Gowri Dinesh Nair
Registration Number: 200968001
B.Tech Data Science and Engineering
Semester V, Section B

# Deep Learning Project:
# Cataract Identification

## Problem Statement:

Cataracts are the leading cause of blindness in nearly 50-80% of the Indian population affected by blindness. Currently[1], the only reliable method of cataract identification is through a doctor's diagnosis. While doctors' diagnosis has the highest accuracy to date, registering for and seeking a doctor's diagnosis can be time-consuming. Considering that the cataract population is expected to double every two decades[2], a more efficient method of detecting one of the world's most common eye diseases can potentially identify more cataract patients faster and allow them to seek the required medical help earlier. A possible solution to this issue is a deep learning model that, given a color fundus photograph, can automatically identify whether a patient is affected by cataract or not. In this project, three different deep learning models will be trained and evaluated for this purpose - a CNN-based model, a VGG-based model, and an IncpetionV3 model.

## Objectives:

- Understand the correlations between individuals diagnosed with cataracts and their age, gender, and other eye diseases.
- Train at least three models for image classification (cataract vs. normal fundus classification using a CNN-based model, a VGG model, a ResNet model, and an Inception model)

- Evaluate the performances of each model to identify the best-performing model out of those trained
- Train at least one model to have a minimum accuracy of 70%
- Create a platform utilizing the best-performing model that allows users to submit fundus photographs and receive a classification of cataract vs. normal
- In the long term, provide a supplement to healthcare services that can speed up the process of providing medical care to the growing population of cataract patients

# Metadata:

The data used in this project is obtained from a Kaggle dataset for ocular diseases[3]. The dataset chosen was limited by compatibility with the computational resources at hand and focuses solely on the data regarding diagnoses of cataract-affected and normal fundi. Given is both a CSV file on patient information, “full_df.csv” and color photographs of fundi. Below is the metadata of the data in CSV format:

Columns in the dataframe:

- Age of patient
- Sex of patient
- File name - Color fundus photograph of left eye
- File name - Color fundus photograph of right eye
- Keywords of left eye diagnosis
- keywords of right eye diagnosis

The next 8 columns are one-hot encoded diagnoses of patients:

- Normal [‘N’]
- Diabetes [‘D’]
- Glaucoma [‘G’]
- Cataract [‘C’]
- AMD - ( Age related Macular Degeneration) [‘A’]
- Hypertension [‘H’]
- Pathological Myopia [‘M’]
- Other diseases/abnormalities [‘O’]

This is followed by:

- Filepath (of each image)
- Labels (assigned by the doctors to either the left or right eye, specified in "filename")
- Target (one-hot encoded labels assigned to each patient)
- Filename (that references the labeled eye)

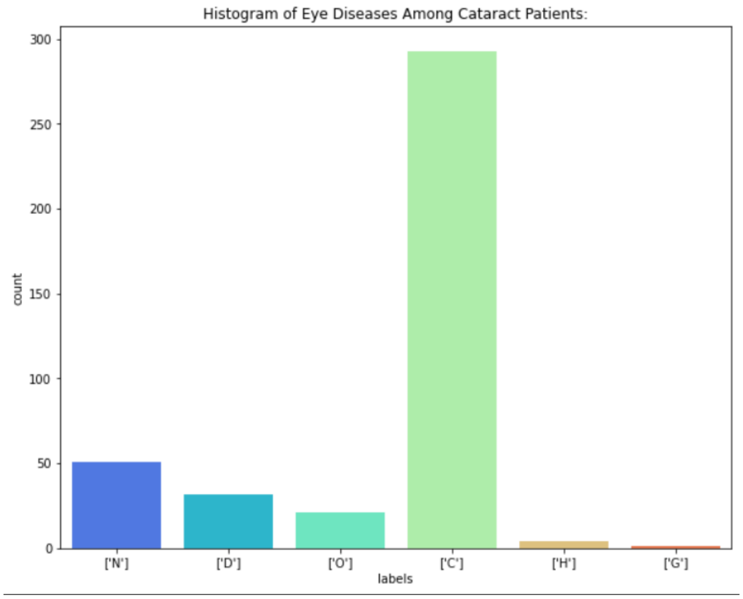
# Exploratory Data Analysis:

