UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Exploring the Diagnostic and Prognostic Potential of OCT Data in Multiple Sclerosis Using Machine Learning Techniques

Tom Nangosyah

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Bioinformatics

SUPERVISOR :
Prof. dr. Dirk VALKENBORG

CO-SUPERVISOR :
dr. Axel-Jan ROUSSEAU

2024
2025

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Exploring the Diagnostic and Prognostic Potential of OCT Data in Multiple Sclerosis Using Machine Learning Techniques

Tom Nangosyah

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Bioinformatics

SUPERVISOR :
Prof. dr. Dirk VALKENBORG

CO-SUPERVISOR :
dr. Axel-Jan ROUSSEAU

Acknowledgment

I extend my heartfelt gratitude to those whose remarkable contributions have been instrumental in the completion of this thesis.

First and foremost, I am deeply thankful to the Lord Almighty for providing me with strength, spiritual guidance, and abundant blessings throughout this journey.

I express my sincere appreciation to Prof.Dr. Dirk VALKENBORG and Dr. Axel-Jan ROUSSEAU for their guidance, patience, expertise. Their unwavering support and insightful guidance have played a vital role in shaping my research and academic growth.

I am grateful for the Ugandan community in Hasselt, whose collective experiences and supportive spirit have significantly enhanced my experience during my masters, making my stay enjoyable and contributing to my personal growth. Their presence has created a nurturing environment that I deeply appreciate.

To the friends I made here in Belgium, your companionship holds significance beyond measure in many ways you may not fully realize. I also deeply appreciate the unwavering support from friends in Uganda and all my masters classmates, with whom I have had the privilege to share classes, work together, enjoy moments, and learn from.

Lastly, I dedicate this work to my family, who have provided moral support for all my life decisions. I am especially grateful to my brother Julius and his wife Meryl, who have taken care of me, opened their household to me, and given me all the encouragement and support along this journey. A special shout out goes to my little niece Alexia, whose company, playfulness and smiles help me escape the hardships of life.

To my parents, siblings and family, their sacrifices, love, and belief in me have been the foundation upon which I have built my life and I am profoundly grateful for their unconditional support and the love that has been my anchor throughout this journey.

Abstract

Multiple Sclerosis (MS) is a chronic neurological disorder with significant diagnostic and prognostic challenges. Optical Coherence Tomography (OCT) has emerged as a valuable tool for assessing retinal changes in MS patients, offering insights into neuro-degeneration. However, variability in OCT measurement techniques and the lack of standardization limit its full potential. Combining OCT data with machine learning (ML) offers an innovative approach for improving diagnostic accuracy and understanding disease mechanisms.

This study had two objectives: (1) to develop and evaluate machine learning models for analyzing OCT data to classify MS severity and (2) to review OCT measurement techniques, focusing on retinal structures, variability across manufacturers, and segmentation methods.

A retrospective analysis was conducted on OCT data from 230 MS patients, derived from an initial cohort of 740 after addressing missing EDSS scores and other pre-processing steps. ML models, including Random Forest (RF), Support Vector Machine (SVM), XGBoost, and k-Nearest Neighbors (KNN), were employed to classify patients into Non-Severe and Severe categories based on EDSS thresholds. Key features were identified using LASSO. The study also reviewed segmentation methods and measurement variability across different OCT devices currently used by ophthalmologists, in comparison with the Canon OCT-HS100 system.

Overall, the Random Forest model emerged as the most robust classifier, achieving an F1-Score of 0.74 for the left eye and 0.72 for the right eye, with high precision and balanced recall across both datasets. SVM and XGBoost also performed well, with F1-Scores of 0.73 and 0.72, respectively, highlighting their potential for OCT-based MS severity classification. KNN showed consistent performance, particularly for the right-eye dataset, achieving an F1-Score of 0.70.

While the overall performance was strong, it was observed that the models struggled to accurately predict the the severe cases due to data imbalance in this category. This imbalance led to lower precision and recall for the severe category.

The Feature importance analysis identified critical predictors of MS severity, including Superior and Temporal sectors of the retina, Central ILM-RPE, retinal asymmetry metrics, and Nasal regions of the nerve fiber layer, ganglion Cell layer and inner pleciform layer. These findings align with existing evidence linking retinal thinning and asymmetry to MS-related neuro-degeneration.

The literature review highlighted significant variability in segmentation methods and device-specific metrics. Proprietary OCT devices, including Canon, Cirrus HD-OCT, and Spectralis, offer unique strengths but differ in scan patterns, resolution, and normative databases. This shows the need for standardization and the development of adaptable tools to address device-specific challenges

This study demonstrates the utility of ML models, particularly Random Forest, for classifying MS severity based on OCT data and emphasizes the importance of retinal structural features in predicting disease progression. Future work should focus on expanding datasets, refining segmentation methods, and addressing standardization to enhance the clinical applicability of OCT-based ML models in MS management.

Contents

1 Introduction **1**
- 1.1 Description of Research . . . 2
- 1.2 Study Objectives . . . 2

2 Literature Review **3**
- 2.1 Segmentation Algorithms . . . 4
- 2.2 OCT Device Comparison . . . 6
- 2.3 Advancements in OCT Analysis . . . 8
- 2.4 Related Works Using OCT Data for MS Diagnosis . . . 9

3 Data **10**
- 3.1 Dataset . . . 10
- 3.2 Data Extraction and Preparation . . . 11
- 3.3 Study Variables . . . 11
- 3.4 OCT Protocols . . . 13
- 3.5 Exploratory Data Analysis . . . 14
- 3.6 Software . . . 14

4 Methodology **15**
- 4.1 Support Vector Machine (SVM) . . . 15
- 4.2 Random Forest . . . 15
- 4.3 K-Nearest Neighbors (KNN) . . . 16
- 4.4 XGBoost . . . 16
- 4.5 Hyperparameter Tuning . . . 17
- 4.6 Model Assessment . . . 17
- 4.7 Machine Learning Pipeline . . . 19

5 Results **21**
- 5.1 Feature Engineering . . . 21
- 5.2 MS Diagnostic Model . . . 22
- 5.3 Model Performance . . . 24
- 5.4 Feature Importance . . . 29

6 Discussion **32**

7 Possible Drawback of the Methods **34**

8 Ethics, Societal Relevance, and Stakeholder Awareness **34**

9 Conclusion **35**

10 Ideas for Future Research **35**

List of Tables

1 Comparison of Optic Disc Measurements Across OCT Devices. **Source:** Compiled from manufacturer specifications, published research papers, and independent analysis. 7

2 Comparison of RNFL and macular measurements across OCT devices, including Cirrus HD-OCT, Spectralis OCT, and Canon HS100 OCT. **Source:** Compiled from manufacturer specifications, published research papers, and independent analysis. 7

3 Comparison of MS studies using OCT data and machine learning 10

4 Summary of Variables 12

5 Demographic and clinical characteristics of the study population. 14

6 Summary of Engineered Features Derived from OCT Data 22

7 Performance metrics for the models evaluated on the left and right eye datasets. 24

8 Performance metrics for the models evaluated on the left and right eye datasets for the severe class 29

9 Top predictive features identified by the models using the left eye dataset. 30

10 Top predictive features identified by the models using the right eye dataset. . . . 31

List of Figures

1 OCT Analysis Sectors (García Mesa et al., 2023) 13
2 Machine learning Pipeline. 20
3 Confusion Matrices for Models on Left Eye Datasets 26
4 Confusion Matrices for Models on Right Eye Datasets 27
5 ROC Curves for Models on Left and Right Eye Datasets 28
6 Visualization of retinal regions and their corresponding OCT scans: (a) Background image of the eye highlighting the macula and optic disc, (b) Optical Coherence Tomography (OCT) scan of the macula, (c) OCT scan of the peripapillary region. Key layers and structures: **mRNFL** – macular retinal nerve fiber layer, **GCL** – ganglion cell layer, **IPL** – inner plexiform layer, **INL** – inner nuclear layer, **ONL** – outer nuclear layer, **BM** – Bruch's membrane, **pRNFL** – peripapillary retinal nerve fiber layer, **PT** – prelaminar tissue, **LC** – lamina cribrosa, **Source:** Olbert and Struhal, 2022 40

1 Introduction

Multiple Sclerosis is an autoimmune disease that affects the central nervous system and, when left untreated, can cause neuro-degeneration, leading to severe disabilities such as partial or total blindness, sensory loss, and other impairments. Although relatively uncommon, it occurs most frequently in individuals aged 20 to 40, often resulting in long-term disability (Goodin, 2014). Early diagnosis of Multiple Sclerosis is essential in helping patients with the disease make better decisions about their future concerning issues such as family planning, lifelong healthcare, and maintaining employment (Ramagopalan and Sadovnick, 2011). This makes it essential to gain insights into the underlying causes of the disease and the principal mechanisms of disease evolution that would help in finding a cure or building therapies for patients, as well as ways of primary prevention of future disease.

The current diagnostic methods for multiple Sclerosis are extensively explained in the 2017 McDonald Criteria, with the latest revisions placing emphasis on increased sensitivity to allow earlier diagnosis (Thompson et al., 2018). The introduction over the years of diagnostic tools and methods such as Magnetic Resonance Imaging (MRI) and neuro-imaging have become important in the monitoring and detection of Multiple Sclerosis and are usually the tools for choice in the clinical setting (Wattjes et al., 2021). However, these alone might not be as effective in the diagnosis of the disease, given multiple sclerosis manifestations make following the disease over time challenging and require the incorporation of various clinical, laboratory, and radiological data. Access to this type of data requires advanced models to analyze and interpret it effectively. The data includes long-term neuro-performance measures, blood and cerebrospinal fluid (CSF) biomarkers, imaging results, electrodiagnostic data, patient-reported outcomes, and optical coherence tomography (OCT). In order to achieve these models, Artificial Intelligence offers an avenue for this kind of modeling to derive insights from these data.

Optical Coherence Tomography (OCT) is one of the widely used rapidly developing medical imaging technologies; it serves as a non-invasive imaging tool that provides high-resolution cross-sectional images of the retina; it was first proposed by Huang et al., 1991. The relatively low resolution of the first OCT devices has been gradually improved so that the image quality is now able to resolve more subtle changes in retinal layers. Numerous studies have shown that OCT can be used in monitoring and confirming many common and sight-threatening ocular conditions, such as glaucoma (Geevarghese et al., 2021), diabetic retinopathy (Amoaku et al., 2020), and age-related macular degeneration (Flores et al., 2021).

OCT, being able to provide high-resolution images of the retina, offers a unique window into neurodegeneration because it reflects changes in both the optic nerve and brain pathways. Neurodegeneration and inflammation in MS cause damage to the optic nerves, which leads to the thinning of the retinal layers. OCT can detect and quantify this thinning by measuring the retinal nerve fiber layer (RNFL) and ganglion cell-inner plexiform layer (GCIPL). These layers are particularly sensitive to damage in MS patients, where axonal and neuronal loss are common. The thickness of these retinal layers, as assessed through OCT, has been correlated with disability progression and cognitive impairment in MS patients.

Studies have demonstrated that RNFL and GCIPL thinning, as captured by OCT, are linked with greater physical disability and cognitive decline in MS, particularly in those with progressive forms of the disease (Shi et al., 2019). This association shows OCT's potential as a marker for

disease activity and prognosis.

Machine Learning (ML) has transformed healthcare, particularly in medical imaging, where its ability to process and analyze huge datasets is improving diagnostic and prognostic practices (Ranschaert et al., 2019). ML algorithms, capable of learning from data, help in uncovering intricate patterns that are not immediately visible to the human eye. In the diagnosis of Multiple Sclerosis (MS), these technologies offer the potential to enhance understanding, prediction, and the development of medicines.

Despite the significant progress in understanding and managing Multiple Sclerosis (MS), current diagnostic and prognostic methods using Optical Coherence Tomography (OCT) remain limited by their reliance on traditional statistical analyses and visual inspection by clinicians (Cavaliere et al., 2019). Several studies have employed parameters from OCT to train machine learning algorithms for diagnosis, but not all relevant features have been fully identified, and most studies do not use a wide range of features extracted from the OCT images (Rothman et al., 2019). The current practice of visually inspecting OCT scans also lacks objectivity and scalability, especially in large datasets. Moreover, there are few large-scale longitudinal studies that have thoroughly explored how OCT data can be used in clinical decision-making and its impact on predicting MS progression.

This study focuses on evaluating different machine learning models for the classification of patients with Multiple Sclerosis (MS) based on a collection of features derived from an Optical Coherence Tomography (OCT) imaging device. These features include measurements across various anatomical regions, such as sectoral thicknesses (Four and Twelve Sectors of the optic disc), optic nerve head (ONH) parameters, retinal nerve fiber layer (RNFL) metrics, ganglion cell layer with inner plexiform layer (GCL+IPL), combined nerve fiber layer, ganglion cell layer, and inner plexiform layer (NFL+GCL+IPL), as well as thickness and volume parameters between the inner limiting membrane (ILM) and retinal pigment epithelium (RPE) or Bruch's membrane (BM).

1.1 Description of Research

This study utilizes a comprehensive OCT dataset from MS Centrum Pelt, consisting of eye scans from 230 MS patients for both left and right eye. The primary aim is to analyze retinal and optic nerve parameters to gain insights into Multiple Sclerosis (MS) progression. By combining patient data with OCT-derived features, the research focuses on building machine learning models to enhance the diagnostic potential of OCT in MS. Given the subjective nature of current OCT interpretations by ophthalmologists, this study seeks to develop objective, machine learning driven approaches that can help in the standardizing and supporting decision-making in MS diagnosis.

1.2 Study Objectives

This study has two main objectives:

- To develop machine learning models for analyzing OCT data to improve diagnostic accuracy in MS patients.
- To conduct a literature review on OCT measurement techniques, focusing on retinal struc-

tures and variations across manufacturers. It also involves exploring existing segmentation methods for OCT images.

The thesis is organized as follows: Section 2 provides a comprehensive overview of the state of Optical Coherence Tomography (OCT) and its applications in the context of Multiple Sclerosis (MS), including a review of existing literature on OCT measurement techniques and segmentation methods. Section 3 describes the dataset used in this study, detailing data extraction, preparation, and exploratory analyses, as well as the OCT protocols followed and the study variables considered. Section 4 outlines the machine learning methods employed, including Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Gradient Boosting, along with the pipeline for model development and assessment. Section 5 presents the results of the study, including feature engineering, model performance metrics, feature importance analysis, and the clinical relevance of the findings. Section 8 discusses the ethical considerations, societal relevance, and stakeholder awareness associated with the research. Section 6 interprets the results in detail, highlighting key insights and addressing methodological challenges. Section 7 identifies potential drawbacks of the methods used, while Section 9 summarizes the conclusions of the study. Finally, Section 10 proposes ideas for future research, building on the study's findings and addressing identified limitations.

2 Literature Review

Optical Coherence Tomography (OCT) is a non-invasive imaging technique that generates high-resolution cross-sectional images of the retina (Clinic, 2024). The technique is based on the principle of low-coherence interferometry, where the echo time delay and intensity of back-scattered light from retinal tissue layers are measured and analyzed. By interpreting the resulting interference patterns, OCT systems reconstruct detailed cross-sectional views of the retina (Zeppieri et al., 2023). The resolution of these images depends on the bandwidth of the light source, with broader bandwidths delivering superior axial resolution. Different manufacturers incorporate various light source technologies, impacting both the resolution and depth of tissue penetration. Among the available OCT technologies, spectral-domain OCT (SD-OCT) is particularly stands out for its ability to acquire entire depth profiles in a single scan, offering faster and more efficient imaging compared to older time-domain OCT systems (Liu et al., 2014).

By producing detailed maps of retinal layers, OCT serves as an essential tool for diagnosing and monitoring neurological conditions such as multiple sclerosis (MS) (El Ayoubi et al., 2024). Key OCT-derived metrics, including Retinal Nerve Fiber Layer (RNFL) thickness, Ganglion Cell Layer + Inner Plexiform Layer (GCL+IPL) thickness, and macular volume, are extensively studied in the context of MS (Schneider et al., 2013) where the RNFL thinning correlates with axonal loss, while reductions in GCL+IPL thickness reflect neuro-degeneration, making both metrics critical indicators of MS severity and progression. Furthermore, changes in macular volume are associated with retinal atrophy, further linking retinal alterations to central nervous system damage in MS (Kaushik and Fraser, 2020). The ability of OCT to provide comprehensive insights into retinal structure highlights its critical role in advancing the understanding and management of MS. To enhance clarity, in the Appendix, Figure 6 illustrates the Retinal Nerve Fiber Layer (RNFL), Ganglion Cell Layer + Inner Plexiform Layer (GCL+IPL), and macular regions.

