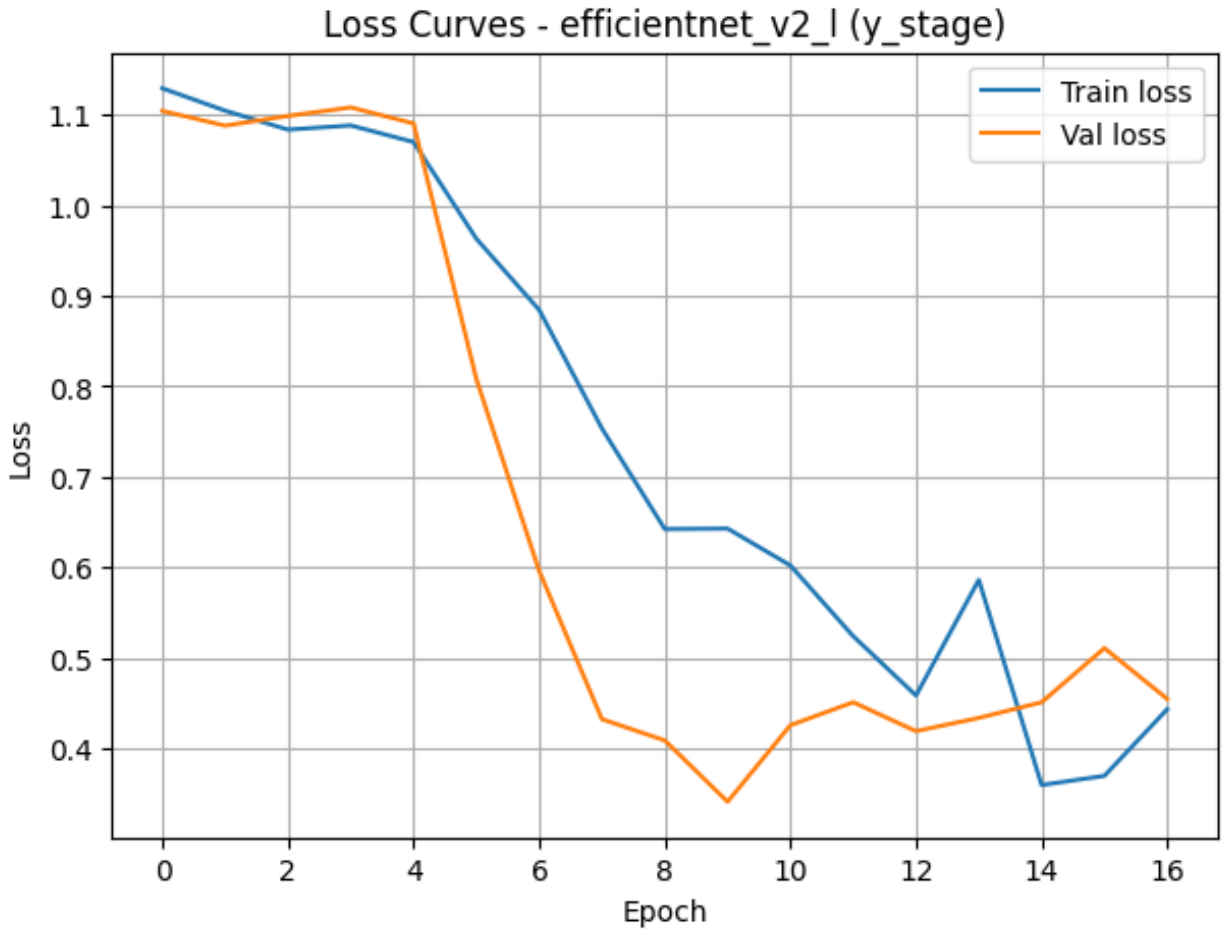

Figure 5.19: EfficientNetV2-L staging loss curves

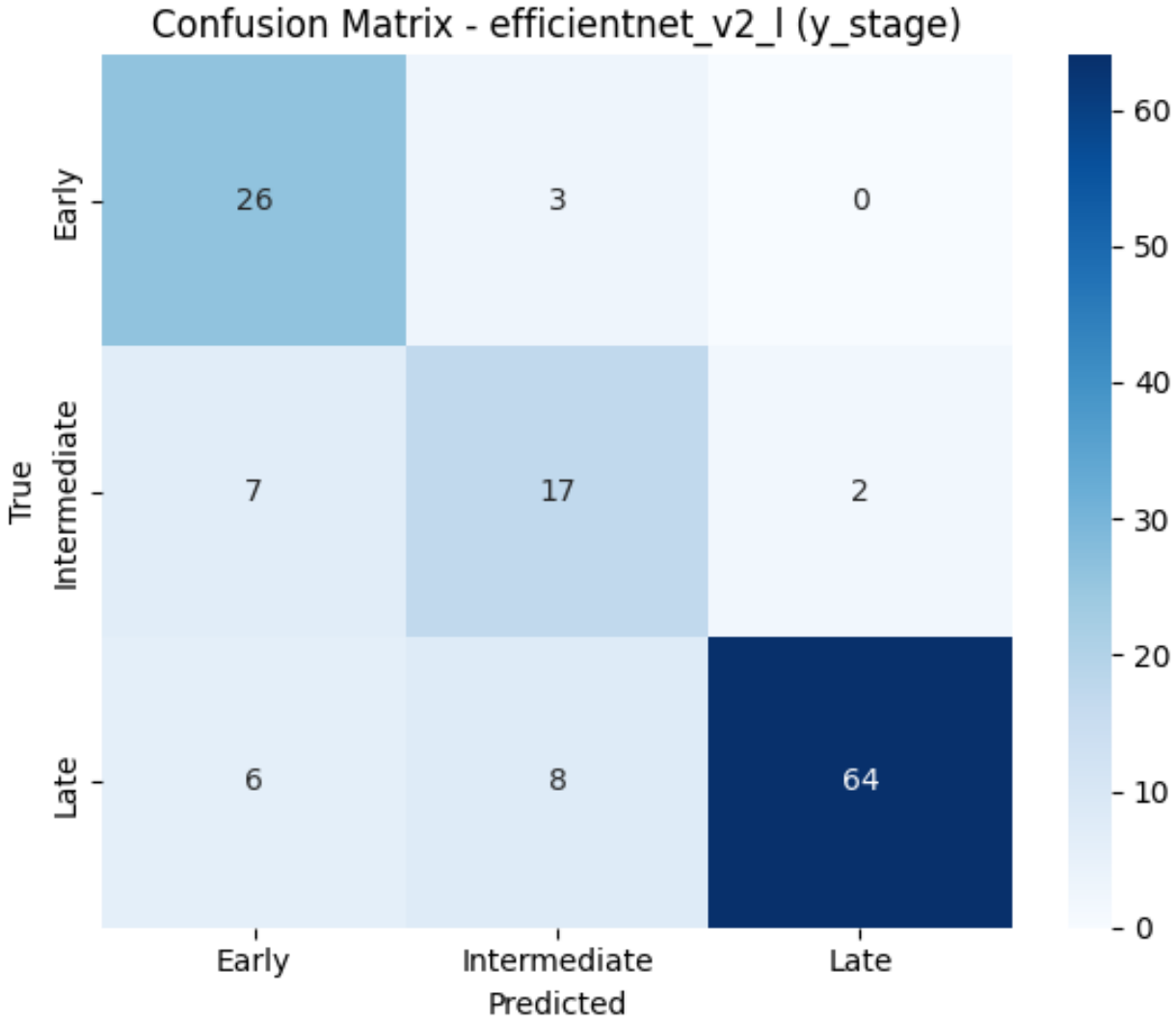

Figure 5.20: EfficientNetV2-L staging confusion matrix

5.5.3 Summary (EfficientNetV2-L)

EfficientNetV2-L delivers competitive staging performance (80.5% accuracy; macro F1 0.761), with relatively balanced recognition across Early, Intermediate, and Late stages. However, for binary detection, performance is slightly lower than the best-performing backbone, with an increased number of false negatives and a less stable validation-loss trajectory. Overall, EfficientNetV2-L is a strong candidate for staging, but ConvNeXt Small remains the most reliable across both tasks in the present experimental configuration.

5.6 Overall Cross-Model Comparison (Final Summary for Thesis)

This subsection consolidates the final test-set performance across all evaluated backbones for both tasks: (i) binary AMD detection (Not-AMD vs AMD) and (ii) AMD staging (Early vs Intermediate vs Late). All models were trained and evaluated

under the same unified pipeline described in Chapter 4 (consistent preprocessing, augmentation policy, two-phase freeze→fine-tune strategy, early stopping, and standardized test evaluation). This supports a fair, implementation-aligned comparison in which observed differences are primarily attributable to backbone capacity and representation quality rather than experimental inconsistencies.

	ResNet50	EfficientNet-B4	ConvNeXt-Tiny	ConvNeXt-Small	EfficientNetV2–L
Accuracy	0.975309	0.975309	0.967078	0.979424	0.967078
Sensitivity (Recall AMD)	0.962406	0.962406	0.962406	0.969925	0.954887
Specificity	0.990909	0.990909	0.972727	0.990909	0.981818
Precision (PPV AMD)	0.992248	0.992248	0.977099	0.992308	0.984496
F1 (AMD)	0.977099	0.977099	0.969697	0.980989	0.969466
Balanced Accuracy	0.976658	0.976658	0.967567	0.980417	0.968353
AUC	0.987560	0.984997	0.988448	0.986945	0.982365

Table 5.1: Binary detection performance comparison across models

model	resnet50	efficientnet_b4	convnext_tiny	convnext_small	efficientnetV2_l
TN	109	109	107	109	108
FP	1	1	3	1	2
FN	5	5	5	4	6
TP	128	128	128	129	127

Table 5.2: Binary detection performance comparison across models

Binary screening performance is uniformly high across all backbones, with accuracies in the range 0.967–0.979 and strong class separation as indicated by AUC values near 0.99. From a clinical screening perspective, the most important metric is typically AMD sensitivity (recall), because false negatives correspond to missed AMD cases. Specificity is also important because it determines false-positive referral burden. Best-performing model (Detection).

- ConvNeXt Small achieved the highest overall detection performance, with:
  - Accuracy: 0.9794
  - Sensitivity (Recall AMD): 0.9699
  - Specificity: 0.9909
  - AUC: 0.9869
  - Confusion counts: TN=109, FP=1, FN=4, TP=129

This model yields the lowest observed false negatives (FN=4) while maintaining extremely low false positives (FP=1), making it the strongest candidate for the first-stage screening component of the hierarchical pipeline.
AUC comparison (Detection). Although ConvNeXt Small provides the best accuracy and recall trade-off, the highest detection AUC was achieved by ConvNeXt Tiny (AUC = 0.9884), indicating excellent probability ranking capability despite slightly lower accuracy due to increased false positives. The full AUC values are:

- ResNet50: AUC = 0.9876
- EfficientNetB4: AUC = 0.9850
- ConvNeXt Tiny: AUC = 0.9884 (highest)
- ConvNeXt Small: AUC = 0.9869
- EfficientNetV2-L: AUC = 0.9824

Error-profile interpretation (Detection).

- ResNet50 and EfficientNetB4 have identical confusion outcomes (TN=109, FP=1, FN=5, TP=128), representing a strong and stable baseline.
- ConvNeXt Tiny increases false positives (FP=3), suggesting a slightly more "sensitive" decision boundary at the cost of additional false alarms.
- EfficientNetV2-L increases false negatives (FN=6), which is less desirable in screening because it corresponds to missed AMD.

5.6.1 AMD Staging (Early / Intermediate / Late): Comparative Results

	ResNet50	EfficientNet-B4	ConvNeXtTiny	ConvNeXtSmall	EfficientNetV2–L
Macro Precision	0.705713	0.700869	0.727679	0.778752	0.747835
Macro Recall	0.746242	0.748895	0.770262	0.773210	0.790304
Macro F1	0.713641	0.688156	0.740385	0.763560	0.761075
Weighted F1	0.779941	0.718819	0.793250	0.830640	0.811129
Balanced Accuracy	0.746242	0.748895	0.770262	0.773210	0.790304
Macro AUC (OvR)	0.920860	0.901088	0.924930	0.907297	0.938520
Recall – Early	0.931034	0.862069	0.862069	0.896552	0.896552
Test N (Early)	29	29	29	29	29
AUC Early (OvR)	0.960544	0.947944	0.955570	0.962865	0.953912
Recall – Intermediate	0.500000	0.769231	0.653846	0.500000	0.653846
Test N (Intermediate)	26	26	26	26	26
AUC Intermediate (OvR)	0.835370	0.801941	0.845794	0.783968	0.892883
Recall – Late	0.807692	0.615385	0.794872	0.923077	0.820513
Test N (Late)	78	78	78	78	78
AUC Late (OvR)	0.966667	0.953380	0.973427	0.975058	0.968765

Table 5.3: Multi-class classification performance comparison (AMD staging)

Staging is a more challenging task than detection due to: (i) subtle inter-stage differences, (ii) heavy class imbalance (Late dominates), and (iii) clinically realistic ambiguity, especially for Intermediate AMD. Therefore, macro-averaged metrics and OvR AUC are emphasized alongside accuracy.
Best-performing model (Staging accuracy and F1).