There is a preexisting problem with the data representation; each row of the dataframe consists of information for one patient, meaning information regarding both of the patient’s eyes is present in each row. However, the “labels” feature appears to be based on the photograph of only one eye per patient. This is addressed in detail in the attached python notebook. Below are insights obtained from the exploratory data analysis:

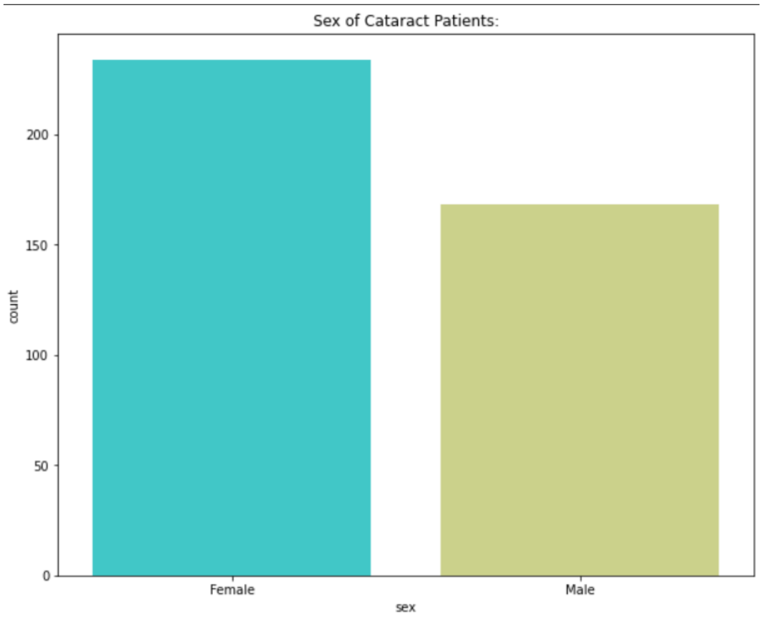
Of the 6392 patients listed in this dataset, around 402 patients (6% of all patients in the dataset) are affected by cataracts in either one or both eyes.

The data primarily consists of diagnoses and photographs for patients with either normal fundi (indicated by ‘N’ by the “labels” feature, 2873 in number), or diabetic retinopathy (indicated by ‘D’ by the “labels” feature, 1608 in number)

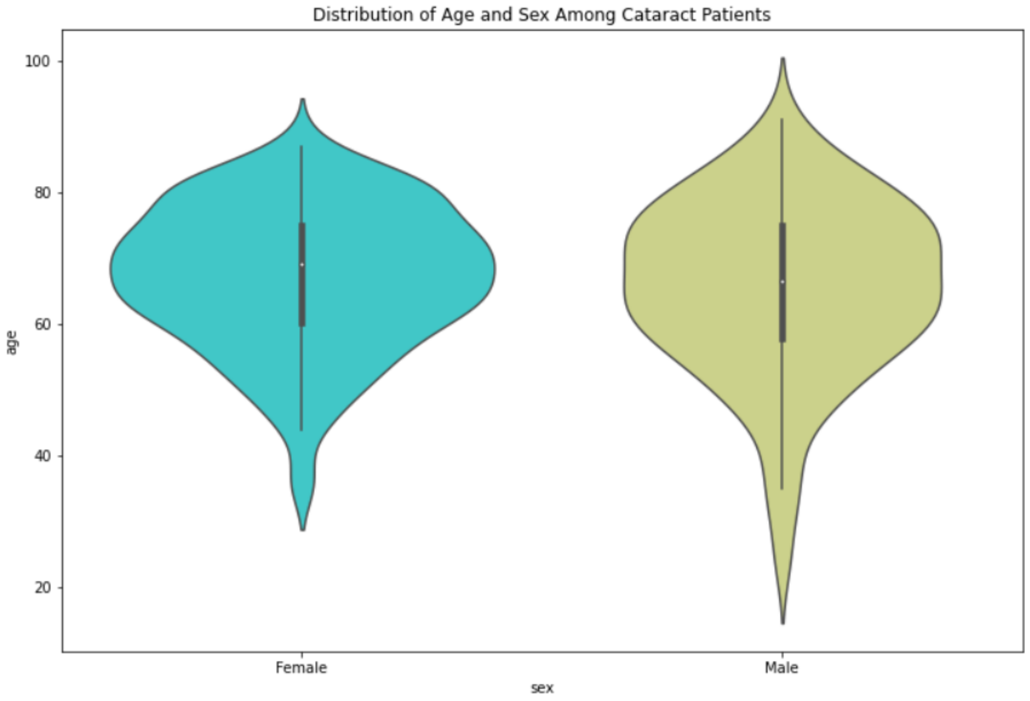
Among cataract patients, nearly 1/6th was also diagnosed with either diabetic retinopathy or other diseases. Surprisingly, no cataract patients were diagnosed with age-related macular degeneration or myopia.