2.1 Segmentation Algorithms

The measurement of retinal structures, such as RNFL, macular, and GCL thickness, relies on automated segmentation algorithms that identify and delineate different retinal layers (Zahavi et al., 2021). The accuracy of these measurements directly depends on the performance of these algorithms. As highlighted in research on segmentation software performance, variations in algorithms between manufacturers can lead to discrepancies in measurements (Tian et al., 2016). Specifically, the Cirrus HD-OCT (Zeiss) utilizes a spectral-domain approach and is known for its robust RNFL analysis, with segmentation algorithms optimized for detecting glaucomatous changes. The Spectralis OCT (Heidelberg Engineering), also an SD-OCT system, offers high image quality and incorporates a confocal scanning laser ophthalmoscope (cSLO) for precise anatomical registration, enhancing image stability and follow-up measurements (Abe et al., 2015). Its segmentation capabilities are particularly strong for macular analysis and detailed layer segmentation. Canon OCT-HS100 and Xephilio OCT-A1 also use SD-OCT, offering high-speed image acquisition and detailed retinal layer analysis, although detailed information on their specific segmentation algorithms is proprietary information.

Differences between manufacturers extend beyond segmentation algorithms to include scan patterns, image processing techniques, and normative databases (Rivas-Villar et al., 2023). Scan patterns dictate the imaged retinal regions and data acquisition density (Society, 2024). Different patterns may emphasize different anatomical regions, potentially influencing measurement outcomes. Image processing techniques, such as noise reduction and motion correction, also vary, affecting image quality and measurement precision. Normative databases, used to compare individual measurements to a healthy population, are manufacturer-specific and may introduce variability in interpretation. Therefore, it is crucial to consider these factors when comparing measurements from different OCT platforms and interpreting clinical data.

The increasing need for independent and customizable OCT image processing has led to the development of open-source segmentation tools. These tools provide flexibility and reproducibility, complementing proprietary software by enabling independent analyses, customized workflows, and cross-platform applications. However, these tools have limitations when applied to specific OCT data formats, such as those extracted from Canon OCT-HS100 and Xephilio OCT-A1 systems, which are not natively supported by these tools.

EyeSeg (Perry and Fernandez, 2020) is a fast and efficient encoder-decoder architecture specifically designed for accurate semantic segmentation tasks in contexts with limited annotated data. EyeSeg's innovative use of a customized loss function—combining categorical cross-entropy and generalized dice loss addresses challenges posed by imbalanced class distributions. Initially validated in the OpenEDS 2020 Semantic Segmentation Challenge, EyeSeg achieved a 94.5% mean Intersection Over Union (mIOU), outperforming state-of-the-art models by 10.5% in accuracy. It incorporates advanced features such as residual connections and dilated convolutions, allowing for computational efficiency without sacrificing performance. EyeSeg is particularly well-suited for tasks requiring precise segmentation, such as eye-tracking in augmented reality and virtual reality environments. However, while it demonstrates robust segmentation capabilities, EyeSeg's applicability to Canon OCT-HS100 or Xephilio OCT-A1 data would require significant adaptation, as it was developed for datasets like OpenEDS2020, which differ in imaging protocols and data structures.

Eyepy, a python-based framework, simplifies the processing of OCT volumes using the unified EyeVolume object, supporting formats such as Heyex (E2E, VOL, XML), Topcon FDA, and public datasets e.g., Duke AMD and RETOUCH Challenge (Morelle, 2023). It enables visualization, quantification (e.g., drusen analysis), and data handling. While highly flexible, its lack of native support for Canon OCT-HS100 or Xephilio OCT-A1 data and data conversion steps or modifications to the codebase might be necessary to bridge compatibility gaps.

PyOCT focuses on spectral-domain OCT (SD-OCT) and digital holography microscopy (DHM), offering tools for reconstruction workflows like background subtraction, spectral resampling, dispersion correction, and Fourier transformations (Yuechuan, 2022). It also includes advanced DHM features like phase retrieval and quantitative phase imaging. PyOCT supports batch processing for large datasets and extensive parameter customization but requires pre-processing or adaptation to handle Canon and Xephilio OCT data, which utilize proprietary imaging protocols and file structures.

ReLayer (Ometto et al., 2019) is an open-access, web-based platform that facilitates retinal layer segmentation and thickness measurement extraction from OCT images. Designed to work with devices, including Heidelberg Spectralis, Topcon 3D OCT-2000, and OptoVue AngioVue, ReLayer employs a two-step segmentation process. Initially, the platform applies Gaussian filtering to reduce noise and the Sobel operator to compute vertical gradients for edge detection. These gradients are analyzed column-wise to identify nodal points for retinal layers such as the inner limiting membrane (ILM), inner/outer segment (ISOS), and retinal pigment epithelium (RPE). The second step refines these boundaries using a one-dimensional active contour model that balances edge adherence with smoothness constraints, ensuring accurate and anatomically consistent segmentation. ReLayer's cross-platform adaptability is achieved through resampling, standardizing micrometer-to-pixel ratios, and device resolutions. Validation studies have shown ReLayer's performance to be highly accurate for healthy retinas and competitive for pathological cases, making it suitable for longitudinal and cross-sectional studies.

Another open-source segmentation algorithm from Bhargava et al., 2015 for multi-device application provides a robust framework for OCT segmentation, addressing the variability between devices such as Cirrus HD-OCT and Spectralis OCT. This algorithm operates through three main stages: preprocessing, pixel classification, and graph-based multilayer segmentation. Preprocessing involves normalizing intensities and flattening scans relative to the Bruch's membrane (BM) boundary. A random forest classifier then assigns probabilities for each pixel to belong to nine retinal layer boundaries, informed by manually segmented training data. Finally, a graph-based segmentation refines these probabilities by enforcing constraints on smoothness and inter-layer distances. Validation across multiple platforms demonstrated excellent agreement for layers such as the ganglion cell and inner plexiform layer (GCIP) and the inner nuclear layer (INL) at the cohort level. However, individual-level variability, particularly for the macular retinal nerve fiber layer (mRNFL), highlighted limitations in tracking longitudinal changes.

A graph-based multi-surface segmentation algorithm from Dufour et al., 2012 takes a novel approach by integrating prior knowledge models and energy minimization techniques. Designed for efficiency and robustness, this algorithm segments retinal layers and pathologies such as drusen. It employs a three-stage process: coarse segmentation of the ILM, simultaneous segmentation of the upper and lower retinal pigment epithelium (RPE) boundaries, and segmentation of inner retinal surfaces guided by prior models. The energy minimization framework incorporates

boundary energy, smoothness energy, and interaction energy, balancing anatomical consistency with flexibility to accommodate noise and pathologies. Hard constraints strictly enforce spatial relationships between layers, while soft constraints allow adaptive handling of morphological variations. The algorithm's efficiency is demonstrated by its execution time of under 15 seconds per volume and memory usage of less than 1 GB. Validation on healthy and pathological datasets showed segmentation errors lower than inter-observer variability and comparable performance to manual annotations in challenging cases.

The existing open-source tools, while robust and highly customizable, face challenges in natively supporting data from Canon OCT-HS100 and Xephilio OCT-A1 devices due to the proprietary imaging protocols and data formats unique to these systems. Adapting these tools requires a multi-step process. First, developing data conversion pipelines to translate proprietary formats into widely supported ones is essential for ensuring compatibility. Next, the algorithms must be adjusted to account for the unique imaging characteristics of these devices, including differences in resolution, scan patterns, and artifact profiles. Additionally, these adaptations need to be rigorously validated through systematic testing to ensure accuracy, reliability, and reproducibility in segmentation outcomes. These modifications are critical for integrating the advanced capabilities of open-source tools—such as EyeSeg, Eyepy, PyOCT, ReLayer, and other graph-based segmentation algorithms with data from Canon and Xephilio OCT devices, thereby extending their utility to a broader range of imaging systems.

2.2 OCT Device Comparison

We examine how these devices acquire measurements in key anatomical regions (optic disc, retinal nerve fiber layer (RNFL), and macula) and highlight the technological and functional differences among manufacturers. The optic disc is a critical region for evaluating glaucomatous damage. The Heidelberg Spectralis OCT employs the Bruch's Membrane Opening-Minimum Rim Width (BMO-MRW) method, which measures the neuroretinal rim width using 24 radial scans centered on the optic nerve head (Mitsch et al., 2019). This method offers higher geometric precision and reliability, particularly for anatomically complex cases. Spectralis also incorporates metrics such as neuroretinal rim volume and ONH asymmetry analysis, enhancing its diagnostic capabilities, for technical details see Chauhan and Burgoyne, 2013.

The Cirrus HD-OCT employs a cube-based approach for optic nerve head analysis, specifically for measuring rim area and cup-to-disc parameters, rather than directly calculating neuroretinal rim width (NRW) using the Bruch's Membrane Opening-Minimum Rim Width (BMO-MRW) method. The device acquires a 6×6 mm Optic Disc Cube scan, consisting of 200 B-scans, each with 200 A-scans, to create a detailed 3D representation of the optic nerve head. Instead of using the BMO-MRW technique, Cirrus HD-OCT employs a unique algorithm that determines the rim area through a three-dimensional calculation (Mitsch et al., 2019). This process involves, the Bruch's Membrane Opening serving as the foundational reference point for defining the inner margin of the neuroretinal rim, Adjacent radial scan axes defining lateral edges, while intersections with the internal limiting membrane (ILM) form the peripheral boundaries and optimization of the cross-sectional area representing the lateral vectors, more details on the measurements are described in the patents for the Cirrus HD-OCT device in Everett and Oakley, 2015.

The Canon HS100 OCT uses a 6×6 mm Disc 3D scan with 256 B-scans (each containing 512

A-scans), in contrast to the BMO-RW radial scan methodology employed by the Spectralis. This process is automated, and specific acquisition details are proprietary. Unlike the Spectralis, Canon does not offer ONH asymmetry analysis or BMO-centric alignment for cup-to-disc ratio measurements, which may limit its diagnostic potential (Brautaset et al., 2016). A summary of the differences in optic disc measurements between the devices is provided in Table 1.

Feature	**Cirrus HD-OCT**	**Spectralis OCT**	**Canon HS100 OCT**
Neuroretinal Rim Width	Cube-based rim area measurement	BMO-MRW radial scans	Disc 3D scan protocol
ONH Asymmetry Analysis	No	Yes	No
Cup-to-Disc Ratio	Derived from cube scan	BMO-centric analysis	Automated cup-to-disc ratio
Neuroretinal Rim Volume	No	Yes	No
Alignment and Tracking	Basic tracking	cSLO-enhanced alignment	Automatic alignment and tracking

Table 1: Comparison of Optic Disc Measurements Across OCT Devices. **Source:** Compiled from manufacturer specifications, published research papers, and independent analysis.

Retinal nerve fiber layer (RNFL) thickness is another vital parameter in multiple sclerosis diagnostics. Cirrus HD-OCT measures RNFL thickness along a single 3.46 mm circular scan centered on the optic nerve head. It provides average, quadrant, and clock-hour thickness measurements, visualized through a TSNIT plot, which is compared against an age-matched normative database. The device also constructs RNFL thickness maps using a 6×6 mm cube scan for enhanced spatial assessment (Cirrus, 2020).

Region	**Cirrus HD-OCT**	**Spectralis OCT**	**Canon HS100 OCT**
RNFL Thickness	Single scan circle	Multiple concentric circles	Single scan circle
Additional RNFL Circles	No	Yes	No
RNFL Thickness Map	Cube-based	BMO-aligned	Single 3D scan
Macular Layer Segmentation	GCL+IPL	Full segmentation of all layers	GCL+IPL & NFL+GCL+IPL
Macular Asymmetry Analysis	No	Intra- and inter-eye analysis	No
Posterior Pole Analysis	No	Yes	No
Normative Database	Age-specific	Age-, race-, and axial-length-matched	Age-specific
Alignment and Tracking	Basic tracking	cSLO-enhanced	Automatic

Table 2: Comparison of RNFL and macular measurements across OCT devices, including Cirrus HD-OCT, Spectralis OCT, and Canon HS100 OCT. **Source:** Compiled from manufacturer specifications, published research papers, and independent analysis.

Spectralis OCT offers RNFL thickness measurements using multiple concentric circles (3.46 mm, 4.1 mm, and 4.7 mm), allowing for a more comprehensive evaluation of regional RNFL variations. This device also aligns RNFL scans with the BMO center, enhancing reproducibility and accuracy. The normative comparison includes adjustments for age, race, and axial length, making it more robust than Cirrus (Spectralis, 2025). Canon HS100 OCT measures RNFL thickness along a single 3.46 mm circle, similar to Cirrus, and generates TSNIT plots based on age-matched normative data. However, it does not support concentric scans or cube-based RNFL mapping (Canon, 2020).

In macular analysis, Cirrus HD-OCT focuses on Ganglion Cell Analysis (GCA) to measure the combined thickness of the ganglion cell layer and inner plexiform layer (GCL+IPL). This measurement has been seen to be a reliable biomarker for early glaucoma detection. However, Cirrus does not segment individual macular layers or provide intra-eye asymmetry analysis. Spectralis OCT initially measured total macular thickness but now includes detailed segmentation of macular layers (e.g., GCL+IPL, RNFL) with the Glaucoma Module Premium Edition (GMPE). It also performs intra-eye and inter-eye asymmetry analyses, as well as posterior pole asymmetry analysis, which compares superior and inferior retinal thickness. Canon HS100 OCT measures GCL+IPL and NFL+GCL+IPL thickness but lacks segmentation of individual layers and asymmetry analysis. The comparative features of RNFL and macula measurements are summarized in Table 2.

2.3 Advancements in OCT Analysis

Beyond the hardware and segmentation methods of the specific devices, advancements in deep learning have introduced significant innovations in OCT image segmentation. This has enabled automated segmentation with higher accuracy and efficiency. Fully supervised learning models, such as U-Net and its variants, have set benchmarks in retinal layer segmentation, achieving Dice coefficients exceeding 91.1% (Lee et al., 2017). These models excel in segmenting layers such as the RNFL and GCL+IPL, which are critical for assessing conditions like multiple sclerosis, macular edema, glaucoma etc. However, the requirement for large, annotated datasets remains a barrier, as manual labeling of OCT scans is labor-intensive and requires expert knowledge.

To mitigate this challenge, semi-supervised and weakly supervised methods have gained traction. Semi-supervised techniques, which combine labeled and unlabeled data during training, reduce the dependence on extensive manual annotations. For instance, adversarial learning frameworks have shown promise in enhancing segmentation accuracy by leveraging unlabeled data to complement limited labeled datasets (Liu et al., 2018). Similarly, weakly supervised models, which rely on image-level annotations, have demonstrated competitive performance, particularly in identifying and segmenting small pathological regions, though they often require specialized pre-processing to adapt to device-specific imaging protocols (Liu et al., 2018).

Another critical area of innovation in OCT imaging is the development of explainable AI models. For clinical applications, transparency in the decision-making process is essential. Tools such as Class Activation Mapping (CAM) and attention mechanisms, (Zhou et al., 2016) enable visualization of the regions influencing model predictions, ensuring alignment with clinical understanding and facilitating adoption in medical practice. These tools are especially valuable in high-stakes diagnostics, where confidence in automated segmentation outcomes is important.

The variability in OCT device outputs stemming from differences in resolution, scan patterns, and proprietary algorithms remains a significant challenge. Open-source tools like EyeSeg, Eyepy, and PyOCT are designed to provide independent and flexible segmentation capabilities but face limitations in handling proprietary data formats from devices like Canon HS100 OCT and Xephilio OCT-A1. Addressing these compatibility issues requires developing data conversion pipelines and adjusting algorithms to accommodate the unique imaging characteristics of these devices. Rigorous validation of these adaptations is critical to ensure the reliability of measurements derived from open-source tools.

In the future, reinforcement learning and hybrid models combining traditional and deep learning techniques present opportunities for further advancement . These models aim to generalize across datasets and devices, overcoming the challenges posed by data variability and device-specific segmentation requirements (Minaee et al., 2021). Additionally, efforts to create standardized imaging protocols and expand the availability of diverse, annotated datasets will be instrumental in fostering collaboration and innovation within the field.

