Ocular Disease Recognition

Sprint 2

October 2, 2023

The Problem

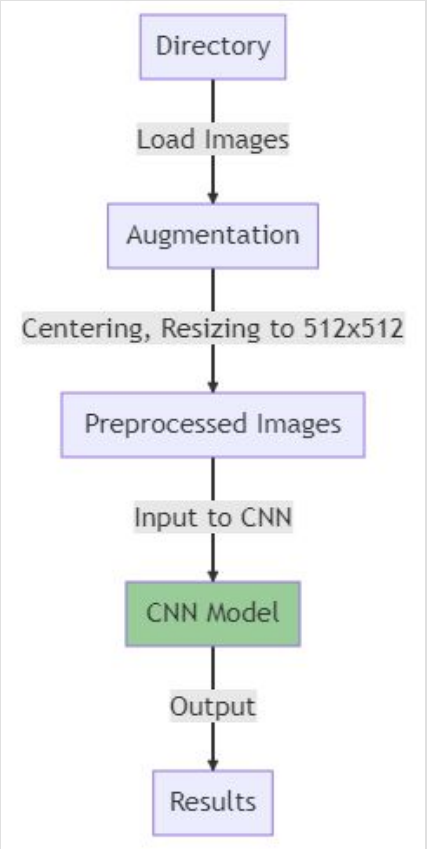
Delayed and Costly Diagnosis

Ocular diseases often go undiagnosed until they are in advanced stages, leading to irreversible vision loss. Traditional diagnostic methods are time-consuming and require specialized equipment and expertise, contributing to delays and high costs.

The Data Science Solution

Early and Accurate Diagnosis

We aim to employ machine learning algorithms trained on a comprehensive dataset of ocular images and patient records. These algorithms can analyze new data in seconds, offering potential diagnoses that can be further verified by healthcare providers.



Potential Impact

Broad Benefits:

- This solution could potential benefit millions of individuals at risk of ocular diseases by facilitating early intervention.
- Assist healthcare providers in making quicker, data-backed decisions, and potentially reduce the financial and time burden on medical institutions.

Data Foundation

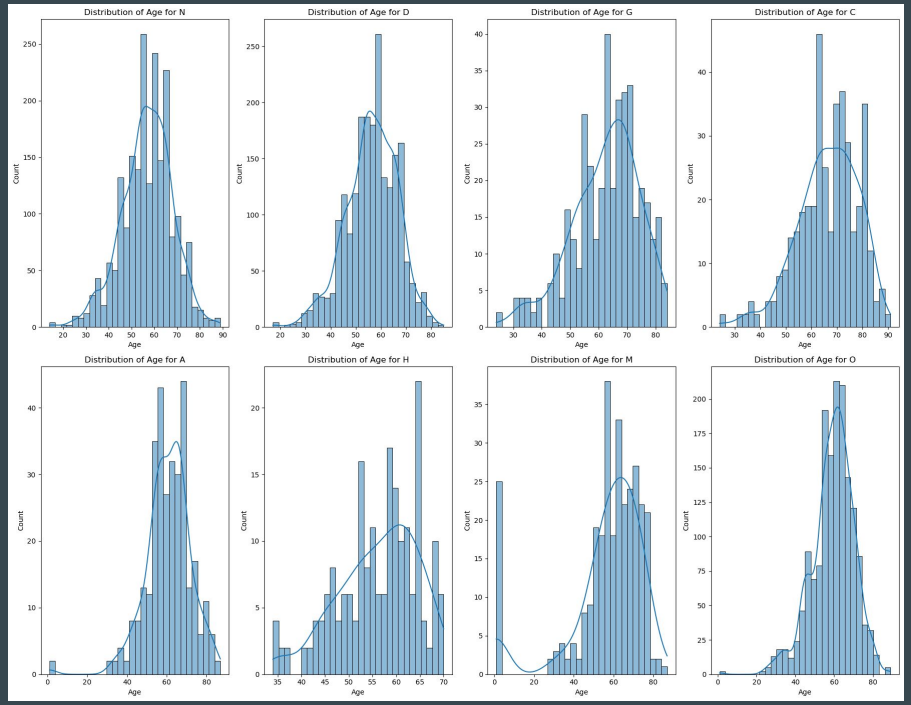
ODIR-5 Dataset

Consists of 5,000 patients' data and 14,400 fundus images.