Primarily, the cataract patients in this dataset identify as female, which is consistent with a study in the British Journal of Opthalmology stating that women are 69% more likely to be cataract blind than men[4].

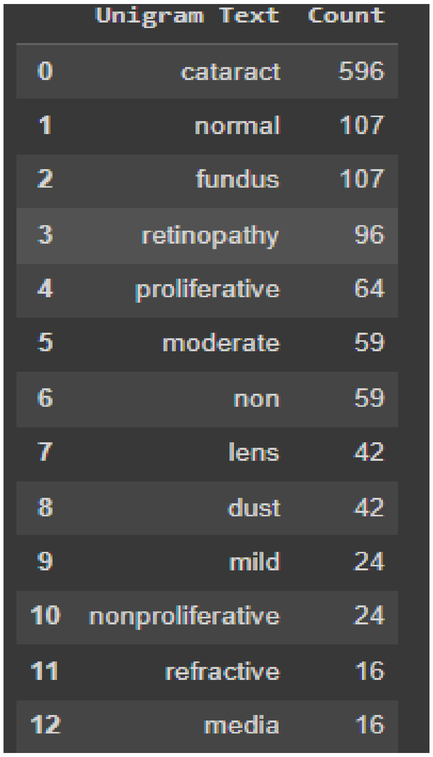


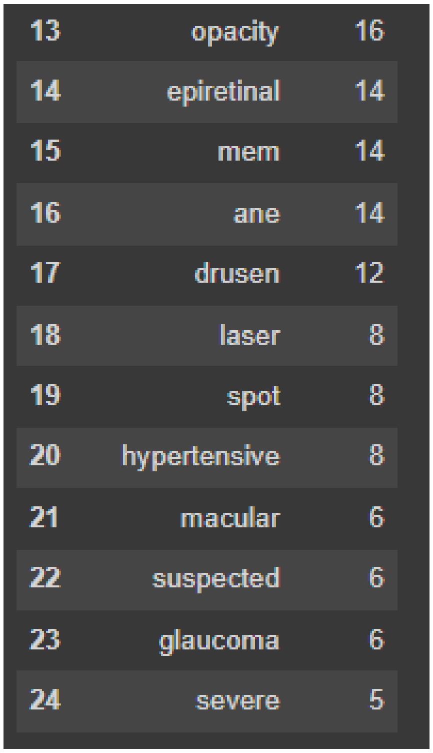
Patients between the ages of 50-80 make up the majority of the cataract patient population regardless of sex. Women in the age range of 60s-70s are the most likely to be affected by cataracts. However, when examining individuals in the age group of 30 and below, males are much more likely to be diagnosed with cataracts than females.