2.4 Related Works Using OCT Data for MS Diagnosis

A number of studies have demonstrated the application of machine learning techniques in the OCT domain for diagnosing multiple sclerosis. These studies demonstrate the potential of ML algorithms in distinguishing MS patients from healthy controls, utilizing OCT derived features. For example García Mesa et al., 2023 achieved 87.3% accuracy using RF, k-NN, and SVM for distinguishing MS patients from controls with Cirrus HD-OCT 5000 imaging.

Similarly, Montolío et al., 2022 reported 95.8% accuracy with SVM, neural networks, and ensemble classifiers with Spectralis OCT data. These results highlight the robustness of RF for OCT-derived feature analysis in MS.

Higher diagnostic accuracies in some studies can be attributed to advanced advanced OCT devices and diverse datasets, for instance Palomar et al., 2019 achieved an accuracy of 95.74% with Swept-Source OCT (SS-OCT) data using SVM, RF, and AdaBoost classifiers. This study also reported high sensitivity (97.22%) and specificity (95.16%), showcasing the utility of different classifiers and imaging technologies.

Furthermore, Garcia-Martin et al., 2021 demonstrated near-perfect diagnostic performance with sensitivity and specificity values of 98% using a convolutional neural network (CNN) applied to Spectralis OCT data. These findings highlight the potential of deep learning approaches for OCT-based MS analysis.

Author	Analysis Type	Subjects	Machine	Classifiers	Results(best model)
García Mesa et al., 2023	Diagnosis	40 MS patients, 27 Controls	Cirrus HD-OCT 5000	SVM, k-NN, Random Forest, Bagging Classifier	Accuracy: 87.3%, F1-score: 87.3%, AUC: 87.6%
Montolío et al., 2022	Diagnosis	72 MS patients, 30 Controls	Spectralis OCT	SVM, k-NN, DT, Naive Bayes, Ensemble Classifier, Neural Networks	Accuracy: 95.8%, AUC: 95.8%
Montolío et al., 2021	Diagnosis	108 MS patients, 104 Controls	Cirrus HD-OCT	Multiple Linear Regression, SVM, DT, k-NN, NB	Accuracy: 87.7%; Sensitivity: 87.0%; Specificity: 88.5%; Precision: 88.7%; AUC: 0.8775)
Garcia-Martin et al., 2021	Diagnosis	48 MS patients, 48 Controls	Spectralis OCT	CNN	Sensitivity = Specificity = 0.98.
Cavaliere et al., 2019	Diagnosis	28 MS patients, 22 Controls, 15 RIS, 31 CIS	Spectralis OCT	k-NN	Accuracy FMC: 54%, HC: 74% MCC FMC: 43%, HC: 68%
Palomar et al., 2019	Diagnosis	80 MS patients, 180 Controls	Swept-Source OCT (SS-OCT)	SVM, Random Forest, AdaBoost	Accuracy: 95.74%, Sensitivity: 97.22%, Specificity: 95.16%

Table 3: Comparison of MS studies using OCT data and machine learning

Montolío et al., 2022 explored the diagnostic capacity of Cirrus HD-OCT data combined with multiple ML models, including SVM and Decision Trees (DT). The study reported an accuracy of 87.7%, along with a sensitivity of 87.0%, specificity of 88.5%, and a precision of 88.7%. In contrast, Cavaliere et al., 2019 evaluated k-NN classifiers for distinguishing between MS, Radiologically Isolated Syndrome (RIS), and Clinically Isolated Syndrome (CIS), obtaining varying accuracies depending on the classification focus, such as 74% for healthy controls.

The variability in diagnostic performance across studies reflects differences in OCT devices, datasets, and ML models. Studies employing high-resolution OCT devices, such as Spectralis, Cirrus-OCT often report superior performance, emphasizing the importance of imaging quality and advanced classification techniques. Table 3 provides a comprehensive summary of these studies.

3 Data

3.1 Dataset

This study utilized OCT scans from patients with Multiple Sclerosis from the MS Centrum Pelt. The dataset included two eyes per patient, providing a detailed representation of retinal features for each individual. Patients underwent comprehensive ophthalmological examinations, with OCT measurements obtained using the Canon OCT-HS100. The scans were performed using the Macula 3D (10 mm x 10 mm scanning area), Glaucoma 3D (10 mm x 10 mm scanning area),

and Disc 3D (6 mm x 6 mm scanning area) protocols. EDSS scores, based on the McDonald criteria, were recorded during these examinations. The OCT dataset initially contained 740 patients prior to data preparation.

3.2 Data Extraction and Preparation

The dataset was securely stored in an encrypted format with anonymized patient IDs and was accessed via a controlled container, ensuring data security and confidentiality. It consisted of two main components: clinical patient data stored in an *imed* file, and OCT scan data provided in *XML* files. The *imed* file contained information such as patient demographics, MS classification, visit details, patient treatments and clinical measures, including the Expanded Disability Status Scale (EDSS). The *XML* files recorded OCT measurements for both eyes, with scans from up to 740 patients across multiple visits.

A data preparation process was carried out to create a clean, integrated dataset suitable for analysis by combining clinical patient data and OCT measurement data while ensuring consistency across all records. This process involved several key steps:

First, clinical patient data from the *imed* file, including patient-specific attributes such as year of birth, onset date, diagnosis date, treatment records, and the start of progression, were linked to the OCT measurement dataset from the *XML* files. Next, EDSS scores and their corresponding visit dates, sourced from the clinical records in the *imed* file, were merged with the OCT dataset. To ensure alignment between the OCT scans and the clinical assessments, the OCT scan dates were matched to the closest corresponding EDSS visit dates. This matching ensured that OCT measurements accurately reflected the clinical context.

To facilitate meaningful analysis, the dataset was further filtered to include only records where both OCT data and EDSS scores were available for both eyes of a patient, with both measurements taken on the same day. Imputation for missing EDSS scores was deemed inappropriate, as it could give a false impression of treatment outcomes and the prognosis of MS, particularly in the context of this retrospective study. The final dataset of 230 patients, classified into two categories of MS severity on consultation with an ophthalmologist, Non-Severe ($EDSS < 6$) and Severe ($EDSS \geq 6$), enabled robust analysis of the relationships between OCT features and MS disease severity.

3.3 Study Variables

Variables encompassing patient demographics, clinical information, and detailed optical coherence tomography (OCT) features were utilized to explore potential predictors of disease severity in Multiple Sclerosis (MS). The primary outcome variable was the EDSS Category, which classified patients into two groups: Not Severe and Severe.

Demographic and clinical variables included patient-specific information such as year of birth, gender, ethnicity, date of onset of MS symptoms and diagnosis date. These variables were critical for understanding the progression of MS and its influence on patients' clinical profiles.

The dataset also included OCT features extracted from the *XML* files, categorized based on the anatomical regions of the eye. These features provided detailed insights into retinal structure, enabling the analysis of changes associated with multiple sclerosis.

The Four Sectors measurements of the Optic Disc were derived from the temporal, superior, nasal, and inferior retinal sectors. These features offered a broad view of the retinal nerve fiber layer (RNFL) health, helping to identify potential localized differences in thickness. Additionally, Twelve Sector measurements of the Optic Disc captured data from twelve distinct regions of the RNFL, providing a more granular structural analysis of the retina.

Variables	**Description**
PatientID, Birth Date, Gender, Ethnicity, Date of Onset, Diagnosis Date, Start of Progression	Patient clinical data
Temporal, Superior, Nasal, Inferior	Retinal Nerve Fiber Layer (RNFL) thickness values for the Four Retinal Sectors of the peripapillary region (Optic Disc)
Temporal (Inferotemporal, Superotemporal), Superior (Superotemporal, Superonasal), Nasal (Superonasal, Inferonasal), Inferior (Inferonasal, Inferotemporal)	Retinal Nerve Fiber Layer (RNFL) thickness values for the Twelve Retinal Sectors of the peripapillary region (Optic Disc).
Disc Area, Rim Area, Cup Volume, Rim Volume, Cup-to-Disc Area Ratio (CDR), Vertical Cup-to-Disc Ratio (VCDR), Horizontal Cup-to-Disc Ratio (HCDR), Rim-to-Disc Minimum Ratio (RDMR), Disc Damage Likelihood Scale (DDLS)	Optic Nerve Head measurements.
TSNIT Average, Standard Deviation	Average and standard deviation of Retinal Nerve Fiber Layer (RNFL) thickness across the Temporal, Superior, Nasal, and Inferior regions.
Total(GCL+IPL), Total(NFL+GCL+IPL), Superior(GCL+IPL, NFL+GCL+IPL), Inferior(GCL+IPL, NFL+GCL+IPL), Paracentral and Peripheral Thickness Values (Para/PeriInferiorNasal(GCL+IPL, NFL+GCL+IPL), Para/PeriInferiorTemporal(GCL+IPL, NFL+GCL+IPL), Para/PeriSuperiorTemporal(GCL+IPL, NFL+GCL+IPL), Para/PeriSuperiorNasal(GCL+IPL, NFL+GCL+IPL))	Combined thickness values of the macula and optic disc, including Retinal Nerve Fiber Layer (RNFL), Ganglion Cell Layer (GCL), and Inner Plexiform Layer (IPL).
Paracentral and Peripheral Thickness(Para/PeriTemporal, Para/PeriNasal, Para/PeriSuperior, Para/PeriInferior), Central Fovea Thickness, Minimum Thickness, Average Thickness, Volume	Thickness values for the Internal Limiting Membrane to Retinal Pigment Epithelium (ILM-RPE).
Paracentral and Peripheral Thickness(Para/PeriTemporal, Para/PeriNasal, Para/PeriSuperior, Para/PeriInferior), Central Fovea Thickness, Minimum Thickness, Average Thickness	Thickness values for the Internal Limiting Membrane to Bruch's Membrane (ILM-BM).

Table 4: Summary of Variables

The Optic Nerve Head (ONH) features included variables such as disc area, rim area, and rim volume. These features were critical for evaluating neuro-degenerative changes that might affect the optic nerve head. Complementary to this, RNFL features such as the TSNIT (Temporal-Superior-Nasal-Inferior-Temporal) average and standard deviation provided summary metrics of RNFL thickness across various sectors of Optic Disc.

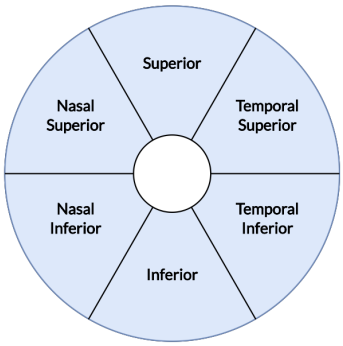
The dataset also contained features representing the combined thickness of retinal layers of the macula and optic disc. GCL+IPL features represented the combined thickness of the ganglion cell layer (GCL) and inner plexiform layer (IPL), while NFL+GCL+IPL features aggregated measurements from the nerve fiber layer (NFL), GCL, and IPL. These combined measurements were particularly useful for assessing retinal layers that could possibly contribute to the neuro-

degenerative processes.

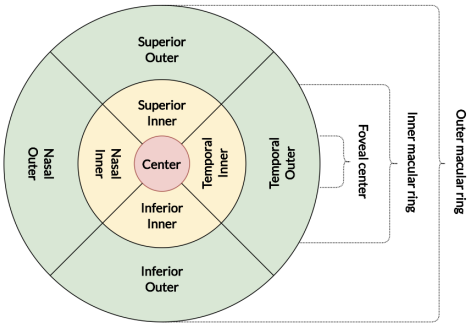
Features such as ILM-RPE and ILM-BM spanned the internal limiting membrane (ILM) to the retinal pigment epithelium (RPE) or Bruch's membrane (BM). These features offered broader insights into retinal and macular health, complementing the more localized metrics provided by the other features. Table 4 summarizes the variables included in the dataset.

3.4 OCT Protocols

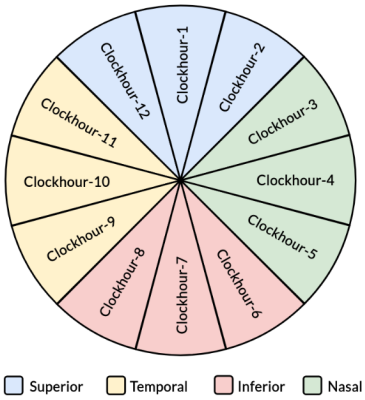
The Canon OCT-HS100 device provides a comprehensive analysis of both retinal and optic-disc regions through a combination of quantitative and qualitative evaluations. This study used three distinct scan modes i.e., Macula 3D, Glaucoma 3D, and Disc 3D. Macula 3D is performed on the region centered on the macula, with a scanning area of 10x10 mm and the primary scanning direction being horizontal. Glaucoma 3D is similarly centered on the macula with the same scan area of 10x10 mm, but the primary scanning direction is vertical. Disc 3D is centered on the optic disc, with a scan area of 6x6 mm and the primary scanning direction being horizontal.



(a) Ganglion Cell Analysis



(b) ETDRS for Macular Thickness Analysis



(c) Optic-Disc RNFL and ONH Analyses

Figure 1: OCT Analysis Sectors (García Mesa et al., 2023)

These modes enable detailed assessments in three key categories, the Ganglion Cell Analysis, which evaluates the Ganglion Cell Layer (GCL) and Inner Plexiform Layer (IPL); Macular Thickness Analysis, which measures the macular thickness; Optic-Disc Retinal Nerve Fiber Layer(RNFL) and Optic Nerve Head (ONH) analyses, which provide a detailed evaluation of the optic disc and Retinal Nerve Fiber Layer. Each of these scan modes have a different focus area, ensuring that both macular and optic-disc regions are thoroughly analyzed.

The Ganglion Cell Analysis utilized six-sector maps centered on the fovea for evaluation as seen in Figure 1a. The Macular Thickness Analysis applied the Early Treatment Diabetic Retinopathy Study (ETDRS) grid, which was positioned at the center of the fovea, dividing the macular region into nine distinct regions as seen in Figure 1b. For the Optic Disc Analysis, a clock-like grid was used, segmenting the optic disc into twelve sectors and four anatomical regions as seen in Figure 1c.

3.5 Exploratory Data Analysis

The study included a diverse cohort of 230 patients, with demographic and clinical characteristics summarized in Table 5. These details show gender distribution, age range, and EDSS scores, providing a comprehensive overview of the sample and the distribution of disease severity.

Demographic Information	**Category**	**Details**
Gender Distribution	Male	77 patients (33.5%)
	Female	153 patients (66.5%)
Age Distribution	Age Range	20–82 years
	Mean Age	49.8 years
	Standard Deviation	11.5 years
	Age Groups	20–30 years: 10 patients (4.3%) 31–40 years: 40 patients (17.3%) 41–50 years: 64 patients (27.8%) 51–60 years: 70 patients (30.4%) 61–70 years: 38 patients (16.5%) 70+ years: 8 patients (3.4%)
EDSS Scores	EDSS Range	0 to 8
	Mean EDSS score	3
	Standard Deviation	1.8

Table 5: Demographic and clinical characteristics of the study population.

3.6 Software

All statistical analyses were conducted using R (version 4.3.3). The normality of numerical variables was evaluated using the Shapiro-Wilk test. Statistical significance was defined as a p-value < 0.05. The Machine learning experiments, including model training and evaluation, were performed using Python (version 3.9.20) with the scikit-learn library.

4 Methodology

A range of machine learning algorithms was utilized to classify multiple sclerosis (MS) patients based on their clinical and demographic data. The selected models Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and XGBoost (eXtreme Gradient Boosting) were chosen for their effectiveness in handling diverse datasets and their ability to provide reliable classification results. Each model was carefully configured and evaluated to determine its suitability for predicting MS severity.

4.1 Support Vector Machine (SVM)

Support Vector Machine is an algorithm for linear and non-linear classification problems. The main objective of SVM is to find the optimal hyperplane that maximizes the margin between classes. The decision function of a linear SVM classifier is given by:

$$f(x) = \mathbf{w}^T x + b$$

where $\mathbf{w}$ is the weight vector representing the direction of the boundary, b is the bias term, a constant that shifts the boundary and x is the input feature vector. The margin is defined as the distance between the decision boundary and the closest data points, called support vectors. The optimization problem for SVM is:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w}^T x_i + b) \geq 1, \quad \forall i \tag{1}$$

where y_i is the label of the i-th data point.