- ConvNeXt Small achieved the best overall staging performance in terms of hard-label classification:
  - Accuracy: 0.8346
  - Macro F1: 0.7636 (highest)
  - Weighted F1: 0.8306 (highest)
  - Balanced Accuracy: 0.7732
  - Macro AUC (OvR): 0.9073

This indicates the best overall trade-off between majority-class performance (Late) and minority-class recognition (Early/Intermediate), making it the most suitable model for the second stage of the pipeline when the objective is maximum overall staging correctness.
Best-performing model (Staging probability separation / AUC).

- EfficientNetV2-L achieved the highest macro one-vs-rest ROC-AUC for staging:
  - Macro AUC (OvR): 0.9385 (highest)
  - Balanced Accuracy: 0.7903 (highest)
  - Accuracy: 0.8045
  - Macro F1: 0.7611 (very close to ConvNeXt Small)

This suggests that EfficientNetV2-L provides the strongest class-wise probability ranking and separation, even though its discrete accuracy is slightly below ConvNeXt Small. In contexts where thresholding, calibration, or risk scoring is required (rather than only top-1 stage), this property can be particularly valuable.
AUC comparison (Staging macro OvR). The macro OvR AUC values across backbones are:

- ResNet50: 0.9209
- EfficientNetB4: 0.9011
- ConvNeXt Tiny: 0.9249
- ConvNeXt Small: 0.9073
- EfficientNetV2-L: 0.9385 (highest)

Stage-wise recall considerations (clinical emphasis).

- Early-stage recall (sensitivity for early AMD) is highest for ResNet50 (0.9310), and remains high for ConvNeXt Small and EfficientNetV2-L (both 0.8966). High Early recall is clinically relevant because it supports timely monitoring and intervention planning.
- Intermediate-stage recall remains the weakest and most variable across models, reflecting the transitional and overlapping nature of intermediate AMD. EfficientNetB4 achieves the highest Intermediate recall (0.7692) but at the cost of reduced Late recall and overall staging accuracy.
- Late-stage recall is highest for ConvNeXt Small (0.9231), contributing substantially to its top accuracy given Late-stage prevalence in the test set.

5.6.2 Final Backbone Selection for the Two-Stage Clinical Pipeline

The proposed deployment follows a hierarchical workflow: Detection → (if AMD) Staging. Therefore, a final backbone must satisfy two criteria: (i) reliable detection with minimal false negatives, and (ii) strong overall staging performance with acceptable minority-class sensitivity.

- Selected best overall model: ConvNeXt Small

- Best detection accuracy (0.9794) and best AMD recall (0.9699) with low false negatives (FN=4).
- Best staging accuracy (0.8346) and strongest F1 performance (Macro F1 0.7636; Weighted F1 0.8306).
- Strong Late-stage recall (0.9231) while maintaining high Early-stage recall (0.8966).

Accordingly, the qualitative interpretability analysis (Grad-CAM) is reported only for ConvNeXt Small as the final selected model, providing clinically oriented visual validation of the learned decision basis without redundant explanation panels for all backbones.

5.6.3 Interpretation of Tables 5.1–5.3

Detection (Table 5.1 and 5.2). All backbones achieved high screening performance (AUC 0.982–0.988). ConvNeXt Small produced the best overall screening performance with the highest accuracy (0.9794) and sensitivity (0.9699) while maintaining high specificity (0.9909). ConvNeXt Tiny achieved the highest detection AUC (0.9884), indicating excellent probability ranking despite slightly more false positives.
Staging (Table 5.3). AMD staging is more challenging and exhibits larger separation between models. ConvNeXt Small achieved the highest staging accuracy (0.8346) and the strongest macro/weighted F1 scores. EfficientNetV2-L achieved the highest macro OvR AUC (0.9385) and balanced accuracy (0.7903), indicating superior probability separation across stages. Intermediate-stage discrimination remains the primary difficulty across backbones.

5.7 Interpretability Analysis (Grad-CAM) — ConvNeXt Small (Best Model)

To complement quantitative evaluation, Grad-CAM was applied to the best-performing backbone (ConvNeXt Small) to provide qualitative evidence that staging predictions are driven by clinically plausible retinal structures. Grad-CAM produces a class-discriminative heatmap by propagating gradients from the predicted class score back to a selected convolutional feature layer, highlighting regions that most influenced the decision. In this study, Grad-CAM is reported only for the final selected model to avoid repetitive interpretability panels across all backbones.

(a) Correct Early-stage prediction In the Early-stage example (True=Early, Pred=Early) figure:5.21, the model's activation concentrates around localized regions within the central retina, consistent with subtle early-stage structural irregularities. The attention is largely confined to retinal tissue rather than background areas, supporting that the classifier is leveraging meaningful anatomical cues.

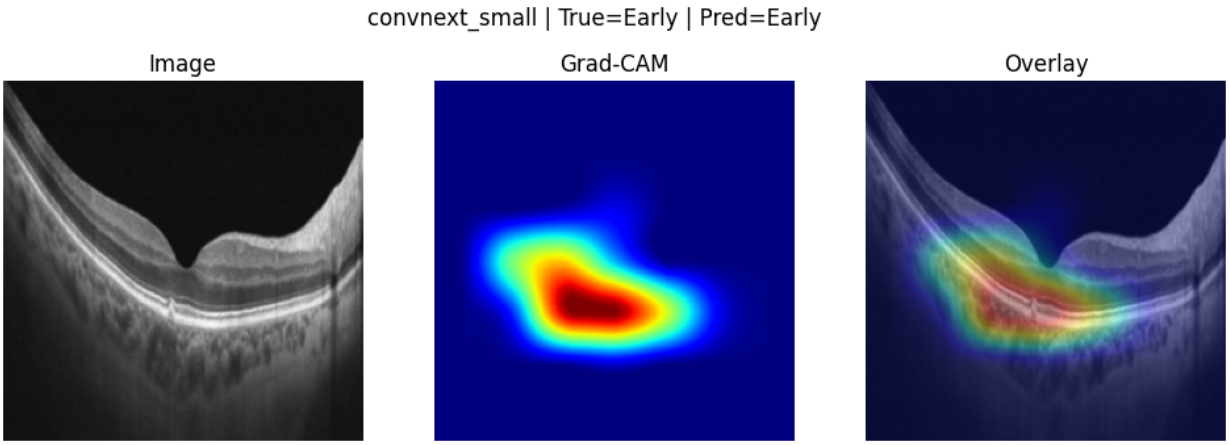

Figure 5.21: ConvNeXt Small Grad-CAM (Early; correct)

(b) Correct Intermediate-stage prediction Figure shows an Intermediate-stage example correctly classified (True=Intermediate, Pred=Intermediate) figure:5.22. The heatmap emphasizes focal retinal regions consistent with intermediate morphological changes, indicating that the network is capable of identifying discriminative patterns beyond the dominant Late class. Notably, the activation is concentrated within retinal structures rather than peripheral margins, suggesting reliance on anatomically relevant information.