Preprocessing the text data used in the “left_diagnosis” and “right_diagnosis” columns and creating a unigram reveals keywords that are associated the most with cataract patients. The text counts show that the keyword “normal” was used 107 times while “cataract” was used 596 times. Thus, it can be inferred that nearly 1/6th of the cataract patient population is affected by cataracts only in one eye while the other is diagnosed as being “normal”.

Additionally, the data below shows the link between cataracts and diabetic retinopathy. According to an NCBI article, diabetes patients are 2-5 times more likely to develop cataracts than nondiabetic patients[5]. The diagnoses in this dataset identify most diabetic retinopathy patients with the keyword “retinopathy”,  which is shown in the top-25 unigram of cataract patients’ diagnoses below:



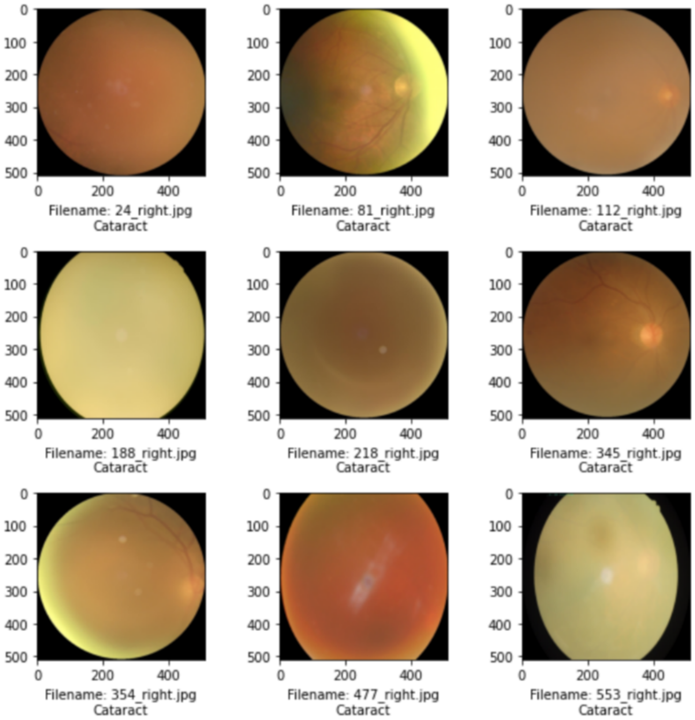


| | Unigram Text | Count |
|---|---|---|
| 0 | cataract | 596 |
| 1 | normal | 107 |
| 2 | fundus | 107 |
| 3 | retinopathy | 96 |
| 4 | proliferative | 64 |
| 5 | moderate | 59 |
| 6 | non | 59 |
| 7 | lens | 42 |
| 8 | dust | 42 |
| 9 | mild | 24 |
| 10 | nonproliferative | 24 |
| 11 | refractive | 16 |
| 12 | media | 16 |
| 13 | opacity | 16 |
| 14 | epiretinal | 14 |
| 15 | mem | 14 |
| 16 | ane | 14 |
| 17 | drusen | 12 |
| 18 | laser | 8 |
| 19 | spot | 8 |
| 20 | hypertensive | 8 |
| 21 | macular | 6 |
| 22 | suspected | 6 |
| 23 | glaucoma | 6 |
| 24 | severe | 5 |

# Preprocessing:

In order to narrow down our data to patients with cataracts and normal fundi, I have used the “left_diagnoses” and “right_diagnoses” features rather than the “labels” feature as more patients can be identified this way. These columns were searched using the keywords “cataract” and “normal” for this purpose, and the appropriate data was initially retrieved into two separate data frames for the purpose of visualization. Below are a few sample photographs of cataracts and normal fundi:

Color photographs of cataract-affected eyes::