To find this optimal boundary, SVM solves the optimization problem in (1) that ensures all data points are correctly classified while maximizing the margin. For more complex, non-linear problems, SVM uses kernel functions (like polynomial or radial basis functions) to map the data into a higher-dimensional space, where it becomes easier to separate the classes. The SVM was chosen for its ability to handle both linear and non-linear classification tasks.

4.2 Random Forest

Random Forest is an ensemble learning method based on decision trees. Decision tree classifiers are supervised methods that model outcomes and predictions using a flowchart–like a tree structure. Trees are composed from top to bottom of a root node, internal nodes (branches) and leaf nodes. The tree is constructed via a process of an if-else statement that identifies ways to split predictor space and classify the outcome based on different conditions of predictors. In a decision tree, each internal node represents a test on a feature of a dataset, each leaf node represents an outcome, and branches represent the decision rules that lead to class labels. These trees are relatively unstable and are normally replaced by a random forest of decision trees.

The goal is to build an ensemble of T decision trees, where each tree $h_t(x)$ is trained on a random subset of the training data by sampling with replacement. The final classification prediction, $\hat{y}$ is obtained by aggregating the outputs of these trees through majority voting:

$$\hat{y} = \text{mode}\left(\{h_1(x), h_2(x), \ldots, h_T(x)\}\right)$$

where

$$h_t(x)$$

represents the prediction of the t-th decision tree for an input x. During the training process of a decision tree, the algorithm iteratively splits the data at each node by selecting the feature that provides the most informative division of the dataset. This selection is based on a criterion that measures the purity of the subsets created after the split. In our case, the Gini impurity was used as the splitting criterion, calculated as:

$$\text{Gini Index} = 1 - \sum_{i=1}^{K} p_i^2,$$

where p_i is the proportion of samples belonging to class i.

To further enhance diversity among the trees, Random Forest limits the number of features considered for splitting at each node to a random subset of the total features. Random Forest efficiently handles datasets with many features, making it ideal for our application.

4.3 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a simple, non-parametric classification algorithm that assigns a label to a data point based on the majority class of its k-nearest neighbors in the feature space. Given a test point x, the KNN classifier finds the k-nearest training points in the feature space, typically using the Euclidean distance:

$$d(x, x') = \sqrt{\sum_{i=1}^{m} (x_i - x_i')^2}.$$

The prediction is made by selecting the majority class among the k-nearest neighbors:

$$\hat{y} = \text{mode}\left(\{y_{i_1}, y_{i_2}, \ldots, y_{i_k}\}\right)$$

where $y_{i_1}, y_{i_2}, \ldots, y_{i_k}$ are the labels of the k-nearest neighbors.

KNN was used for its simplicity and its effectiveness in capturing local patterns in the data. Since MS progression can vary significantly between individuals, KNN is useful for classifying patients based on the similarity to their neighbors in the feature space.

4.4 XGBoost

XGBoost is a gradient boosting algorithm that combines weak learners (decision trees) sequentially to minimize a loss function. At each iteration, a new tree is added to correct the errors of the previous ensemble:

$$f_t(x) = f_{t-1}(x) + h_t(x),$$

where $h_t(x)$ is the new tree and $f_t(x)$ is the updated prediction.

where $F_t(x)$ is the prediction of the model at iteration t, $h_t(x)$ is the weak learner, and η is the learning rate. The final prediction for a data point x is obtained by summing the contributions of all T weak learners:

$$\hat{y} = \sum_{t=1}^{T} \eta h_t(x).$$

In order to prevent over-fitting XGBoost uses an objective function to improve its performance and reduce over-fitting. This objective function includes two parts: one measures how well the model predicts the actual values, and the other adds a penalty for making the model too complex. By adding this penalty, XGBoost prevents the model from fitting the noise in the training data, which helps it perform better on new, unseen data. XGBoost Boosting was used for since it focuses on correcting errors iteratively, it can effectively capture non-linear relationships in the data and provide accurate classifications even with limited data.

4.5 Hyperparameter Tuning

Hyperparameter tuning is a crucial step in the model development process, allowing us to optimize model performance by identifying the best combination of parameters for each algorithm. For this study, hyperparameter tuning was done using a grid search approach. This method involves systematically searching through a predefined range of hyperparameters for each model to find the configuration that yields the best cross-validation performance. Cross-validation was employed to ensure that the selected parameters generalize well to unseen data by evaluating the model across multiple data splits.

The grid search explored combinations of hyperparameters specific to each model. For Random Forest, parameters such as the number of estimators, minimum samples required for splitting and leaf nodes, and tree depth were tuned. For Support Vector Machines (SVM), the kernel type, gamma, and regularization parameter C were adjusted, along with class weighting in some cases. XGBoost required tuning of parameters like the number of estimators, learning rate, maximum tree depth, and subsample ratio. For K-Nearest Neighbors (KNN), the number of neighbors, distance metric, and weighting scheme were varied to identify the optimal configuration.

4.6 Model Assessment

The performance of the classification models was evaluated using several metrics derived from the confusion matrix. The confusion matrix is useful for calculation of the accuracy, sensitivity, specificity and precision. It includes four parameters True Positives (TP), which represents the positive samples correctly classified as positive; True Negatives (TN), which represents the negative samples correctly classified as negative; False Positives (FP), which shows the negative samples incorrectly classified as positive; and False Negatives (FN), which represents the positive samples incorrectly classified as negative.

Accuracy was calculated to measure the overall correctness of the models, representing the proportion of correctly classified instances, both positive and negative, out of the total instances. It was determined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Sensitivity, also referred to as recall, quantified the proportion of actual positive cases correctly identified by the models. Sensitivity represented the models' ability to correctly classify severe MS patients and was calculated as

$$\text{Sensitivity} = \frac{TP}{TP + FN}.$$

Specificity was used to evaluate the proportion of actual negative cases correctly identified, reflecting the ability to classify non-severe MS patients correctly. It was expressed as

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

Balanced accuracy was used to provide a fair evaluation of the models on the imbalanced dataset. It was calculated as the average of sensitivity and specificity, ensuring equal weight for both classes:

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}.$$

Precision, or positive predictive value, measured the proportion of predicted positive cases that were truly positive, assessing the reliability of the models' positive predictions. Precision was computed as

$$\text{Precision} = \frac{TP}{TP + FP}.$$

The F1 score was used to provide a balanced measure of the models' performance by calculating the mean of precision and sensitivity. This metric was particularly important for addressing the trade-off between false positives and false negatives, especially given the class imbalance in the dataset. The F1 score was defined as

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}.$$

In addition to these metrics, the Receiver Operating Characteristic (ROC) curve was analyzed to evaluate the models' diagnostic performance across various classification thresholds. The ROC curve plotted the true positive rate (TPR), or sensitivity, against the false positive rate (FPR), which was calculated as

$$\text{FPR} = \frac{FP}{FP + TN}.$$

This curve illustrated the trade-off between sensitivity and specificity as the discrimination threshold varied. The area under the curve (AUC) was used as a single scalar value to summarize the model's performance across all thresholds. An AUC value closer to 1 indicated superior classification performance, with higher sensitivity and specificity across a range of thresholds.

4.7 Machine Learning Pipeline

A machine learning pipeline was implemented to classify MS patients and predict disability status using clinical and OCT data. The pipeline consisted of several key steps: data pre-processing, feature selection, model building, cross-validation, and model evaluation as shown in Figure 2.

During data pre-processing, two issues were addressed: missing values in numerical OCT measurements and class imbalance in the outcome variable, EDSS Category. Missing values were handled through imputation using sector means for the features. To enhance the quality of the dataset, normalization was applied to numerical variables, and one-hot encoding was applied to categorical variables in the patient demographic data. The numerical variables in the training set were scaled to have a mean of 0 and a standard deviation of 1. The test set was then normalized using the mean and standard deviation of the training set to prevent information leakage during training.

After pre-processing, the dataset was split into training and test sets to facilitate model development and evaluation. The training set was used to build the predictive models, while the test set served as an independent dataset for performance evaluation.

To address the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied on the training set. SMOTE is commonly used in clinical machine learning studies to create a balanced dataset by generating synthetic examples for the minority class. This technique involves selecting a random sample from the minority class and identifying its k-nearest neighbors (k-NN) (Chawla et al., 2002). For this study, the default values k=5 was used. Synthetic data points were generated by interpolating between the selected sample and its neighbors, augmenting the minority class to be equal to the majority class.

Feature selection is a critical step in the development of predictive models, offering several benefits such as reducing over-fitting, improving predictive accuracy, and lowering computational costs. To identify the most relevant features in our analysis, feature selection was performed using a Logistic Regression model with Elastic Net regularization, optimized through Grid Search. Elastic Net, which combines L1 and L2 penalties, was chosen for its ability to handle feature selection by shrinking less relevant coefficients to zero. The Grid Search optimized several hyperparameters, including the regularization strength and the ratio between L1 and L2 penalties. After cross-validation, features with non-zero coefficients in the final model were retained, and the original data was filtered accordingly. This resulted in new training set containing only the selected features, which were used across all models.

Model building was carried out using the filtered training set, four different models were fitted to the data. Classifier performance was optimized through hyperparameter tuning using Grid Search, which systematically explored combinations of hyperparameter values to minimize misclassification rates and maximize overall performance.

To address the limitations of a relatively small dataset that resulted from our pre-processing and reduce the risk of over-fitting, k-fold cross-validation was employed. This method ensures that results are not influenced by the initial data split, providing a more robust assessment of model performance. The dataset was randomly partitioned into k folds, with one fold serving as the test set and the remaining k-1 folds used as the training set. This process was repeated k times, allowing each fold to act as the test set once. The overall performance of the model was calculated by averaging the results across all k iterations. A 10-fold cross-validation approach

was adopted, as it offers a balance between computational efficiency and reliable performance estimation.

The final step involved evaluating the models on the test set. This independent evaluation provided an unbiased assessment of model performance. The pipeline steps, from data preprocessing to evaluation, are illustrated in Figure 2.

Figure 2: Machine learning Pipeline.

5 Results

5.1 Feature Engineering

Feature engineering plays a crucial role in transforming raw data into meaningful variables that can enhance the performance of machine learning models. This process involved creating new features to capture structural patterns in the eye and patient-specific metrics, which are crucial for diagnosing and monitoring multiple sclerosis (MS). By summarizing retinal and optic nerve characteristics, the engineered features provide a comprehensive understanding of the underlying changes associated with MS progression.

Key features derived from the OCT data include measurements from four quadrants: Temporal, Superior, Nasal, and Inferior. These measurements were used to calculate averages for each quadrant, offering a spatial understanding of the retinal nerve fiber layer (RNFL) thickness. The averages, named the Temporal , Superior , Nasal , and Inferior Quadrants, are essential for identifying thinning or thickening in specific areas, which are early signs of MS-related changes. Additionally, asymmetry metrics such as Superior-Inferior Asymmetry and Temporal-Nasal Asymmetry were developed to measure differences between these regions. These features are useful for detecting localized changes that might not be evident in overall averages.

Other important features focus on the optic nerve head. The Normalized Rim Area, which is the ratio of the rim area to the disc area, provides a way to assess changes in the optic nerve. Similarly, the Cup-to-Disc Area Ratio measures the size of the cup relative to the disc, helping to track optic nerve health. The Rim-to-Cup Volume Ratio, which compares the rim volume to the cup volume, offers additional insight into changes in the optic nerve.

Variance features were created to capture variability within specific retinal layers. These include Temporal (NFL+GCL+IPL) Variance, Nasal (NFL+GCL+IPL) Variance, Superior (NFL+GCL+IPL) Variance, and Inferior (NFL+GCL+IPL) Variance. These features highlight differences in the thickness of the Nerve Fiber Layer (NFL), Ganglion Cell Layer (GCL), and Inner Plexiform Layer (IPL) in various regions of the retina, helping to identify irregularities linked to MS.

Features related to retinal volume were also included. Total ILM-RPE Volume and Total ILM-BM Volume measure the overall thickness of the retina from the inner limiting membrane (ILM) to the retinal pigment epithelium (RPE) and Bruch's membrane (BM), respectively. Ratios such as the Ratio ILM-RPE Central-Peripheral and Ratio ILM-BM Para-Peri were added to compare differences between central and peripheral regions of the retina, highlighting specific patterns in thickness.

Finally, global metrics i.e., the Normalized TSNIT Average, which normalizes RNFL thickness along the TSNIT (Temporal-Superior-Nasal-Inferior-Temporal) profile relative to disc area, and TSNIT Coefficient of Variation (TSNIT CoV), which measures variability in this profile, provide an overall view of RNFL patterns and help identify disruptions linked to MS. The engineered features are summarized in Table 6.

Feature	Description
Temporal Quadrant, Superior Quadrant, Nasal Quadrant, Inferior Quadrant	Average RNFL thickness measurements across the Temporal, Superior, Nasal, and Inferior quadrants, summarizing regional RNFL health.
Superior-Inferior Asymmetry, Temporal-Nasal Asymmetry	Quantify imbalances between superior and inferior quadrants, and between temporal and nasal quadrants, highlighting localized structural changes.
Normalized Rim Area	Ratio of the optic nerve rim area to the disc area, indicating structural integrity of the optic nerve head.
Cup-to-Disc Area Ratio	Proportion of the cup area to the disc area, used to assess optic nerve degeneration.
Rim-to-Cup Volume Ratio	Ratio of rim volume to cup volume, reflecting the distribution of optic nerve volume.
Temporal NFL+GCL+IPL Variance, Nasal NFL+GCL+IPL Variance, Superior NFL+GCL+IPL Variance, Inferior NFL+GCL+IPL Variance	Variability in the thickness of the Nerve Fiber Layer (NFL), Ganglion Cell Layer (GCL), and Inner Plexiform Layer (IPL) across different regions, providing insights into retinal layer irregularities.
Total ILM-RPE Volume, Total ILM-BM Volume	Overall retinal volume from the Inner Limiting Membrane (ILM) to the Retinal Pigment Epithelium (RPE) and Bruch's Membrane (BM), capturing global retinal structure.
Ratio ILM-RPE Central-Peripheral, Ratio ILM-BM Para-Peri	Ratios capturing structural differences between central and peripheral regions of the retina, highlighting localized patterns.
Normalized TSNIT Average	RNFL thickness profile along the TSNIT (Temporal-Superior-Nasal-Inferior-Temporal) curve, normalized to the optic disc area.
TSNIT Coefficient of Variation (TSNIT CoV)	Variability in the TSNIT thickness profile, indicating consistency or irregularity in RNFL patterns.

Table 6: Summary of Engineered Features Derived from OCT Data

5.2 MS Diagnostic Model

A diagnostic model was developed to classify the severity of multiple sclerosis (MS) using data from 230 patients, comprising 154 Not Severe cases and 76 Severe cases. The analysis involved a greedy approach, where various dataset configurations were evaluated to determine the most effective combination of features. These configurations included inter-eye differences for all OCT features to capture asymmetrical optic nerve damage, total OCT features alone to assess global structural changes, and localized OCT features alone to focus on specific anatomical regions such as the optic disc, macula, and RNFL.

We also analyzed individual protocols separately to evaluate the predictive value of each imaging technique and explored datasets with treatment information to assess the potential influence on MS. While these approaches aimed to capture distinct aspects of MS pathology, we do not present results from these intermediate analyses. Instead, we focus exclusively on the diagnostic model derived from the final dataset configuration, which demonstrated the strongest performance and

interpretability.

The analysis utilized two datasets: Dataset 1 for the left eye and Dataset 2 for the right eye, each containing 75 features derived from optical coherence tomography (OCT) measurements using three distinct protocols. The datasets were analyzed separately because MS often presents asymmetrically, with optic nerve damage or retinal nerve fiber layer (RNFL) thinning differing between the two eyes. This independent analysis allows the model to capture eye-specific patterns of damage.

The first protocol, Ganglion Cell Layer and Inner Plexiform Layer (GCIPL), included 15 OCT features, capturing the structure of the retinal layers, which are commonly affected in MS. The second protocol, Macular Thickness, provided 24 OCT features, offering detailed measurements of the macula to evaluate potential damage caused by MS. The third protocol, Optic Disc and Retinal Nerve Fiber Layer (RNFL) measurements from the Optic Nerve Head, contributed 27 OCT features, essential for assessing the health of the optic nerve and detecting its degeneration in MS patients.