Rich Data Source:

- Retinal scans and optical coherence tomography (OCT) images
- Along with corresponding medical records such as patient age, gender, and disease labels

Data Foundation



Preprocessing

Missing values and inconsistencies were identified and resolved to ensure data integrity.

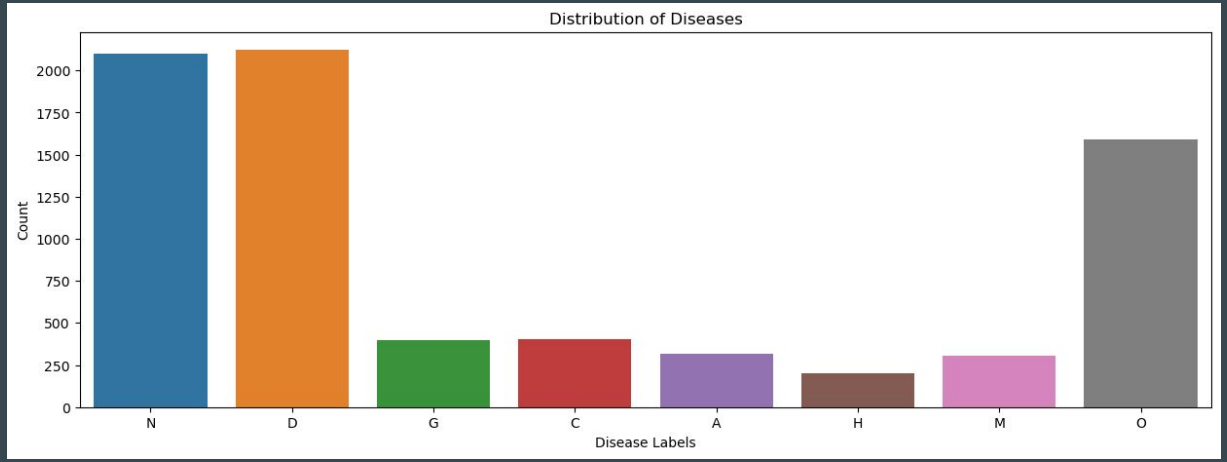
Feature Engineering:

- Outliers such as young age were addressed and found to have no statistical significance
- Images were flattened and normalized to prepare them for baseline modeling