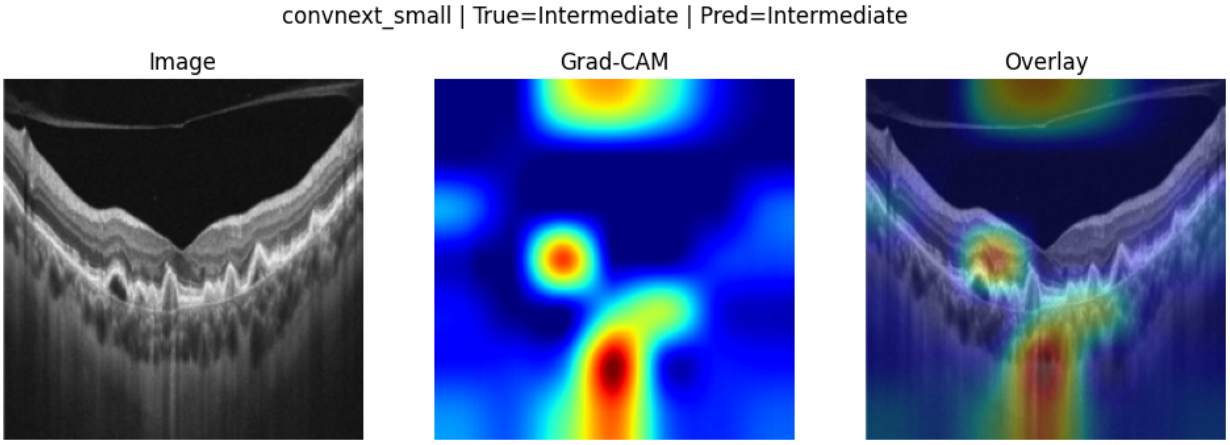

Figure 5.22: ConvNeXt Small Grad-CAM (Intermediate; correct)

(c) Correct Late-stage predictions (multiple examples) Multiple Late-stage examples (True=Late, Pred=Late) are shown in Figures figure:5.23, figure:5.24, figure:5.25, and also the previously reported Late examples in figure:5.26. Across these cases, Grad-CAM consistently highlights large, highly abnormal regions with pronounced structural disruption and lesion-like morphology. This aligns with clinical expectations for late AMD, where macular architecture is substantially distorted and therefore provides strong discriminative evidence.

A consistent pattern across Late examples is that the model's attention aligns with regions of visible deformation and abnormal reflectivity, reinforcing that ConvNeXt Small is not relying on irrelevant background or acquisition artifacts. This quali-

tative evidence is consistent with the strong Late-stage quantitative results for this backbone (high Late precision and recall).

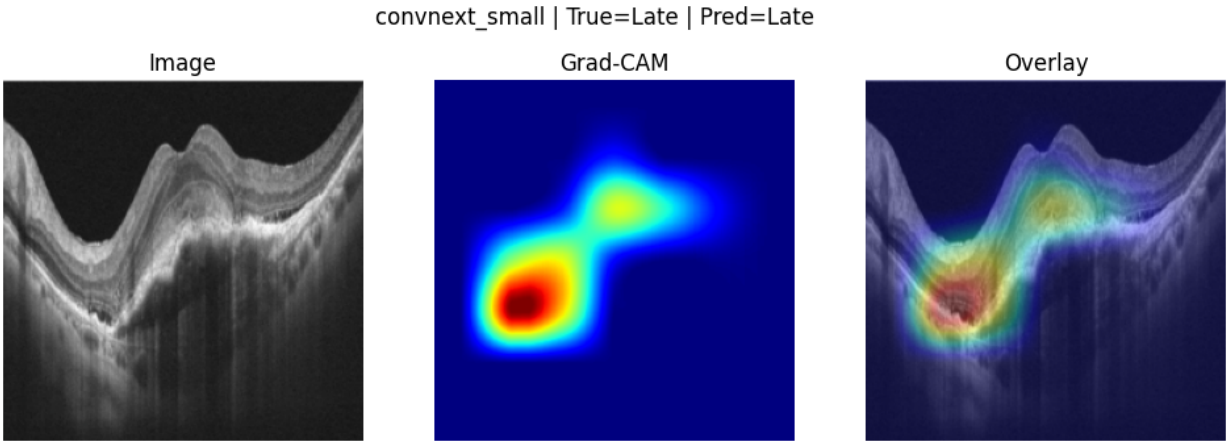

Figure 5.23: ConvNeXt Small Grad-CAM (Late; correct))

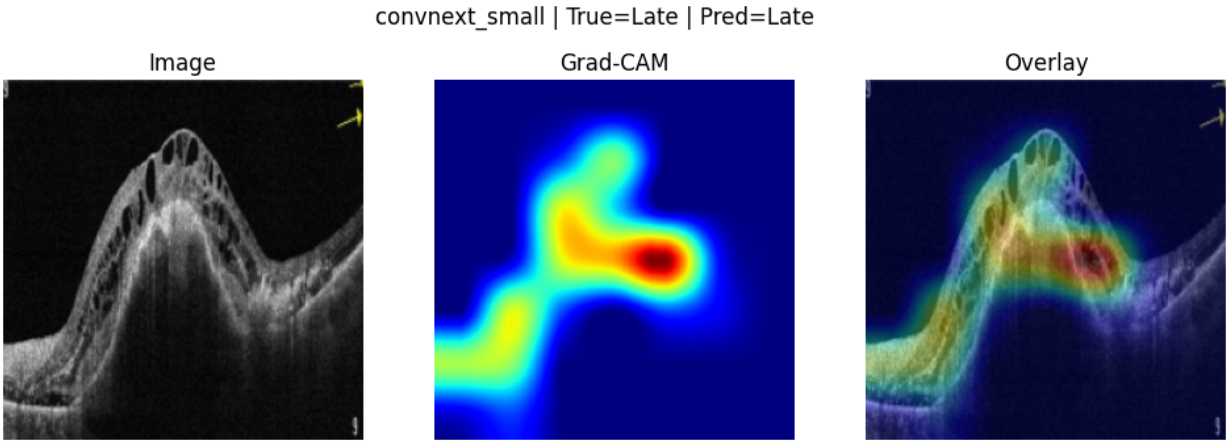

Figure 5.24: ConvNeXt Small Grad-CAM (Late; correct)