Here, "protocol" refers to the specific method used to obtain measurements for a particular aspect of OCT analysis, with each protocol generating a unique set of features for the dataset. For the analysis, the features derived from all three protocols were combined into a single dataset for each eye, creating a comprehensive dataset that included not only the OCT measurements but also patient clinical data and engineered features. This resulted in a total of 95 features in the dataset, allowing the model to leverage information from all protocols to classify MS severity.

A systematic approach was implemented for feature selection to optimize the classification models. Initially, a correlation analysis was conducted on all numeric features to identify pairs with a Spearman's rank correlation coefficient of 0.9 or higher. Features exhibiting high correlations were deemed redundant, as they contributed similar information to the models, possibly adding unnecessary complexity without improving predictive accuracy. As a result, the total number of features was reduced from 94 to 27 in both Dataset 1 (left eye) and Dataset 2 (right eye), for subsequent modeling.

The data was then split into training and test sets. The training set comprised 161 patients, with 108 classified as "Not Severe" and 53 as "Severe." The test set included 69 patients, with 46 "Not Severe" and 23 "Severe." To address the imbalance in the training set, the Synthetic Minority Oversampling Technique (SMOTE) was employed. SMOTE generated synthetic samples for the minority class ("Severe"), resulting in a balanced training set with 108 patients in each class. This step was crucial for mitigating potential biases that could arise from class imbalance.

To further refine the feature set, feature selection was performed using a Logistic Regression model with Elastic Net regularization, optimized through Grid Search. Elastic Net, which combines L1 and L2 penalties, was chosen for its ability to handle feature selection by shrinking less relevant coefficients to zero. The Grid Search optimized several hyper-parameters, including the regularization strength and the ratio between L1 and L2 penalties. After cross-validation, features with non-zero coefficients in the final model were retained, resulting in a reduced dataset containing only the most relevant features.

For the left eye dataset, Random Forest, XGBoost, and KNN each selected nine features: the four-sector measurements of the optic disc (Temporal, Superior, Nasal, Inferior), Superior (GCL+IPL), ParaInferiorNasal (NFL+GCL+IPL), Central (ILM-RPE), Superior-Inferior

Asymmetry, and Temporal-Nasal Asymmetry. SVM retained 13 features, including these core features but also adding CupVolume (ONHParameters), ParaInferiorNasal (GCL+IPL), Superior (NFL+GCL+IPL) Variance, Ratio ILM-BM Para-Peri, and Age.

For the right eye dataset, Random Forest and KNN retained seven features: Temporal and Superior sector measurements of the Optic disc, Volume (ILM_RPE), Age, Superior-Inferior Asymmetry, Temporal-Nasal Asymmetry, and Superior NFL-GCL-IPL Variance. XGBoost retained nine features, overlapping with Random Forest but adding CupVolume (ONHParameters) and Rim-to-Cup Volume Ratio. SVM retained the largest subset of 29 features, including additional structural features like DiscArea (ONHParameters) and demographic indicators such as Patient Sex and Ethnic Group.

5.3 Model Performance

The performance metrics for each model evaluated on both the left and right eye datasets are summarized in Table 7. These metrics, which include precision, recall, F1-score, and accuracy, are averaged across both classes and provide a comprehensive assessment of the models' ability to classify the severity of multiple sclerosis (MS). The results below are based on evaluations conducted on the test set, ensuring an unbiased assessment of model performance.

The Random Forest (RF) model demonstrated consistent performance across both datasets. The confusion matrix on the left eye dataset shows that the model correctly identified 37 "Not Severe" cases (TN) and 14 "Severe" cases (TP). However, it misclassified 9 "Not Severe" cases as "Severe" (FP) and 9 "Severe" cases as "Not Severe" (FN). This balanced performance is reflected in its precision, recall, and F1-score of 0.74, indicating that the model performed equally well in identifying both classes. On the right eye dataset, Random Forest correctly classified 45 "Not Severe" cases (TN) and 5 "Severe" cases (TP). However, the model showed a drop in sensitivity, as it misclassified 18 "Severe" cases as "Not Severe" (FN), while only 1 "Not Severe" case was incorrectly identified as "Severe" (FP). This resulted in a lower F1-score of 0.67, with recall reduced to 0.72.

Model	**Dataset**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Random Forest	Left Eye	0.74	0.74	0.74	0.74
	Right Eye	0.75	0.72	0.67	0.72
SVM	Left Eye	0.73	0.72	0.73	0.72
	Right Eye	0.79	0.70	0.60	0.70
XGBoost	Left Eye	0.72	0.72	0.72	0.72
	Right Eye	0.65	0.68	0.63	0.68
KNN	Left Eye	0.71	0.71	0.71	0.71
	Right Eye	0.75	0.74	0.70	0.74

Table 7: Performance metrics for the models evaluated on the left and right eye datasets.

The Support Vector Machine (SVM) model showed competitive performance. On the left eye dataset, the confusion matrix indicates that the model correctly classified 36 "Not Severe" cases (TN) and 14 "Severe" cases (TP). However, 10 "Not Severe" cases were misclassified as "Severe" (FP), and 9 "Severe" cases were misclassified as "Not Severe" (FN). This performance aligns with

its F1-score of 0.73 and accuracy of 0.72. On the right eye dataset, SVM correctly identified 46 "Not Severe" cases (TN) but only 2 "Severe" cases (TP). The model misclassified a substantial 21 "Severe" cases as "Not Severe" (FN), indicating a significant drop in sensitivity to severe cases. Despite this, no "Not Severe" cases were misclassified as "Severe" (FP), which explains its high precision of 0.79, though the recall dropped to 0.70.

XGBoost's performance was generally consistent but slightly lower compared to other models. On the left eye dataset, the confusion matrix shows that the model correctly identified "Not Severe" cases (TN) and 12 "Severe" cases (TP), while 8 "Not Severe" cases were misclassified as "Severe" (FP) and 11 "Severe" cases were misclassified as "Not Severe" (FN). This resulted in an F1-score of 0.72, reflecting moderate performance. On the right eye dataset, the model correctly classified 42 "Not Severe" cases (TN) and 5 "Severe" cases (TP). However, it misclassified 18 "Severe" cases as "Not Severe" (FN) and 4 "Not Severe" cases as "Severe" (FP). The increase in both false positives and false negatives contributed to its lower F1-score of 0.63 and overall accuracy of 0.68.

The KNN model demonstrated stable performance across both datasets. On the left eye dataset, the confusion matrix reveals that the model correctly identified 36 "Not Severe" cases (TN) and 15 "Severe" cases (TP). However, it misclassified 10 "Not Severe" cases as "Severe" (FP) and 10 "Severe" cases as "Not Severe" (FN). This resulted in an F1-score of 0.71, with precision and recall both at 0.71, and an overall accuracy of 0.71. On the right eye dataset, the model correctly classified 44 "Not Severe" cases (TN) and 7 "Severe" cases (TP). However, it misclassified 16 "Severe" cases as "Not Severe" (FN) and 2 "Not Severe" cases as "Severe" (FP). This performance yielded an F1-score of 0.70, with slightly higher precision at 0.75 and recall at 0.74, and an accuracy of 0.74.

The confusion matrices for each model are provided in Figures 3 and 4. Random Forest and KNN showed balanced classification results with relatively fewer false positives and false negatives, particularly on the left eye dataset. SVM exhibited strong precision, especially on the right eye dataset, but struggled with recall, reflecting a difficulty in identifying severe cases. XGBoost showed moderate performance, with higher misclassification rates for both false positives and false negatives, particularly on the right eye dataset.

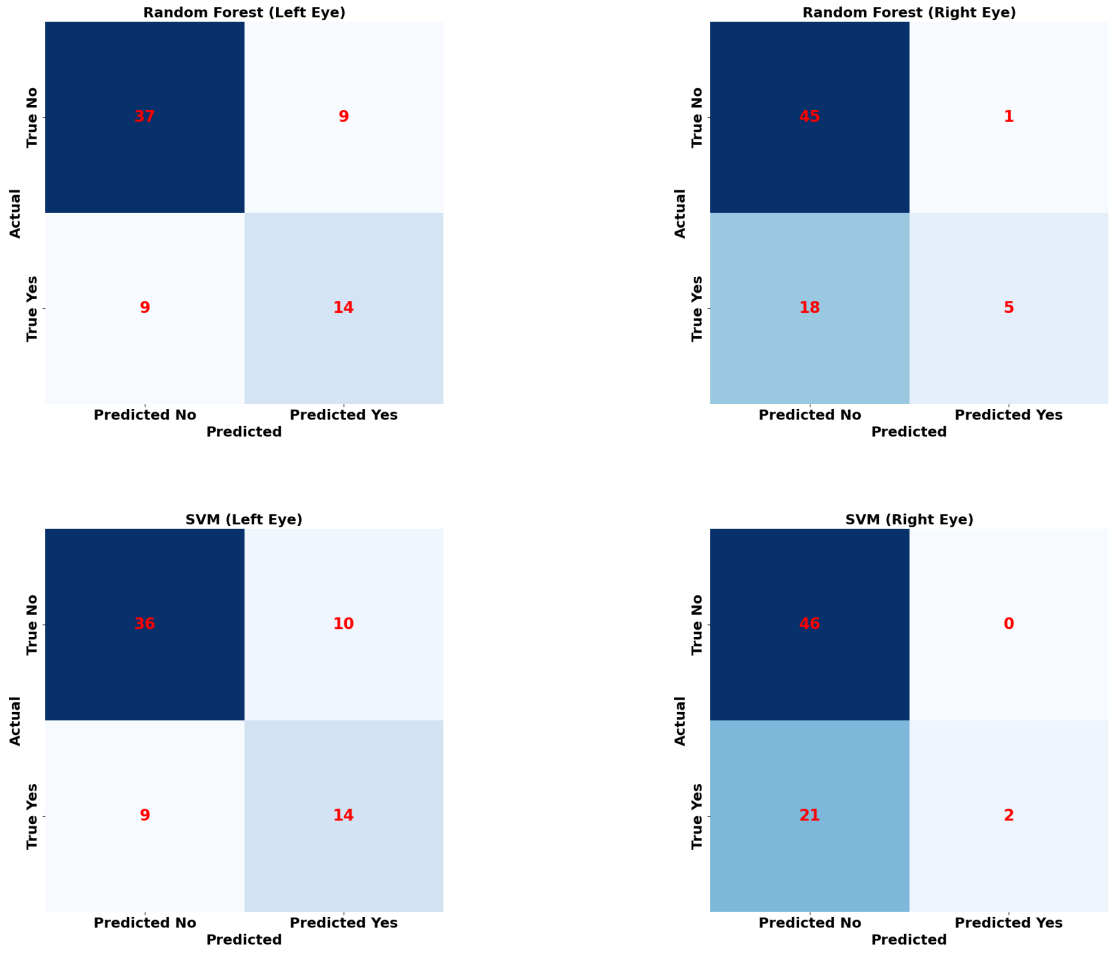


Figure 3: Confusion Matrices for Models on Left Eye Datasets

For the Right Eye dataset, RF and KNN both demonstrated the strongest performance, each achieving an AUC of 0.74. XGBoost followed with an AUC of 0.70, showing consistent but slightly lower performance compared to the top models. SVM had the lowest performance with an AUC of 0.67. These results indicate that RF consistently outperformed other models on both datasets, while SVM struggled the most in both contexts.

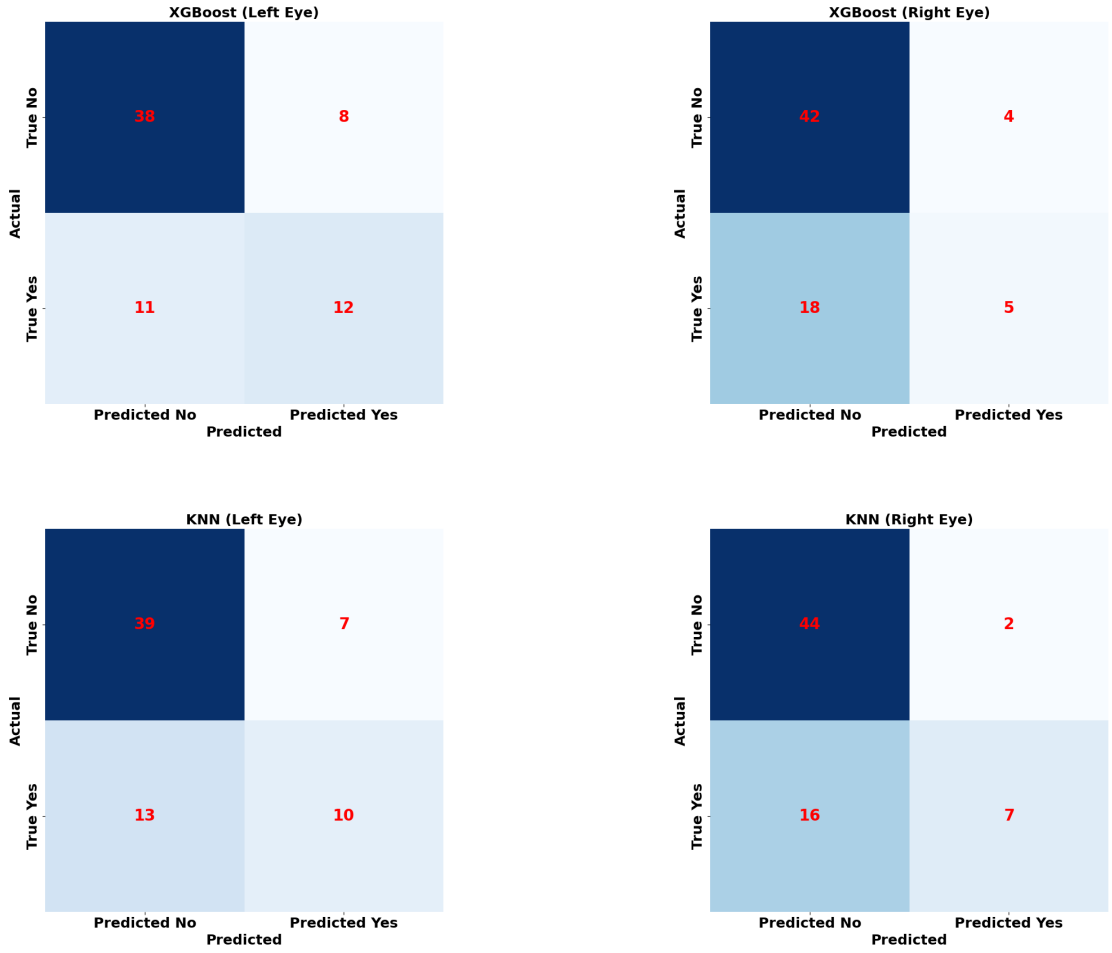
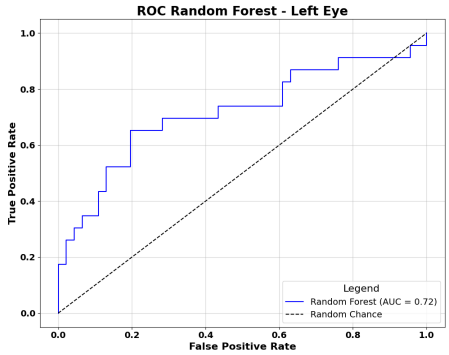
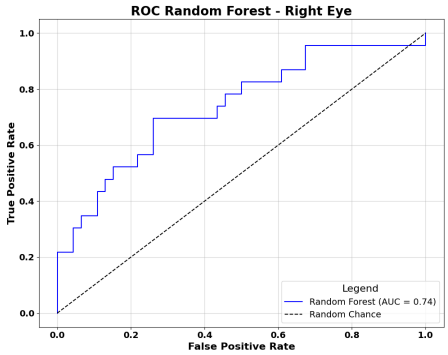
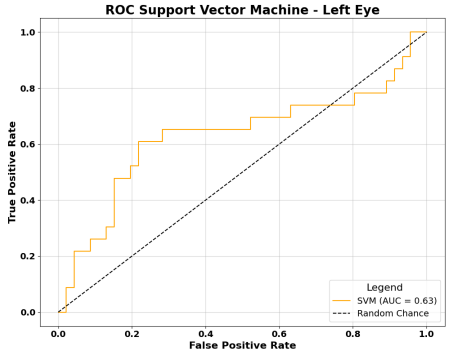


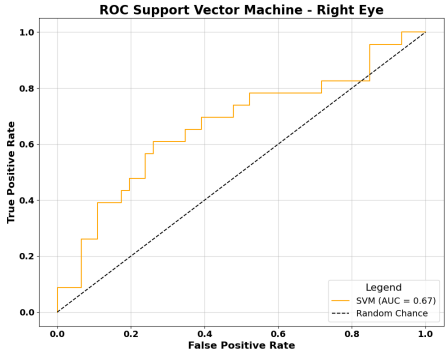
Figure 4: Confusion Matrices for Models on Right Eye Datasets

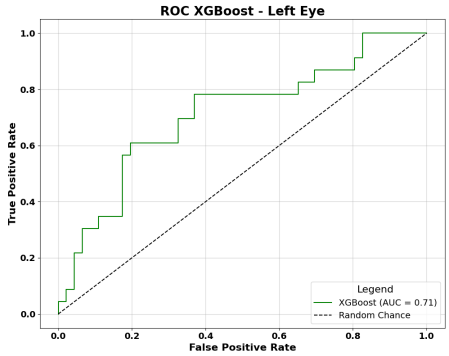
The Receiver Operating Characteristic (ROC) curves in Figure 5 highlight the performance of the models across the datasets. For the Left Eye dataset, the Random Forest (RF) model achieved the highest area under the curve (AUC) at 0.72, closely followed by XGBoost at 0.71. KNN demonstrated moderate performance with an AUC of 0.67, while SVM had the lowest classification capability on this dataset, achieving an AUC of 0.63.