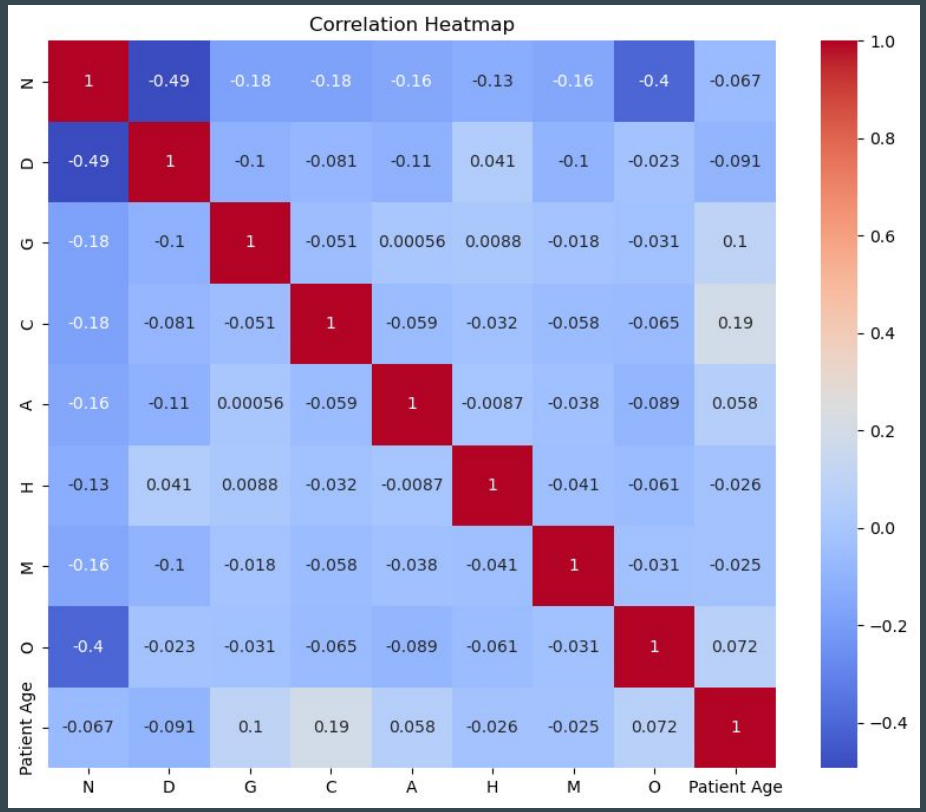
Exploratory Analysis

Key Finding 1

Analysis revealed a significant imbalance in the dataset concerning disease labels, suggesting that some conditions are underrepresented



Exploratory Analysis



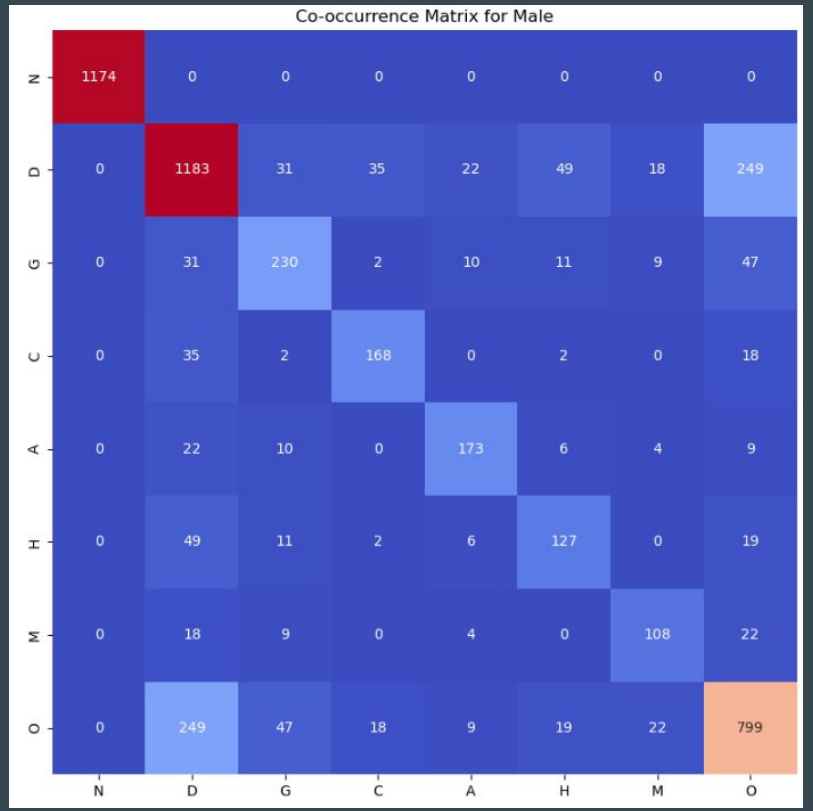
Key Finding 2

We found a notable correlation between age and the occurrence of ocular diseases such as glaucoma and age-related macular degeneration.

Exploratory Analysis

Key Finding 3

EDA also uncovered varying prevalence rates for certain diseases between genders, indicating that gender could be a significant feature for the model.