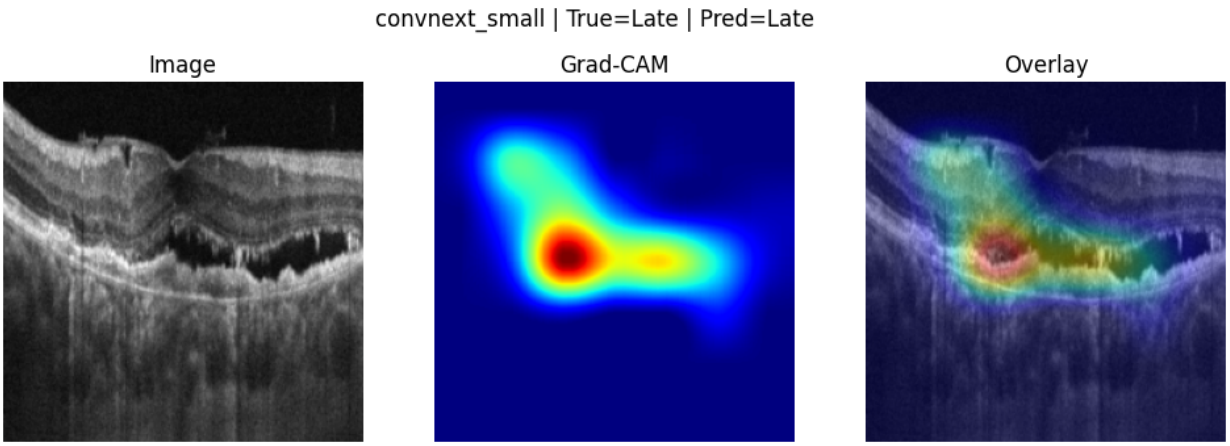

Figure 5.25: ConvNeXt Small Grad-CAM (Late; correct)

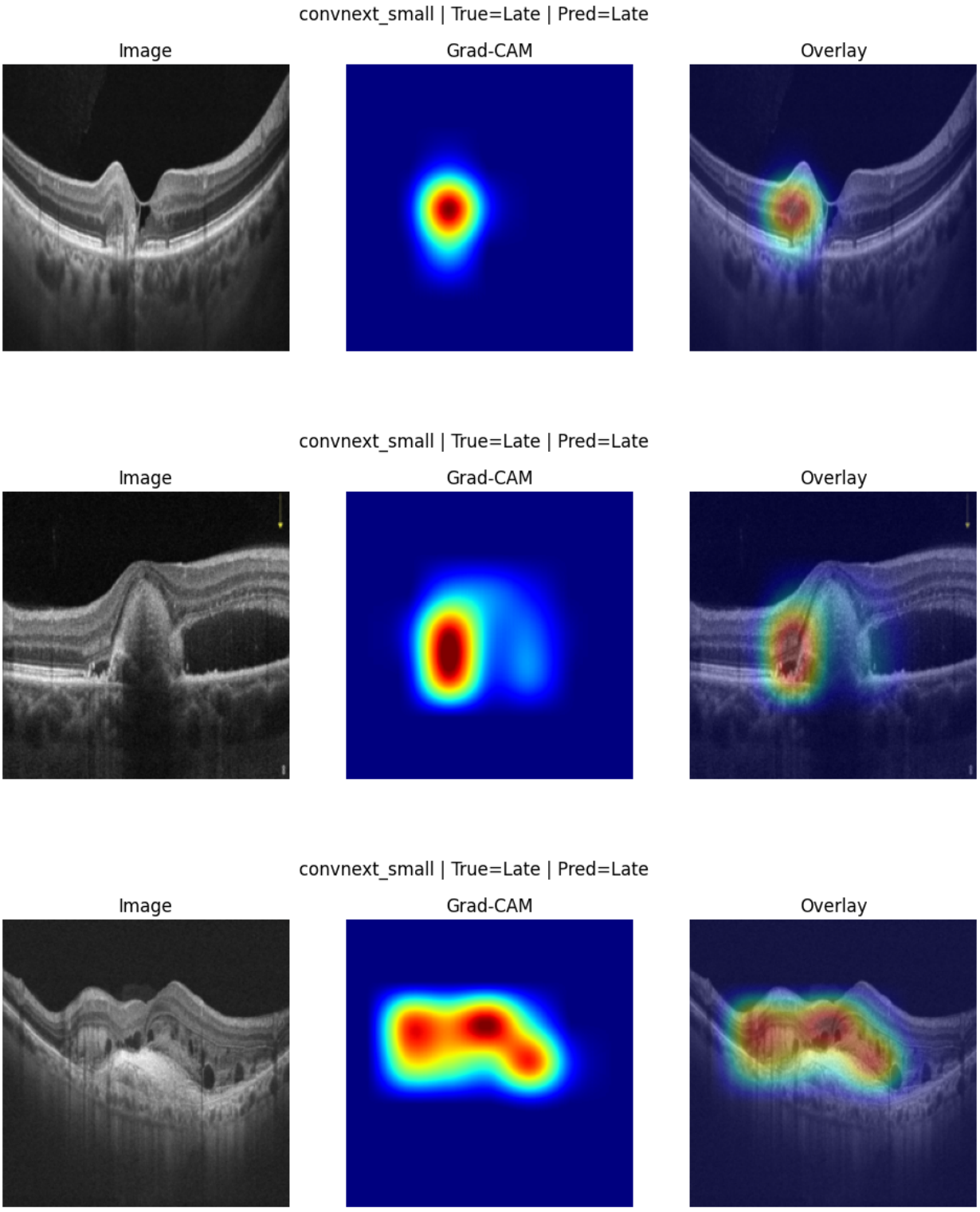

Figure 5.26: ConvNeXt Small Grad-CAM (Late; correct)

Chapter 6

Limitations and Future Work

This chapter discusses the main limitations of the proposed OCT-based deep learning framework for AMD detection and staging and outlines concrete future research directions to address them. The intent is to clarify the scope of validity of the reported findings and to identify methodological improvements required for clinical translation.