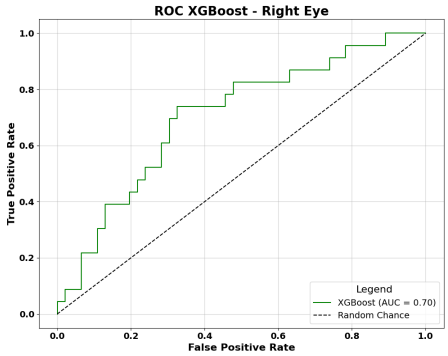


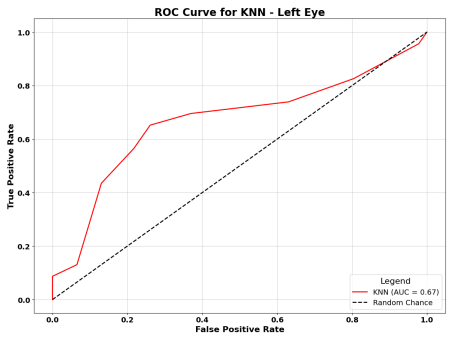












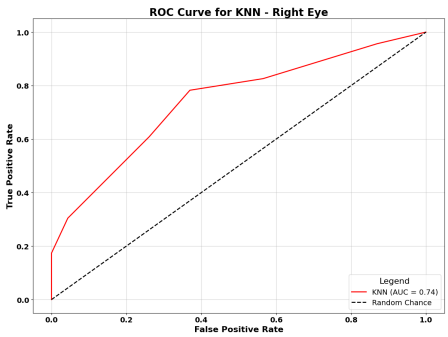


Figure 5: ROC Curves for Models on Left and Right Eye Datasets

While the overall results presented earlier provided a broad view of model performance, careful interpretation is required. The results discussed below are derived by evaluating the models specifically on the severe class, with metrics i.e., precision, recall, and F1-score calculated exclusively for this class rather than averaging performance across both classes. A closer analysis of the severe cases, as detailed in Table 8, highlights notable variability and significant limitations in the models' ability to effectively distinguish severe cases.

For the left eye, the random forest model demonstrated the highest balanced accuracy of 0.7065, with all metrics for the severe class (precision, recall, and F1 score) at 0.6087, indicating consistent moderate performance in identifying severe cases. The support vector machine (SVM) had a balanced accuracy of 0.6957, and F1 score of 0.5957 for the severe class, with precision and recall values of 0.5833 and 0.6087, respectively. The XGBoost model and k-nearest neighbors (KNN) both had a balanced accuracy of 0.6739; however, XGBoost underperformed in recall (0.5217) for severe cases, resulting in an F1 score of 0.5581 while the KNN model, showed consistent metrics across precision, recall, and F1 score at 0.5652.

Model	**Dataset**	**Precision**	**Recall**	**F1-Score**	**Balanced Accuracy**
Random Forest	Left Eye	0.6087	0.6087	0.6087	0.7065
	Right Eye	0.8333	0.2174	0.3448	0.5978
SVM	Left Eye	0.5833	0.6087	0.5957	0.6957
	Right Eye	1.0000	0.0870	0.1600	0.5435
XGBoost	Left Eye	0.6000	0.5217	0.5581	0.6739
	Right Eye	0.5556	0.2174	0.3125	0.5652
KNN	Left Eye	0.5652	0.5652	0.5652	0.6739
	Right Eye	0.7778	0.3043	0.4375	0.6304

Table 8: Performance metrics for the models evaluated on the left and right eye datasets for the severe class

For the right eye, the overall results showed reduced performance across all models compared to the left eye. The random forest model, despite achieving a balanced accuracy of 0.5978, showed a poor performance on the recall for severe cases at 0.2174, although it attained a high precision of 0.8333 with the F1 score of only 0.3448, showing limitation in identifying severe cases. The SVM model had the lowest balanced accuracy of 0.5435 and struggled significantly with severe cases, achieving perfect precision (1.0000) but a very low recall of 0.0870, with an F1 score of 0.1600. XGBoost exhibited slightly better performance with a balanced accuracy of 0.5652 and an F1 score of 0.3125 for severe cases, though its recall (0.2174) remained low. The KNN model achieved the highest balanced accuracy for the right eye at 0.6304, with moderate performance in identifying severe cases, reflected by a precision of 0.7778, recall of 0.3043, and F1 score of 0.4375.

5.4 Feature Importance

Feature importance analysis revealed critical contributions of specific features to classifying multiple sclerosis (MS) severity. Distinct patterns emerged when analyzing left and right eye datasets, highlighting key clinical and OCT-derived features.

For dataset 1, the Random Forest model identified retinal features from the superior and temporal optic disc sectors, Central ILM-RPE, and ParaInferiorNasal (NFL+GCL+IPL) as significant predictors. These findings indicate the role of central retinal thickness and nasal GCL+IPL thinning in MS progression, consistent with Shi et al., 2019.

The SVM model showed the temporal and superior sectors of the optic disc and Temporal-Nasal Asymmetry as important predictors of severity. Asymmetry measures are particularly valuable for capturing localized retinal changes, which are characteristic of MS-related neurodegeneration. This is supported by research demonstrating that inter-eye differences in retinal measurements can serve as diagnostic indicators for MS (Petzold et al., 2021).

Rank	**Random Forest**	**SVM**
1	Superior (Four Sectors)	Temporal-Nasal Asymmetry
2	Temporal (Four Sectors)	Temporal (Four Sectors)
3	Central (ILM-RPE)	Nasal (Four Sectors)
4	ParaInferiorNasal (NFL+GCL+IPL)	Superior (Four Sectors)
5	Superior-Inferior Asymmetry	ParaInferiorNasal (NFL+GCL+IPL)
6	Superior (GCL+IPL)	Superior-Inferior Asymmetry (NFL+GCL+IPL)
7	Inferior (Four Sectors)	Superior (GCL+IPL)
8	Nasal (Four Sectors)	Central (ILM-RPE)
9	Temporal-Nasal Asymmetry	Inferior (Four Sectors)
Rank	**XGBoost**	**KNN**
1	ParaInferiorNasal (NFL+GCL+IPL)	Inferior (Four Sectors)
2	Superior (Four Sectors)	Temporal (Four Sectors)
3	Temporal (Four Sectors)	Superior (Four Sectors)
4	Central (ILM-RPE)	Superior-Inferior Asymmetry
5	Superior (GCL+IPL)	Nasal (Four Sectors)
6	Superior-Inferior Asymmetry	Temporal-Nasal Asymmetry
7	Inferior (Four Sectors)	ParaInferiorNasal (NFL+GCL+IPL)
8	Temporal-Nasal Asymmetry	Superior (GCL+IPL)
9	Nasal (Four Sectors)	Central (ILM-RPE)

Table 9: Top predictive features identified by the models using the left eye dataset.

XGBoost placed significant importance on regions like the ParaInferiorNasal (NFL+GCL+IPL), superior sectors, and temporal sectors. These findings show the relevance of inferior nasal regions, which may exhibit signs of MS-related retinal damage as highlighted in García Mesa et al., 2023 which showed progressive thinning of the GCIPL is associated with MS.

The KNN model relied on features related to retinal sectors, i.e., the inferior, temporal, and superior regions of the optic disc. The top predictive features using dataset 1 are showed in Table 9.

For the dataset 2, the Random Forest model identified the temporal and superior sectors of the optic disc, as well as retinal volume (ILM-RPE), as the most important features. The SVM model identified regions, including the Superior (GCL+IPL) and PeriInferiorNasal (GCL+IPL), as key features. These features relating to the ganglion cell and inner plexiform layers, have

shown in previous research that they are associated with MS as highlighted in Garcia-Martin et al., 2014.

XGBoost showed clinical features, particularly Age, alongside superior sectors of the optic disc and structural variance metrics were the most important. This suggests that age, shows cumulative disease burden, and also interacts with structural changes to influence disease severity. The focus on variance metrics also reflects the model's ability to detect subtle and uneven retinal changes that are a characteristic of MS. The KNN model for the right eye dataset ranked structural features such as the superior sectors, retinal volume, and asymmetry metrics as most important, according to Petzold et al., 2021 these differences in retinal measurements can serve as diagnostic indicators for MS. The top predictive features using dataset 2 are indicated in Table 10.

Rank	**Random Forest**	**SVM**
1	Temporal (Four Sectors)	Superior (GCL+IPL)
2	Volume (ILM-RPE)	PeriInferiorNasal (GCL+IPL)
3	Superior (Four Sectors)	ParaInferiorNasal (GCL+IPL)
4	Age	Temporal (Four Sectors)
5	Superior (NFL+GCL+IPL) Variance	Superior (NFL+GCL+IPL) Variance
6	Superior-Inferior Asymmetry	Temporal(NFL+GCL+IPL) Variance
7	Temporal-Nasal Asymmetry	Volume (ILM_RPE)
Rank	**XGBoost**	**KNN**
1	Age	Superior (Four Sectors)
2	Superior (Four Sectors)	Volume (ILM-RPE)
3	Superior (NFL+GCL+IPL) Variance	Temporal-Nasal Asymmetry
4	Temporal (Four Sectors)	Age
5	Superior-Inferior Asymmetry	Temporal (Four Sectors)
6	Rim-to-Cup Volume Ratio	Superior-Inferior Asymmetry
7	CupVolume (ONH Parameters)	Superior (NFL+GCL+IPL) Variance

Table 10: Top predictive features identified by the models using the right eye dataset.

Among the models evaluated, the Random Forest algorithm showed the best overall performance for both left and right eye datasets. For the left eye, the optimal model parameters included 100 estimators, a minimum of 18 samples per split, 8 samples per leaf, and no restriction on tree depth. This configuration achieved an F1-score and accuracy of 0.74, reflecting a strong balance between precision and recall. For the right eye, the best-performing Random Forest model utilized 75 estimators, a minimum of 12 samples per split, 8 samples per leaf, and a maximum tree depth of 20. Although its performance was slightly lower, with an F1-score of 0.67 and accuracy of 0.72, it remained the top model for this dataset.

The differences observed between the left and right eyes in the results are likely attributable to the asymmetric progression of Multiple Sclerosis (MS), a characteristic feature of the disease. MS-related damage is inherently multifocal and localized, often resulting in varying degrees of retinal thinning and structural alterations between the eyes. This asymmetry is primarily driven by the random distribution of lesions in the central nervous system, including the optic nerves, which

may differentially affect the left and right eyes. Optic neuritis, a common early manifestation of MS, frequently occurs unilaterally, leading to more pronounced damage in one eye compared to the other. This phenomenon has been extensively documented in studies such as Saidha et al., 2015. Other factors, such as natural biological variability, differences in treatment responses, and technical variations in OCT measurements, may further contribute to these discrepancies.

6 Discussion

This study aimed to develop machine learning models to analyze OCT data for improving diagnostic accuracy in MS patients. Additionally, it sought to review the literature on OCT measurement techniques, emphasizing retinal structures, variations across manufacturers, and existing segmentation methods for OCT images.

The results highlight that the models used have the ability to classify MS severity, however, they face significant difficulty, particularly in identifying severe cases, which shows the need for further development before these models can be considered for clinical application and use.

The overall performance metrics based on both classes indicate that Random Forest (RF) and K-Nearest Neighbors (KNN) showed consistent performance across the left and right eye datasets. For the left eye, RF achieved precision, recall, F1-score, and accuracy of 0.74, while KNN recorded values of 0.71. Both models demonstrated a moderate performance, with RF slightly performing better than KNN. Support Vector Machine (SVM) performed similarly to RF, with precision of 0.73, recall of 0.72, and accuracy of 0.72, while XGBoost showed slightly lower results (precision, recall, F1-score, and accuracy of 0.72).

In the right eye dataset, all models faced challenges. RF had a precision of 0.75 and recall of 0.72, but its F1-score dropped to 0.67, indicating difficulties in identifying all true positives. KNN performed slightly better with precision of 0.75, recall of 0.74, and F1-score of 0.70, but still struggled with severe cases. SVM achieved the highest precision (0.79), but had recall of 0.70 and F1-score of 0.60. XGBoost had the lowest performance with a precision of 0.65, recall of 0.68, and F1-score of 0.63. Despite the use of SMOTE to address class imbalance, all models struggled to accurately identify severe cases in the right eye, potentially due to the dataset's inability to capture subtle structural differences in OCT features related to MS severity. Given that all patients in the study had MS, such subtle differences between severe and non-severe cases are expected. Introducing a control group could have potentially enhanced the distinction of these OCT features and improve classification outcomes.

The performance of the models on the severe class revealed significant weaknesses. When focusing on the right eye dataset, Random Forest (RF) showed high precision (0.83) but had poor recall (0.22), resulting in a low F1-score of 0.34. This suggests RF could identify severe cases but missed many true positives. Similarly, Support Vector Machine (SVM) had perfect precision (1.00) but an extremely low recall (0.09), leading to a very low F1-score of 0.16, suggesting that this model often failed to correctly identify severe cases.

XGBoost and KNN also struggled with recall for severe cases. XGBoost had a precision of 0.56 and recall of 0.22, with an F1-score of 0.31, while KNN had better precision (0.78) and recall (0.30), resulting in an F1-score of 0.44. These results highlight that all models had difficulty identifying severe cases reliably, a critical limitation for clinical decision-making. Despite the

use of SMOTE, the models were unable to capture the subtle features linked to severe MS cases, emphasizing the need for improvements in recall to support accurate detection of severe cases in clinical settings.

A key question arising from these findings is whether OCT data alone can reliably differentiate severe from non-severe cases. The performance observed, particularly for the severe class, suggests that OCT data, while valuable, may not be sufficient as a standalone diagnostic tool. The results indicate that these models have the potential but are not yet ready for standalone use in clinical practice. With further refinement and the adoption of deep learning models that incorporate advanced image recognition techniques specific to OCT data, there is potential to improve performance. Such advancements could enable these models to be effectively used within an integrated diagnostic framework. Rather than functioning independently, these models may be more impactful when combined with other clinical tools and data sources, such as longitudinal patient records, advanced imaging techniques, and clinical biomarkers. This integrated approach could provide a more comprehensive understanding of MS severity and enhance diagnostic accuracy.

The performance differences between the left and right eye datasets for severe cases were notable. For example, Random Forest achieved a balanced accuracy of 0.7065 on the left eye dataset but dropped to 0.5978 for the right eye. These differences may be attributed to biological or clinical factors, as MS often presents asymmetrically, with retinal damage being more pronounced in one eye. Alternatively, the discrepancies could be due to the dataset's under-representation of severe cases, which might have biased the models towards the majority class. Additionally, the models' limited sensitivity to subtle structural variations in the OCT data highlights the need for larger datasets and enhanced training to improve performance and address discrepancies effectively. Incorporating domain-specific knowledge into the feature selection process could also help capture more nuanced patterns related to MS severity. Examining and including inter-eye relationships may provide additional insights into MS-related retinal changes, thereby enhancing the models' generalizability, particularly given the observed differences in performance between the left and right eye datasets in this study.