Preliminary Modeling

Baseline Models

Logistic Regression:

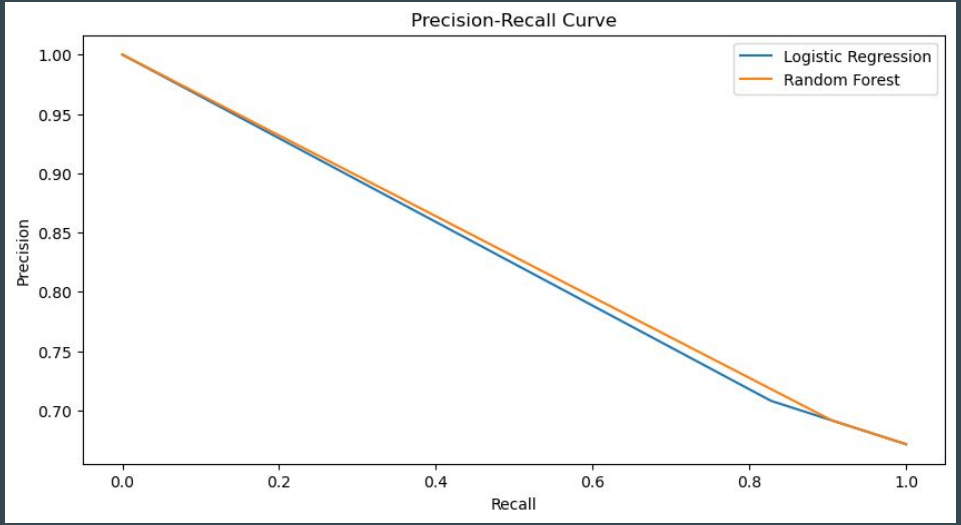
- Used as a simple yet effective model to get initial insights

Random Forest:

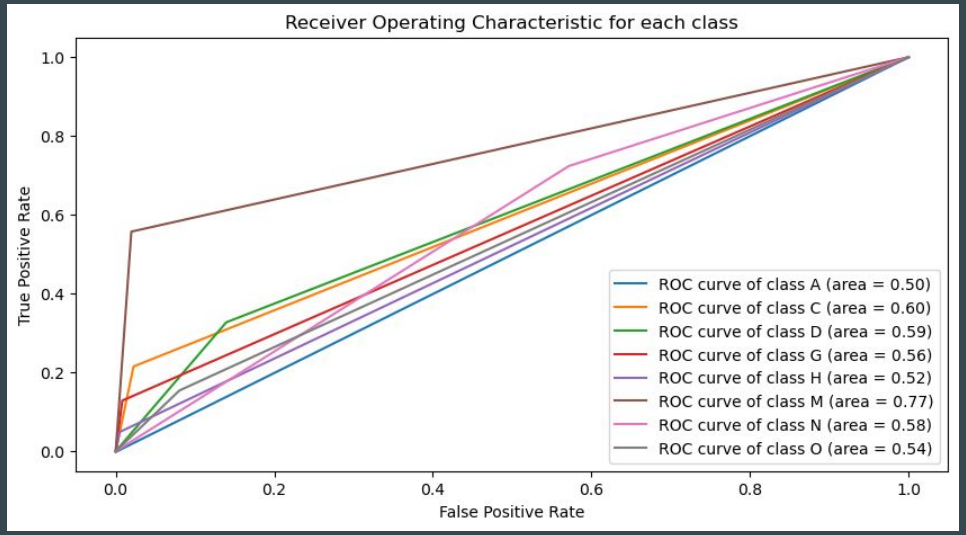
- Employed for its ability to handle complex feature interactions and provide feature importance insights

CNN:

- Used for is strength in handling image data



Preliminary Modeling



Evaluation Metrics

F1-Score:

- Chosen for its balanced consideration of both precision and recall, essential due to the uneven distribution of disease labels

ROC-AUC:

- Used to evaluate the model's ability to distinguish between classes
- Provides a single performance metric that summarizes the trade-off between true positive rate and false positive rate

Loss (Cross-Entropy):

- Evaluate how well the model estimates actual labels

Future Directions

CNNs	• Deepen CNN architecture • Employ Transfer Learning
Optimization	• Hyperparameter Tuning • Data Augmentation
Evaluation	• Cross Validation • F1-Scores, ROC-AUC
Scaling	• Integrate into Electronic Health Records • Create data pipeline from OCTs

Thank You

Presentation by:
William Bergman

—

BrainStation®