6.1 Limitations

6.1.1 Dataset scope and external validity

A key limitation is that the models were trained and evaluated on a single dataset source and under a fixed split protocol. While the achieved test performance is strong, it may not fully reflect real-world variability, including differences in OCT device vendors, scanning protocols, image quality, and patient demographics. Domain shifts of this kind can substantially reduce generalization in clinical deployment.

6.1.2 Class imbalance and intermediate-stage ambiguity

The staging task exhibits class imbalance, with Late AMD dominating the AMD subset. Although imbalance-aware strategies were applied, the Intermediate class remains the most challenging and is frequently confused with Early or Late stages. This limitation is also clinically plausible because intermediate AMD is a transitional phenotype and may not be separable using a single B-scan in borderline cases.

6.1.3 Limited contextual information (single B-scan input)

The framework operates on individual OCT B-scans, which restricts the available spatial and volumetric context. In clinical practice, graders often rely on multiple adjacent slices, en-face views, or full volumes to confirm subtle findings and distinguish stages. This limitation may reduce the robustness of staging, particularly for cases where biomarkers are not fully visible in a single slice.

6.1.4 Label noise and clinical ground truth constraints

The study relies on provided labels, which may contain ambiguity or inter-grader variability. AMD staging can be influenced by multiple biomarkers and clinical con-

ventions; therefore, label noise can limit achievable performance and inflate apparent confusion between adjacent stages.

6.1.5 Interpretability scope

Explainability analysis was conducted using Grad-CAM only for the best model (ConvNeXt Small). While this provides qualitative evidence that predictions are driven by relevant retinal regions, Grad-CAM is not a causal explanation and may vary with layer choice and image characteristics. Additional interpretability methods and systematic error-case inspection would strengthen clinical trust.

6.1.6 Computational and deployment considerations

Although ConvNeXt Small achieved the best overall performance, model size, inference latency, and hardware constraints were not comprehensively benchmarked across clinical deployment environments. Further engineering is required for reliable integration into real-time workflows, including optimization, monitoring, and quality-control mechanisms.

6.2 Future Work

6.2.1 Multi-domain evaluation and domain adaptation

Future work should validate the framework on external datasets and across multiple OCT vendors to quantify generalization under domain shift. When performance degradation occurs, domain adaptation strategies (e.g., feature alignment, test-time adaptation, and robust normalization) can be used to improve transfer.

6.2.2 Volumetric and multi-slice modeling

To better reflect clinical grading, the staging component can be extended to incorporate multi-slice context (2.5D approaches) or full 3D volumes. Sequence models (e.g., temporal pooling across slices) and 3D CNN architectures may improve detection of biomarkers that are sparse across individual slices.

6.2.3 Calibration, thresholding, and risk-aware outputs

For clinical decision support, probability calibration (temperature scaling, isotonic regression) and risk-aware thresholding are important. Future versions should report calibration metrics (e.g., ECE) and explore operating points that optimize sensitivity for screening and stability for staging.

6.2.4 Integration of segmentation or biomarker-guided learning

A clinically grounded extension is to incorporate retinal layer segmentation or lesion-aware supervision. Multi-task learning (classification + segmentation/biomarker

detection) can encourage the model to attend to relevant anatomy and improve interpretability.

6.2.5 Expanded explainability and clinician-in-the-loop validation

Future interpretability work should:

- include systematic Grad-CAM analysis for misclassified cases,
- compare multiple explainability methods (e.g., Integrated Gradients, LRP),
- involve clinician review to assess whether highlighted regions align with accepted biomarkers.

6.2.6 Prospective evaluation and clinical workflow integration

Ultimately, prospective testing is required to measure real-world benefit, including referral decisions, time savings, and performance under noisy acquisition conditions. Practical deployment would also require monitoring for dataset drift, uncertainty estimation, and clear failure-mode handling.

Chapter 7

Conclusion

7.1 Summary of Findings

In this thesis, a deep learning framework for binary detection and multi class staging of Age-Related Macular Degeneration (AMD) utilizing B-scan images from retinal optical coherence tomography (OCT) was demonstrated. The methodology was created in close conjunction with the implementation code and to reflect a clinically meaningful workflow using a two-stage hierarchical pipeline: (i) screening/detection (AMD vs. Non-AMD), subsequent to (ii) conditional staging (Early / Intermediate / Late) for scans identified as AMD.

A consistent transfer-learning methodology was utilized across different ImageNet-pretrained CNN backbones, with uniform head substitution and regularization to ensure fair comparison. All assessed models performed well on the detection challenge, according to experimental data, with high specificity and high AUC values that indicate dependable class separation. ConvNeXt Small demonstrated the most satisfactory overall balance between the two objectives among the researched backbones, attaining the maximum staging accuracy, F1 scores, and detection performance under the applied training procedure. The gradient-CAM visuals for ConvNeXt Small were assessed for interpretability, and the results showed that the model focused its attention on anatomically significant retinal areas, providing qualitative evidence for clinical credibility.

Despite these findings, staging in particular, differentiation of the intermediate class which remains difficult because of phase overlap, class imbalance, and the lack of contextual data found in specific B-scans. As a result, the thesis identifies specific areas for development, such as volumetric or multi-slice modeling, more robust awareness of imbalance goals for learning, assessment that is oriented toward calibration, and more extensive external validation across equipment and datasets.

Overall, the recommended method offers the basis for future research toward reliable, operational clinical decision-support systems and indicates that existing convolutional backbones might provide precise and clinically oriented AMD screening and staging from OCT B-scans.

7.2 Contributions to the Field

This research's contributions to the fields of ophthalmology and deep learning are focused on improving the development of a functional, effective, and clinically aligned

system for addressing Age-Related Macular Degeneration (AMD).