The feature importance analysis from the fitted models identified structural retinal features as critical predictors of MS severity. Specifically, measurements from the superior and temporal sectors, central retinal thickness (Central ILM-RPE), and nasal regions related to the combined NFL, GCL, and IPL layers (NFL+GCL+IPL) were highly influential. These findings are consistent with previous studies, emphasizing the importance of these features in understanding retinal changes associated with MS progression and diagnostics.

This study differs from most previous research by focusing exclusively on an MS-only cohort and stratifying patients by severity rather than comparing MS patients to healthy controls. While the absence of a control group limits direct comparisons to other studies, the findings highlight the clinical relevance of intra-disease classification. Unlike studies aimed at diagnosing MS, this research provides actionable insights into disease progression, an area of significant clinical importance, by identifying retinal features that differentiate between severe and non severe patients.

Comparable studies in the literature have reported similar diagnostic performance using OCT-based ML models. For instance, García Mesa et al., 2023 achieved an 87.3% accuracy using RF, k-NN, and SVM for distinguishing MS patients from controls, while Montolío et al., 2022 and Palomar et al., 2019 reported diagnostic accuracies of 87.7% and 95.8%, respectively, with

advanced OCT devices like Spectralis and Cirrus HD-OCT. The superior performance in these studies likely reflects the advantages of higher-resolution imaging, larger datasets, and the inclusion of advanced classifiers, such as convolutional neural networks (CNNs), as demonstrated by Garcia-Martin et al., 2021, who achieved near-perfect sensitivity and specificity (98%).

In summary, this study demonstrates the potential of machine learning models in leveraging OCT data to classify MS severity, while also highlighting key limitations that must be addressed for clinical application. The observed challenges, particularly in identifying severe cases across the models used, emphasizes the need for enhanced model sensitivity, larger and more balanced datasets, and the integration of advanced imaging techniques. These improvements, combined with approaches that incorporate other clinical tools and biomarkers, could significantly enhance the utility of these models in providing a comprehensive assessment of MS severity.

7 Possible Drawback of the Methods

The application of machine learning in this study faced several limitations. A small sample size, combined with missing EDSS scores, reduced model robustness and generalizability. The absence of a control group further hindered validation and the ability to establish comparative baselines. Class imbalances, with under-representation of certain EDSS categories, biased the models toward majority classes and reduced sensitivity for minority classes.

A significant challenge in broader clinical application lies in the models' limited transferability. The dataset, specific to a subset of MS patients, may not represent diverse demographics or clinical settings. Variations in imaging protocols, device manufacturers, and population characteristics could hinder application of models effectively in other settings, For example, differences in OCT devices can result in inconsistent image resolutions and feature measurements, complicating model applicability. Additionally, the lack of a healthy control group limits the assessment of the models' performance in heterogeneous populations.

8 Ethics, Societal Relevance, and Stakeholder Awareness

This study prioritized ethical considerations in handling patient data and implementing machine learning (ML) models in healthcare. Patient privacy was safeguarded through the anonymization of identifiers and secure, encrypted data storage, with controlled access ensuring compliance with data protection regulations.

Ethical considerations extended to the use of ML models, emphasizing transparency, fairness, and the minimization of bias. The model training process was carefully evaluated to ensure no biases were introduced from clinical variables, maintaining equitable treatment across patient subgroups.

The integration of ML in diagnosing Multiple Sclerosis (MS) has significant societal and clinical implications. Early diagnosis through advanced OCT-based ML models enables timely treatment, potentially improving outcomes by slowing disease progression. This is especially valuable in resource-limited settings, where early access to specialized care is limited. By improving diagnostic accuracy and reducing misdiagnoses, these tools can optimize healthcare workflows, allowing providers to focus on complex cases and efficiently allocate resources.

The outcomes of this research have direct implications for various stakeholders, including neurologists, healthcare providers, and patients. For neurologists and clinicians, the study offers a data-driven approach to enhance diagnostic precision. OCT features identified as possible predictive biomarkers can be tested in existing diagnostic frameworks, providing actionable insights that support clinical judgment.

The outcomes directly benefit neurologists, healthcare providers, and patients. For clinicians, the study offers a possible data-driven framework to enhance diagnostic precision, identifying actionable OCT features as potential biomarkers for MS severity. Healthcare institutions can leverage the scalability of AI-driven tools to enhance care quality while reducing costs. Patients benefit most from improved diagnostic accuracy, which builds trust, reduces uncertainty, and supports the development of personalized treatment plans, empowering them to make informed decisions about their healthcare.

9 Conclusion

This study demonstrates the potential of machine learning (ML) models to classify multiple sclerosis (MS) severity using OCT-derived features. Among the models evaluated, Random Forest (RF) showed the most robust and consistent performance, particularly for the left eye dataset, while Support Vector Machines (SVM) and k-Nearest Neighbors (KNN) also exhibited competitive results. Despite these advancements, challenges remain, particularly in classifying severe cases and ensuring model generalizability across diverse datasets.

Key findings highlight the importance of structural retinal features, such as measurements from the superior and temporal sectors, central retinal thickness, and asymmetry metrics, in predicting MS severity. These features align with known MS-related retinal changes and provide valuable insights into disease progression. However, the limited sensitivity of the models to severe cases shows the need for larger, balanced datasets and advanced techniques to enhance performance.

Standardization of OCT imaging protocols and segmentation tools is essential to address device-specific variability and facilitate cross-platform applicability. Proprietary devices like Canon HS100 OCT, while powerful, require further compatibility with open-source tools to enable broader accessibility. Innovations in deep learning and explainable AI offer promising pathways to improve segmentation accuracy and diagnostic transparency.

This study highlights the need for an integrated diagnostic framework that combines OCT data with clinical variables and biomarkers to provide a comprehensive understanding of MS severity. While the ML models developed here show promise, further research should focus on improving sensitivity for severe cases, incorporating control groups for better benchmarking, and leveraging advanced imaging techniques to achieve clinically viable solutions.

10 Ideas for Future Research

Future research should aim to address the limitations of the current study by expanding to larger, more diverse datasets. Longitudinal studies tracking patients over extended periods would be particularly valuable in assessing the prognostic value of OCT features. Such studies could help discern how structural changes in retinal layers correlate with disease progression and

EDSS scores over time, thus improving model interpretability and clinical utility. Incorporating healthy control groups would provide critical comparative insights to enhance the robustness of findings.

To overcome limitations related to manual feature selection and extraction, future work should explore advanced AI methods, such as convolutional neural networks (CNNs) or other deep learning architectures. These approaches can automate segmentation and feature extraction processes, potentially capturing subtler patterns in OCT images. Furthermore, integrating multi-modal data combining clinical, imaging, and even genetic information could lead to more comprehensive models. This holistic approach may significantly enhance diagnostic accuracy and predictive power.

References

Abe, R. Y., Gracitelli, C. P., and Medeiros, F. A. (2015). The use of spectral-domain optical coherence tomography to detect glaucoma progression. *The Open Ophthalmology Journal*, 9:78.

Amoaku, W. M., Ghanchi, F., Bailey, C., Banerjee, S., Banerjee, S., Downey, L., Gale, R., Hamilton, R., Khunti, K., Posner, E., et al. (2020). Diabetic retinopathy and diabetic macular oedema pathways and management: Uk consensus working group. *Eye*, 34(Suppl 1):1–51.

Bhargava, P., Lang, A., Al-Louzi, O., Carass, A., Prince, J., Calabresi, P. A., and Saidha, S. (2015). Applying an open-source segmentation algorithm to different oct devices in multiple sclerosis patients and healthy controls: implications for clinical trials. *Multiple sclerosis international*, 2015(1):136295.

Brautaset, R., Birkeldh, U., Frehr Alstig, P., Wiken, P., and Nilsson, M. (2016). Repeatability using automatic tracing with canon oct-hs100 and zeiss cirrus hd-oct 5000. *PLoS One*, 11(2):e0149138.

Canon (2020). *OCT HS100 User Manual*. Canon User Manual. Available at https://www.cmi.sk/sites/default/files/oct-hs100.pdf.

Cavaliere, C., Vilades, E., Alonso-Rodríguez, M. C., Rodrigo, M. J., Pablo, L. E., Miguel, J. M., López-Guillén, E., Morla, E. M. S., Boquete, L., and Garcia-Martin, E. (2019). Computer-aided diagnosis of multiple sclerosis using a support vector machine and optical coherence tomography features. *Sensors*, 19(23):5323.

Chauhan, B. C. and Burgoyne, C. F. (2013). From clinical examination of the optic disc to clinical assessment of the optic nerve head: a paradigm change. *American journal of ophthalmology*, 156(2):218–227.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Cirrus (2020). *How to Read Cirrus Reports*. Cirrus User Manual. Available at https://www.asta-tec.com/wp-content/uploads/2020/05/how_to_read_cirrus_reports-2.pdf.

Clinic, C. (2024). Definition of oct. https://my.clevelandclinic.org/health/diagnostics/optical-coherence-tomography-oct. Accessed: (December, 26th 2024).

Dufour, P. A., Ceklic, L., Abdillahi, H., Schroder, S., De Dzanet, S., Wolf-Schnurrbusch, U., and Kowal, J. (2012). Graph-based multi-surface segmentation of oct data using trained hard and soft constraints. *IEEE transactions on medical imaging*, 32(3):531–543.

El Ayoubi, N. K., Ismail, A., Fahd, F., Younes, L., Chakra, N. A., and Khoury, S. J. (2024). Retinal optical coherence tomography measures in multiple sclerosis: a systematic review and meta-analysis. *Annals of Clinical and Translational Neurology*, 11(9):2236–2253.

Everett, M. J. and Oakley, J. D. (2015). Automated analysis of the optic nerve head: measurements, methods and representations. US Patent 9,101,293.

Flores, R., Carneiro, Â., Tenreiro, S., and Seabra, M. C. (2021). Retinal progression biomarkers of early and intermediate age-related macular degeneration. *Life*, 12(1):36.

Garcia-Martin, E., Ortiz, M., Boquete, L., Sánchez-Morla, E. M., Barea, R., Cavaliere, C., Vilades, E., Orduna, E., and Rodrigo, M. J. (2021). Early diagnosis of multiple sclerosis by oct analysis using cohen's d method and a neural network as classifier. *Computers in Biology and Medicine*, 129:104165.

Garcia-Martin, E., Polo, V., Larrosa, J. M., Marques, M. L., Herrero, R., Martin, J., Ara, J. R., Fernandez, J., and Pablo, L. E. (2014). Retinal layer segmentation in patients with multiple sclerosis using spectral domain optical coherence tomography. *Ophthalmology*, 121(2):573–579.

García Mesa, P., Rojas Lozano, P., Díaz Gutiérrez, N., Cadena Santoyo, M., Beltrán Carrero, A. J., Gómez Valverde, J. J., et al. (2023). Evaluation of machine learning algorithms and relevant biomarkers for the diagnosis of multiple sclerosis based on optical coherence tomography.

Geevarghese, A., Wollstein, G., Ishikawa, H., and Schuman, J. S. (2021). Optical coherence tomography and glaucoma. *Annual review of vision science*, 7(1):693–726.

Goodin, D. S. (2014). The epidemiology of multiple sclerosis: insights to disease pathogenesis. *Handbook of clinical neurology*, 122:231–266.

Huang, D., Swanson, E. A., Lin, C. P., Schuman, J. S., Stinson, W. G., Chang, W., Hee, M. R., Flotte, T., Gregory, K., Puliafito, C. A., et al. (1991). Optical coherence tomography. *science*, 254(5035):1178–1181.

Kaushik, M. and Fraser, C. L. (2020). Optical coherence tomography in compressive lesions of the anterior visual pathway. *Annals of Eye Science*, 5:15–15.

Lee, C. S., Tyring, A. J., Deruyter, N. P., Wu, Y., Rokem, A., and Lee, A. Y. (2017). Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomedical optics express*, 8(7):3440–3448.

Liu, M. M., Wolfson, Y., Bressler, S. B., Do, D. V., Ying, H. S., and Bressler, N. M. (2014). Comparison of time-and spectral-domain optical coherence tomography in management of diabetic macular edema. *Investigative Ophthalmology & Visual Science*, 55(3):1370–1377.

Liu, X., Cao, J., Fu, T., Pan, Z., Hu, W., Zhang, K., and Liu, J. (2018). Semi-supervised automatic segmentation of layer and fluid region in retinal optical coherence tomography images using adversarial learning. *IEEE Access*, 7:3046–3061.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542.

Mitsch, C., Holzer, S., Wassermann, L., Resch, H., Urach, S., Kiss, B., Hommer, A., Vass, C., and Schmidt-Erfurth, U. (2019). Comparison of spectralis and cirrus spectral domain optical coherence tomography for the objective morphometric assessment of the neuroretinal rim width. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 257:1265–1275.

Montolío, A., CEGONino, J., Garcia-Martin, E., and Pérez del Palomar, A. (2022). Comparison of machine learning methods using spectralis oct for diagnosis and disability progression prognosis in multiple sclerosis. *Annals of biomedical engineering*, 50(5):507–528.

Montolío, A., Martín-Gallego, A., Cegoñino, J., Orduna, E., Vilades, E., Garcia-Martin, E., and Del Palomar, A. P. (2021). Machine learning in diagnosis and disability prediction of multiple sclerosis using optical coherence tomography. *Computers in Biology and Medicine*, 133:104416.

Morelle, O. (2023). eyepy.

Olbert, E. and Struhal, W. (2022). Retinal imaging with optical coherence tomography in multiple sclerosis: novel aspects. *Wiener Medizinische Wochenschrift*, 172(15):329–336.

Ometto, G., Moghul, I., Montesano, G., Hunter, A., Pontikos, N., Jones, P. R., Keane, P. A., Liu, X., Denniston, A. K., and Crabb, D. P. (2019). Relayer: a free, online tool for extracting retinal thickness from cross-platform oct images. *Translational vision science & technology*, 8(3):25–25.

Palomar, A. P. d., Cegoñino, J., Montolío, A., Orduna, E., Vilades, E., Sebastián, B., Pablo, L. E., and Garcia-Martin, E. (2019). Swept source optical coherence tomography to early detect multiple sclerosis disease. the use of machine learning techniques. *PLOS ONE*, 14(5):1–18.

Perry, J. and Fernandez, A. (2020). Eyeseg: Fast and efficient few-shot semantic segmentation. In *European Conference on Computer Vision (ECCV) Workshops*.

Petzold, A., Chua, S. Y., Khawaja, A. P., Keane, P. A., Khaw, P. T., Reisman, C., Dhillon, B., Strouthidis, N. G., Foster, P. J., Patel, P. J., et al. (2021). Retinal asymmetry in multiple sclerosis. *Brain*, 144(1):224–235.

Ramagopalan, S. V. and Sadovnick, A. D. (2011). Epidemiology of multiple sclerosis. *Neurologic clinics*, 29(2):207–217.

Ranschaert, E. R., Morozov, S., and Algra, P. R. (2019). *Artificial intelligence in medical imaging: opportunities, applications and risks*. Springer.

Rivas-Villar, D., Motschi, A. R., Pircher, M., Hitzenberger, C. K., Schranz, M., Roberts, P. K., Schmidt-Erfurth, U., and Bogunović, H. (2023). Automated inter-device 3d oct image registration using deep learning and retinal layer segmentation. *Biomedical Optics Express*, 14(7):3726–3747.

Rothman, A., Murphy, O. C., Fitzgerald, K. C., Button, J., Gordon-Lipkin, E., Ratchford, J. N., Newsome, S. D., Mowry, E. M., Sotirchos, E. S., Syc-Mazurek, S. B., et al. (2019).

Retinal measurements predict 10-year disability in multiple sclerosis. *Annals of Clinical and Translational Neurology*, 6(2):222–232.

Saidha, S., Al-Louzi, O., Ratchford, J. N., Bhargava, P., Oh, J., Newsome, S. D., Prince, J. L., Pham, D., Roy, S., Van Zijl, P., et al. (2015). Optical coherence tomography reflects brain atrophy in multiple sclerosis: a four-year study. *Annals of neurology*, 78(5):801–813.

Schneider, E., Zimmermann, H., Oberwahrenbrock, T., Kaufhold, F., Kadas, E. M., Petzold, A., Bilger, F., Borisow, N., Jarius, S., Wildemann, B., et al. (2013). Optical coherence tomography reveals distinct patterns of retinal damage in neuromyelitis optica and multiple sclerosis. *PloS one*, 8(6):e66151.