Filename: 24_right.jpg
Cataract

Filename: 81_right.jpg
Cataract

Filename: 112_right.jpg
Cataract

Filename: 188_right.jpg
Cataract

Filename: 218_right.jpg
Cataract

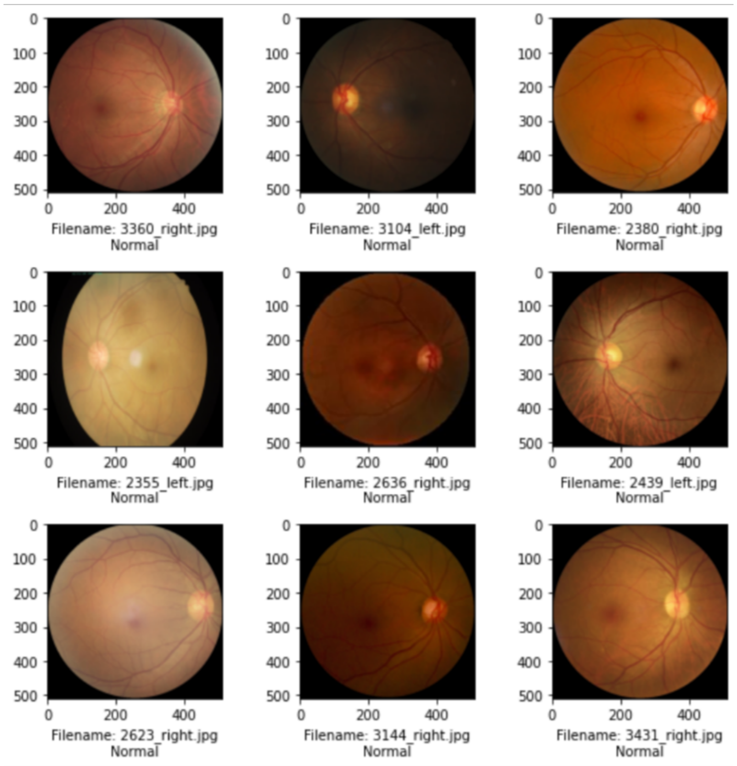
Filename: 345_right.jpg
Cataract

Filename: 354_right.jpg
Cataract

Filename: 477_right.jpg
Cataract

Filename: 553_right.jpg
Cataract

Color photographs of normal eyes:



Filename: 3360_right.jpg
Normal

Filename: 3104_left.jpg
Normal

Filename: 2380_right.jpg
Normal

Filename: 2355_left.jpg
Normal

Filename: 2636_right.jpg
Normal

Filename: 2439_left.jpg
Normal

Filename: 2623_right.jpg
Normal

Filename: 3144_right.jpg
Normal

Filename: 3431_right.jpg
Normal

From direct observation, the cataract photographs appear more cloudy and whitish in color, while the normal fundi are darker in color. However, further domain knowledge is required to accurately determine whether a color photograph depicts a cataract eye or not from observation alone.

Next, the concatenated dataframe of cataract and normal eyes were sampled to create train, test, and validation data frames. Initially, we split our data only into train and test data frames (80% of the concatenated dataframe is separated for training, and 20% for testing).

The image data referenced in these data frames are further augmented using Keras's ImageDataGenerator to provide our models with diverse training data for better generalizability. We also use IamgeDataGenerator's properties to create a validation dataset. As I will be working with three different models, I have created three different sets of augmented training, testing, and validation image data as required by the models' natures. For the VGG-based model, the images are of size 224x224 pixels, while for the CNN-based and InceptionV3 models, the images are of size 256x256 pixels and 299x299 pixels, respectively.

Links to References:

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4064227/#:~:text=%5B1%2C2%2C3%2C,our%20rural%20population%20(78.6%25)
2. https://www.pantechsolutions.net/cataract-detection-system-using-deep-convolutional-neural-network
3. https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k?datasetId=611716&sortBy=voteCount&searchQuery=inception
4. https://www.bmj.com/company/newsroom/cataract-blindness-significantly-more-common-in-women-than-men-in-india/
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3589218/#:~:text=Cataracts%20are%20among%20the%20earliest,than%2040%20years%20of%20age.