- **Implementation of a Hierarchical Clinical Pipeline:** This study developed a two-stage hierarchical workflow, which is opposed to many current models that are strictly limited to binary "AMD vs. Normal" identification. The method proceeds with a conditional multi-class staging (Early, Intermediate, or Late) after performing a binary screening to differentiate AMD from other retinal disorders. This concept is comparable to real-world clinical practice, where a diagnosis is verified prior to grading severity.
- **Identification of an Optimal Efficient Architecture:** The research offered an in-depth comparative evaluation of contemporary minimalist backbones such as ResNet50, EfficientNetB4, ConvNeXt Small, and ConvNeXt Tiny. It determined that ConvNeXt Small was the superior model, with the highest detection accuracy (97.94%) and the best overall staging performance (83.46%).
- **Enhanced Sensitivity for Early-Stage Detection :** A potentially significant gap in the field is that initial features are frequently too minor for successful automated identification.This research used sophisticated methods including transfer learning, weighted sampling, and MixUp/CutMix regularization to enhance the model's sensitivity (recall) for early-stage AMD. For example, the ResNet50 backbone has a high early-stage recall of 0.9310.
- **Emphasis on Deployment in Low-Resource Settings :** This research addressed the obstacle presented by resource demanding models requiring significant processing capability by concentrating on efficient structures. This system is scalable and efficient, making it ideal for use in distant or low-resource clinics.
- **Validation of Clinical Credibility with Explainable AI (XAI):** This research utilized Grad-CAM interpretability to qualitatively ensure that the model's judgments are influenced by anatomically significant indicators including drusen-associated structural disruption and retinal pigment epithelium (RPE) abnormalities. This contributes to the fundamental clinical acceptance for implementing AI-driven recommendation systems.
- **Robustness regarding Real-World Clinical Noise:** The models were trained and tested using the OCTDL dataset, consisting of real-world clinical noise and variability, particularly speckle noise and intensity inhomogeneity. The inclusion of Gaussian noise injection during training improved the model's resilience to these commonly encountered OCT aberrations.

Bibliography

[1] F. L. Ferris, M. D. Davis, T. E. Clemons, *et al.*, "A simplified severity scale for age-related macular degeneration: Areds report no. 18," *Archives of Ophthalmology*, vol. 123, no. 11, pp. 1570–1574, 2005. DOI: 10.1001/archopht.123.11.1570.

[2] F. L. Ferris, C. P. Wilkinson, A. Bird, *et al.*, "Clinical classification of age-related macular degeneration," *Ophthalmology*, vol. 120, no. 4, pp. 844–851, 2013. DOI: 10.1016/j.ophtha.2012.10.036. [Online]. Available: https://doi.org/10.1016/j.ophtha.2012.10.036.

[3] A.-R. E. D. S. 2. R. Group, "Lutein + zeaxanthin and omega-3 fatty acids for age-related macular degeneration: The age-related eye disease study 2 (areds2) randomized clinical trial," *JAMA*, vol. 309, no. 19, pp. 2005–2015, 2013, This is the primary reference for nutritional supplements in intermediate AMD; often cited as AREDS2. DOI: 10.1001/jama.2013.4412.

[4] R. Klein, S. M. Meuer, C. E. Myers, N. M. Bressler, B. E. K. Klein, and S. B. Bressler, "Harmonizing the classification of age-related macular degeneration in the three-continent AMD consortium," *Ophthalmic Epidemiology*, vol. 21, no. 1, pp. 14–23, 2014. DOI: 10.3109/09286586.2013.867512. [Online]. Available: https://doi.org/10.3109/09286586.2013.867512.

[5] W. L. Wong, X. Su, X. Li, *et al.*, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: A systematic review and meta-analysis," *The Lancet Global Health*, vol. 2, no. 2, e106–e116, 2014. DOI: 10.1016/S2214-109X(13)70145-1. [Online]. Available: https://doi.org/10.1016/S2214-109X(13)70145-1.

[6] Z. Yehoshua, P. J. Rosenfeld, G. Gregori, E. Cowen, and W. J. Feuer, "Spectral domain optical coherence tomography imaging of dry age-related macular degeneration," *Ophthalmic Surgery, Lasers and Imaging Retina*, vol. 45, no. 1, pp. 6–14, 2014. DOI: 10.3928/23258160-20140108-01. [Online]. Available: https://doi.org/10.3928/23258160-20140108-01.

[7] F. G. Holz, E. C. Strauss, S. Schmitz-Valckenberg, and M. van Lookeren Campagne, "Geographic atrophy: Clinical features and potential therapeutic approaches," *Ophthalmology*, vol. 124, no. 6, Supplement, S1–S10, 2017. DOI: 10.1016/j.ophtha.2017.01.023. [Online]. Available: https://doi.org/10.1016/j.ophtha.2017.01.023.

[8] K. Alsaih, M. Z. Yusoff, T. B. Tang, I. Faye, and F. Mériaudeau, "Deep learning architectures analysis for age-related macular degeneration segmentation on optical coherence tomography scans," *Computer Methods and Programs in*

Biomedicine, vol. 189, p. 105 566, 2020. DOI: 10.1016/j.cmpb.2020.105566. [Online]. Available: https://doi.org/10.1016/j.cmpb.2020.105566.

[9] S. R. Sadda, R. Guymer, F. G. Holz, *et al.*, "Consensus definition for atrophy associated with age-related macular degeneration on OCT: Classification of atrophy report 5," *Ophthalmology*, vol. 127, no. 3, pp. 394–409, 2020. DOI: 10.1016/j.ophtha.2019.09.028. [Online]. Available: https://doi.org/10.1016/j.ophtha.2019.09.028.

[10] R. F. Spaide, G. J. Jaffe, D. Sarraf, *et al.*, "Consensus nomenclature for reporting neovascular age-related macular degeneration data: Consensus on neovascular age-related macular degeneration nomenclature study group," *Ophthalmology*, vol. 127, no. 5, pp. 616–636, 2020. DOI: 10.1016/j.ophtha.2019.11.036. [Online]. Available: https://doi.org/10.1016/j.ophtha.2019.11.036.

[11] E. Vaghefi, S. Hill, H. M. Kersten, and D. Squirrell, "Multimodal retinal image analysis via deep learning for the diagnosis of intermediate dry age-related macular degeneration: A feasibility study," *Journal of Ophthalmology*, vol. 2020, no. 1, p. 7 493 419, 2020.

[12] A. Arrigo, A. Bordato, E. Aragona, *et al.*, "Macular neovascularization in AMD, CSC and best vitelliform macular dystrophy: Quantitative OCTA detects distinct clinical entities," *Eye*, vol. 35, no. 12, pp. 3266–3276, 2021. DOI: 10.1038/s41433-021-01396-2. [Online]. Available: https://doi.org/10.1038/s41433-021-01396-2.

[13] E. Y. Chew and M. L. Klein, "Nonexudative macular neovascularization in age-related macular degeneration," *Retinal Physician*, Jun. 2021. [Online]. Available: https://www.retinalphysician.com/issues/2021/june/nonexudative-macular-neovascularization-in-age-related-macular-degeneration.

[14] S. Sotoudeh-Paima, A. Jodeiri, F. Hajizadeh, and H. Soltanian-Zadeh, "Multi-scale convolutional neural network for automated amd classification using retinal oct images," *arXiv preprint arXiv:2110.03002*, 2022. DOI: 10.48550/arXiv.2110.03002.