Shi, C., Jiang, H., Gameiro, G. R., Hu, H., Hernandez, J., Delgado, S., and Wang, J. (2019). Visual function and disability are associated with focal thickness reduction of the ganglion cell-inner plexiform layer in patients with multiple sclerosis. *Investigative ophthalmology & visual science*, 60(4):1213–1223.

Society, O. P. (2024). Retinal oct imaging. https://www.opsweb.org/page/RetinalOCT. Accessed: (December, 26th 2024).

Spectralis (2025). *SPECTRALIS Product Information*. Spectralis User Manual. Available at https://business-lounge.heidelbergengineering.com/us/en/products/spectralis/spectralis/.

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M. S., et al. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria. *The Lancet Neurology*, 17(2):162–173.

Tian, J., Varga, B., Tatrai, E., Fanni, P., Somfai, G. M., Smiddy, W. E., and Debuc, D. C. (2016). Performance evaluation of automated segmentation software on optical coherence tomography volume data. *Journal of biophotonics*, 9(5):478–489.

Wattjes, M. P., Ciccarelli, O., Reich, D. S., Banwell, B., de Stefano, N., Enzinger, C., Fazekas, F., Filippi, M., Frederiksen, J., Gasperini, C., et al. (2021). 2021 magnims–cmsc–naims consensus recommendations on the use of mri in patients with multiple sclerosis. *The Lancet Neurology*, 20(8):653–670.

Yuechuan, L. (2022). PyOCT.

Zahavi, O., Domínguez-Vicent, A., Brautaset, R., and Venkataraman, A. P. (2021). Evaluation of automated segmentation algorithm for macular volumetric measurements of eight individual retinal layer thickness. *Applied Sciences*, 11(3):1250.

Zeppieri, M., Marsili, S., Enaholo, E. S., Shuaibu, A. O., Uwagboe, N., Salati, C., Spadea, L., and Musa, M. (2023). Optical coherence tomography (oct): a brief look at the uses and technological evolution of ophthalmology. *Medicina*, 59(12):2114.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

Appendix

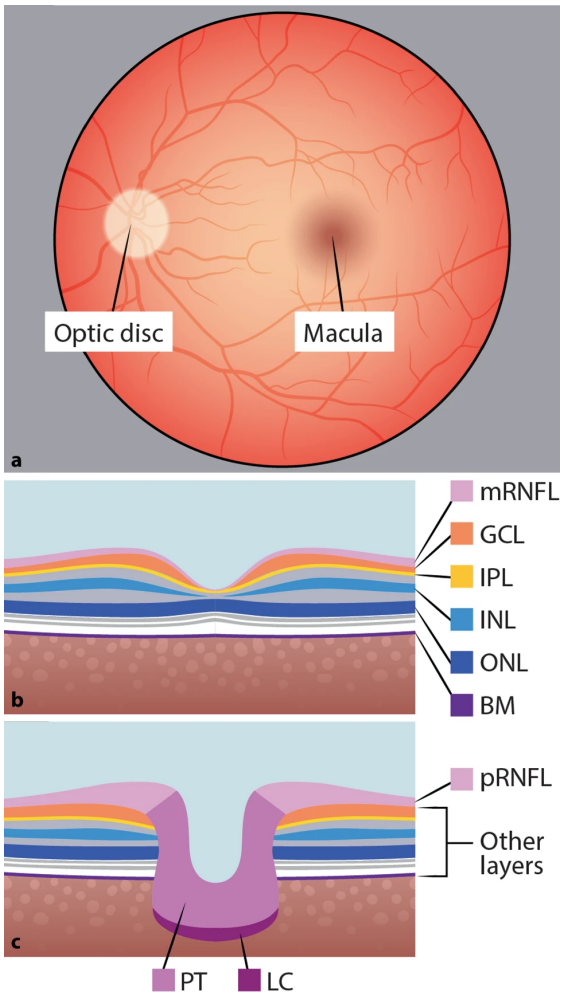


Figure 6: Visualization of retinal regions and their corresponding OCT scans: (a) Background image of the eye highlighting the macula and optic disc, (b) Optical Coherence Tomography (OCT) scan of the macula, (c) OCT scan of the peripapillary region. Key layers and structures: **mRNFL** – macular retinal nerve fiber layer, **GCL** – ganglion cell layer, **IPL** – inner plexiform layer, **INL** – inner nuclear layer, **ONL** – outer nuclear layer, **BM** – Bruch's membrane, **pRNFL** – peripapillary retinal nerve fiber layer, **PT** – prelaminar tissue, **LC** – lamina cribrosa,
Source: Olbert and Struhal, 2022

Code

The code for all the models fitted to the analysis can be found at the following GitHub repository: MS-Diagnostics

Tests and feature Engineering

```
# Extract all numeric features
features_to_test <- setdiff(colnames(numeric_features_df), "EDSS_
    Binary")

# Apply Shapiro-Wilk test to each feature
shapiro_results <- lapply(features_to_test, function(feature) {
  test_result <- shapiro.test(numeric_features_df[[feature]])
  list(
    Feature = feature,
    P_Value = test_result$p.value,
    W_Statistic = test_result$statistic
  )
})

# results
shapiro_results_df <- do.call(rbind, lapply(shapiro_results,
    function(result) {
  data.frame(
    Feature = result$Feature,
    P_Value = result$P_Value,
    W_Statistic = result$W_Statistic,
    Significance = ifelse(result$P_Value < 0.05, "Significant", "
    Non-Significant"),
    stringsAsFactors = FALSE
  )
}))

# Feature Engineering for Twelve Sectors
left <- left %>%
  mutate(
    # Quadrant Averages (optic Disc)
    Temporal_Quadrant = rowMeans(select(., `Temporal (
    TwelveSectors)`, `TemporalSuperiorTemporal (TwelveSectors)`, `
    TemporalInferiorTemporal (TwelveSectors)`), na.rm = TRUE),
    Superior_Quadrant = rowMeans(select(., `Superior (
    TwelveSectors)`, `SuperiorSuperiorTemporal (TwelveSectors)`, `
    SuperiorSuperiorNasal (TwelveSectors)`), na.rm = TRUE),
    Nasal_Quadrant = rowMeans(select(., `Nasal (TwelveSectors)`, `
    NasalSuperiorNasal (TwelveSectors)`, `NasalInferiorNasal (
```

```
    TwelveSectors)`), na.rm = TRUE),
     Inferior_Quadrant = rowMeans(select(., `Inferior (
    TwelveSectors)`, `InferiorInferiorNasal (TwelveSectors)`, `
    InferiorInferiorTemporal (TwelveSectors)`), na.rm = TRUE))


right <- right %>%
  mutate(
    # Quadrant Averages (optic Disc)
    Temporal_Quadrant = rowMeans(select(., `Temporal (
    TwelveSectors)`, `TemporalSuperiorTemporal (TwelveSectors)`, `
    TemporalInferiorTemporal (TwelveSectors)`), na.rm = TRUE),
     Superior_Quadrant = rowMeans(select(., `Superior (
    TwelveSectors)`, `SuperiorSuperiorTemporal (TwelveSectors)`, `
    SuperiorSuperiorNasal (TwelveSectors)`), na.rm = TRUE),
     Nasal_Quadrant = rowMeans(select(., `Nasal (TwelveSectors)`, `
    NasalSuperiorNasal (TwelveSectors)`, `NasalInferiorNasal (
    TwelveSectors)`), na.rm = TRUE),
     Inferior_Quadrant = rowMeans(select(., `Inferior (
    TwelveSectors)`, `InferiorInferiorNasal (TwelveSectors)`, `
    InferiorInferiorTemporal (TwelveSectors)`), na.rm = TRUE))

# Feature Engineering for Four Sectors
left <- left %>%
  mutate(
    # Asymmetry Features
    Superior_Inferior_Asymmetry = `Superior (FourSectors)` - `
    Inferior (FourSectors)`,
    Temporal_Nasal_Asymmetry = `Temporal (FourSectors)` - `Nasal (
    FourSectors)`)

right <- right %>%
  mutate(
    # Asymmetry Features
    Superior_Inferior_Asymmetry = `Superior (FourSectors)` - `
    Inferior (FourSectors)`,
    Temporal_Nasal_Asymmetry = `Temporal (FourSectors)` - `Nasal (
    FourSectors)`)

# Feature Engineering for ONH Parameters
left <- left %>%
  mutate(
    # Normalized Rim Area
    Normalized_Rim_Area = `RimArea (ONHParameters)` / `DiscArea (
    ONHParameters)`,
    # Cup-to-Disc Ratios
```

```
    Cup_to_Disc_Area_Ratio = `CDArea (ONHParameters)` / `DiscArea
   (ONHParameters)`,
    # Rim-to-Cup Ratios
    Rim_to_Cup_Volume_Ratio = `RimVolume (ONHParameters)` / `
   CupVolume (ONHParameters)`)

right <- right %>%
  mutate(
    # Normalized Rim Area
    Normalized_Rim_Area = `RimArea (ONHParameters)` / `DiscArea (
   ONHParameters)`,
    # Cup-to-Disc Ratios
    Cup_to_Disc_Area_Ratio = `CDArea (ONHParameters)` / `DiscArea
   (ONHParameters)`,
    # Rim-to-Cup Ratios
    Rim_to_Cup_Volume_Ratio = `RimVolume (ONHParameters)` / `
   CupVolume (ONHParameters)`)

# Feature Engineering for GCL, IPL, and NFL Layers
left <- left %>%
  mutate(

# Quadrant Variances for NFL_GCL_IPL
    Temporal_NFL_GCL_IPL_Variance = apply(select(., `
   ParaInferiorTemporal (NFL_GCL_IPL)`, `ParaSuperiorTemporal (NFL
   _GCL_IPL)`), 1, var, na.rm = TRUE),
    Nasal_NFL_GCL_IPL_Variance = apply(select(., `
   ParaInferiorNasal (NFL_GCL_IPL)`, `ParaSuperiorNasal (NFL_GCL_
   IPL)`), 1, var, na.rm = TRUE),
    Superior_NFL_GCL_IPL_Variance = apply(select(., `
   ParaSuperiorTemporal (NFL_GCL_IPL)`, `ParaSuperiorNasal (NFL_
   GCL_IPL)`), 1, var, na.rm = TRUE),
    Inferior_NFL_GCL_IPL_Variance = apply(select(., `
   ParaInferiorTemporal (NFL_GCL_IPL)`, `ParaInferiorNasal (NFL_
   GCL_IPL)`), 1, var, na.rm = TRUE))

right <- right %>%
 mutate(
# Quadrant Variances for NFL_GCL_IPL
    Temporal_NFL_GCL_IPL_Variance = apply(select(., `
   ParaInferiorTemporal (NFL_GCL_IPL)`, `ParaSuperiorTemporal (NFL
   _GCL_IPL)`), 1, var, na.rm = TRUE),
    Nasal_NFL_GCL_IPL_Variance = apply(select(., `
   ParaInferiorNasal (NFL_GCL_IPL)`, `ParaSuperiorNasal (NFL_GCL_
   IPL)`), 1, var, na.rm = TRUE),
```

```
    Superior_NFL_GCL_IPL_Variance = apply(select(., `
   ParaSuperiorTemporal (NFL_GCL_IPL)`, `ParaSuperiorNasal (NFL_
   GCL_IPL)`), 1, var, na.rm = TRUE),
    Inferior_NFL_GCL_IPL_Variance = apply(select(., `
   ParaInferiorTemporal (NFL_GCL_IPL)`, `ParaInferiorNasal (NFL_
   GCL_IPL)`), 1, var, na.rm = TRUE))

# Feature Engineering for ILM_RPE and ILM_BM
left <- left %>%
  mutate(
    # Total Volumes
    Total_ILM_RPE_Volume = rowSums(select(., `ParaTemporal (ILM_
   RPE)`, `PeriTemporal (ILM_RPE)`, `ParaNasal (ILM_RPE)`, `
   PeriNasal (ILM_RPE)`,
                                          `ParaSuperior (ILM_RPE)
   `, `PeriSuperior (ILM_RPE)`, `ParaInferior (ILM_RPE)`, `
   PeriInferior (ILM_RPE)`),
                                   na.rm = TRUE),
    Total_ILM_BM_Volume = rowSums(select(., `ParaTemporal (ILM_BM)
   `, `PeriTemporal (ILM_BM)`, `ParaNasal (ILM_BM)`, `PeriNasal (
   ILM_BM)`,
                                         `ParaSuperior (ILM_BM)`,
   `PeriSuperior (ILM_BM)`, `ParaInferior (ILM_BM)`, `PeriInferior
    (ILM_BM)`),
                                  na.rm = TRUE),
    # Regional Ratios
    Ratio_ILM_RPE_Central_Peripheral = `Central (ILM_RPE)` /
   rowMeans(select(., `ParaTemporal (ILM_RPE)`, `PeriTemporal (ILM
   _RPE)`, `ParaNasal (ILM_RPE)`, `PeriNasal (ILM_RPE)`, `
   ParaSuperior (ILM_RPE)`, `PeriSuperior (ILM_RPE)`, `
   ParaInferior (ILM_RPE)`, `PeriInferior (ILM_RPE)`), na.rm =
   TRUE),

    Ratio_ILM_BM_Para_Peri = rowMeans(select(., `ParaTemporal (ILM
   _BM)`, `ParaNasal (ILM_BM)`, `ParaSuperior (ILM_BM)`, `
   ParaInferior (ILM_BM)`),
                                     na.rm = TRUE) /
                              rowMeans(select(., `PeriTemporal (ILM
   _BM)`, `PeriNasal (ILM_BM)`, `PeriSuperior (ILM_BM)`, `
   PeriInferior (ILM_BM)`),
                                     na.rm = TRUE))
right <- right %>%
  mutate(
    # Total Volumes
    Total_ILM_RPE_Volume = rowSums(select(., `ParaTemporal (ILM_
   RPE)`, `PeriTemporal (ILM_RPE)`, `ParaNasal (ILM_RPE)`, `
```

```
    PeriNasal (ILM_RPE)`,
                                                `ParaSuperior (ILM_RPE)
    `, `PeriSuperior (ILM_RPE)`, `ParaInferior (ILM_RPE)`, `
    PeriInferior (ILM_RPE)`),
                                       na.rm = TRUE),

     Total_ILM_BM_Volume = rowSums(select(., `ParaTemporal (ILM_BM)
    `, `PeriTemporal (ILM_BM)`, `ParaNasal (ILM_BM)`, `PeriNasal (
    ILM_BM)`,
                                             `ParaSuperior (ILM_BM)`,
    `PeriSuperior (ILM_BM)`, `ParaInferior (ILM_BM)`, `PeriInferior
     (ILM_BM)`),
                                     na.rm = TRUE),
     # Regional Ratios
     Ratio_ILM_RPE_Central_Peripheral = `Central (ILM_RPE)` /
    rowMeans(select(., `ParaTemporal (ILM_RPE)`, `PeriTemporal (ILM
    _RPE)`, `ParaNasal (ILM_RPE)`, `PeriNasal (ILM_RPE)`, `
    ParaSuperior (ILM_RPE)`, `PeriSuperior (ILM_RPE)`, `
    ParaInferior (ILM_RPE)`, `PeriInferior (ILM_RPE)`), na.rm =
    TRUE),

     Ratio_ILM_BM_Para_Peri = rowMeans(select(., `ParaTemporal (ILM
    _BM)`, `ParaNasal (ILM_BM)`, `ParaSuperior (ILM_BM)`, `
    ParaInferior (ILM_BM)`),
                                       na.rm = TRUE) /
                               rowMeans(select(., `PeriTemporal (ILM
    _BM)`, `PeriNasal (ILM_BM)`, `PeriSuperior (ILM_BM)`, `
    PeriInferior (ILM_BM)`),
                                       na.rm = TRUE))
left <- left %>%
  mutate(
    # Normalized TSNIT Average
    Normalized_TSNIT_Average = `TSNITAverage (RNFLParameters)` / `
    DiscArea (ONHParameters)`,
    # TSNIT Coefficient of Variation
    TSNIT_CoV = `StandardDeviation (RNFLParameters)` / `
    TSNITAverage (RNFLParameters)`)

right <- right %>%
  mutate(
    # Normalized TSNIT Average
    Normalized_TSNIT_Average = `TSNITAverage (RNFLParameters)` / `
    DiscArea (ONHParameters)`,
    # TSNIT Coefficient of Variation
    TSNIT_CoV = `StandardDeviation (RNFLParameters)` / `
    TSNITAverage (RNFLParameters)`)
```