[15] J. Wu and T. Y. A. Liu, "Application of deep learning to retinal-image-based oculomics for evaluation of systemic health: A review," *Journal of Clinical Medicine*, vol. 12, no. 1, p. 152, 2022.

[16] U. Chakravarthy, C. Bailey, and V. Chong, "Towards a better understanding of non-exudative choroidal and macular neovascularization," *Progress in Retinal and Eye Research*, vol. 92, p. 101 113, 2023. DOI: 10.1016/j.preteyeres.2022.101113. [Online]. Available: https://doi.org/10.1016/j.preteyeres.2022.101113.

[17] O. Leingang, S. Riedl, J. Mai, *et al.*, "Automated deep learning-based AMD detection and staging in real-world OCT datasets (PINNACLE study report 5)," *Scientific Reports*, vol. 13, no. 1, p. 19 545, 2023. DOI: 10.1038/s41598-023-46626-7. [Online]. Available: https://doi.org/10.1038/s41598-023-46626-7.

[18] X. Leng, R. Shi, Y. Wu, *et al.*, "Deep learning for detection of age-related macular degeneration: A systematic review and meta-analysis of diagnostic test accuracy studies," *PLOS ONE*, vol. 18, no. 4, e0284060, 2023. DOI: 10.1371/journal.pone.0284060. [Online]. Available: https://doi.org/10.1371/journal.pone.0284060.

[19] B. Liefers, P. Taylor, A. Alsaafin, *et al.*, "Automated deep learning-based AMD detection and staging in real-world OCT datasets (PINNACLE study report 5)," *Scientific Reports*, vol. 13, p. 19 545, 2023. DOI: 10.1038/s41598-023-46626-7. [Online]. Available: https://doi.org/10.1038/s41598-023-46626-7.

[20] M. Schranz, R. Told, V. Hacker, *et al.*, "Correlation of vascular and fluid-related parameters in neovascular age-related macular degeneration using deep learning," *Acta Ophthalmologica*, vol. 101, no. 1, e95–e105, 2023, Published online 2022 Aug 1. DOI: 10.1111/aos.15219. [Online]. Available: https://doi.org/10.1111/aos.15219.

[21] A. Alenezi, H. Alhamad, A. Brindhaban, Y. Amizadeh, A. Jodeiri, and S. Danishvar, "Enhancing readability and detection of age-related macular degeneration using optical coherence tomography imaging: An AI approach," *Bioengineering*, vol. 11, no. 4, p. 300, 2024. DOI: 10.3390/bioengineering11040300. [Online]. Available: https://doi.org/10.3390/bioengineering11040300.

[22] C. R. Clemens, N. Eter, and F. Alten, "Current perspectives on type 3 macular neovascularization due to age-related macular degeneration," *Ophthalmologica*, vol. 247, no. 2, pp. 73–84, 2024. DOI: 10.1159/000536278. [Online]. Available: https://doi.org/10.1159/000536278.

[23] Y. Gao, Z. Li, Y. Sun, Y. Wang, and X. Zhang, "Recent advances in the application of artificial intelligence in age-related macular degeneration," *BMJ Open Ophthalmology*, vol. 9, no. 1, e001903, 2024, PMC11580293. DOI: 10.1136/bmjophth-2024-001903. [Online]. Available: https://doi.org/10.1136/bmjophth-2024-001903.

[24] L. Hamid, A. Elnokrashy, E. H. Abdelhay, and M. M. Abdelsalam, "A deep learning lstm-based approach for amd classification using oct images," *Neural Computing and Applications*, vol. 36, pp. 19 531–19 547, 2024. DOI: 10.1007/s00521-024-10149-7.

[25] M. Kulyabin, A. Zhdanov, A. Nikiforova, *et al.*, "Octdl: Optical coherence tomography dataset for image-based deep learning methods," *Scientific Data*, vol. 11, no. 1, p. 365, Apr. 2024, Data Descriptor. Dataset available at https://doi.org/10.17632/sncdhf53xc. DOI: 10.1038/s41597-024-03182-7. [Online]. Available: https://doi.org/10.1038/s41597-024-03182-7.

[26] A. Miladinović, A. Biscontin, M. Ajčević, and A. Accardo, "Evaluating deep learning models for classifying OCT images with limited data and noisy labels," *Scientific Reports*, vol. 14, p. 30 321, 2024. DOI: 10.1038/s41598-024-81127-1. [Online]. Available: https://doi.org/10.1038/s41598-024-81127-1.

[27] S. Agarwal, A. Kumar, and R. Sharma, "HDL-ACO hybrid deep learning and ant colony optimization for ocular optical coherence tomography image classification," *Scientific Reports*, vol. 15, p. 5888, 2025. DOI: 10.1038/s41598-025-89961-7. [Online]. Available: https://doi.org/10.1038/s41598-025-89961-7.

[28] GBD 2021 Global AMD Collaborators, "Global burden of vision impairment due to age-related macular degeneration, 1990–2021, with forecasts to 2050: A systematic analysis for the global burden of disease study 2021," *The Lancet Global Health*, vol. 13, no. 7, e1175–e1190, 2025. DOI: 10.1016/S2214-109X(25)00143-3. [Online]. Available: https://doi.org/10.1016/S2214-109X(25)00143-3.

[29] A. Gupta and R. Singh, *Improving diagnostic performance on small and imbalanced datasets using class-based input image composition*, arXiv preprint arXiv:2511.03891, 2025. arXiv: 2511.03891 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2511.03891.

[30] A. Hlali, M. Ben Yakhlef, and S. El Hazzat, *Improving diagnostic performance on small and imbalanced datasets using class-based input image composition*, arXiv preprint arXiv:2511.03891, 2025. DOI: 10.48550/arXiv.2511.03891. arXiv: 2511.03891 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2511.03891.

[31] G. Neri, C. Rebecchi, J. D. Oakley, *et al.*, "Deep learning model for automated classification of macular neovascularization subtypes in AMD," *Investigative Ophthalmology & Visual Science*, vol. 66, no. 9, p. 55, 2025. DOI: 10.1167/iovs.66.9.55. [Online]. Available: https://doi.org/10.1167/iovs.66.9.55.

[32] P. Zhang, L. Wang, and J. Li, "A robust deep learning classifier for screening multiple retinal diseases on optical coherence tomography," *Scientific Reports*, vol. 15, p. 35 334, 2025. DOI: 10.1038/s41598-025-19286-y. [Online]. Available: https://doi.org/10.1038/s41598-025-19286-y.