Convolutional Neural Network-Based Cataracts Prediction from Fundus Imagery

Group 21
Andrea Clark (aoc2111)
Sam Friedman (smf2240)
Yufei Guo (yg2892)
Junqi Zou (jz3506)

Introduction

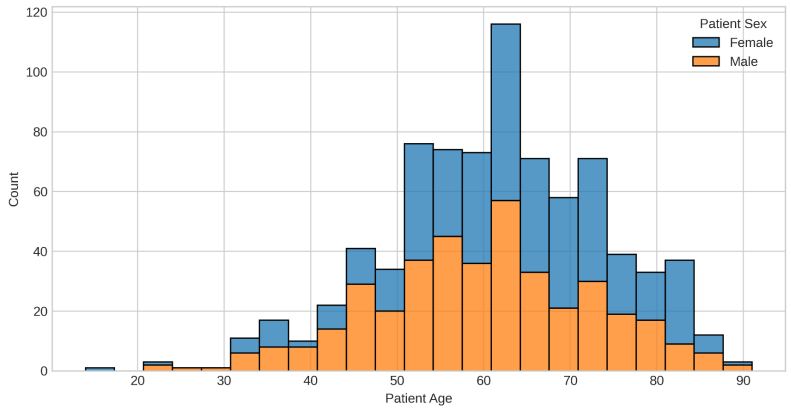
Our group wanted to create a model that would classify patients as having cataracts or not, based solely on images of the back of the patients' eyes (fundus images). To accomplish this task, our group trained a convolutional neural network as a predictive model.

We found a dataset on Kaggle titled "Ocular Disease Intelligent Recognition" (ODIR), collected by Shanggong Medical Technology Co., Ltd. from different hospitals/medical centers in China, that provided us with ample data for our project. ODIR contains colored right and left eye fundus photographs of 5000 patients, along with diagnostic data from physicians. In addition to the diagnostic data, trained personnel annotated each set of right and left eye fundus images with up to eight classification labels, including: Normal (N), Diabetes (D), Glaucoma (G), Cataract (C), Age related Macular Degeneration (A), Hypertension (H), Pathological Myopia (M), Other diseases/abnormalities (O).

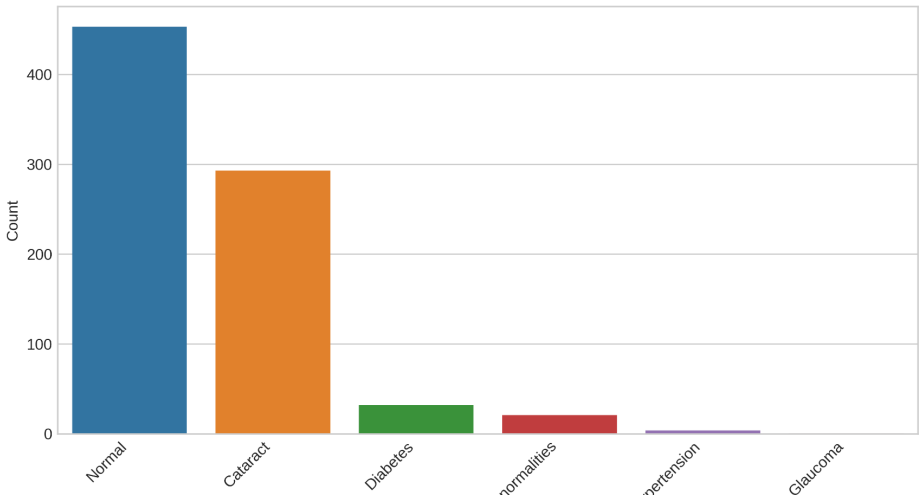
Exploratory Data Analysis

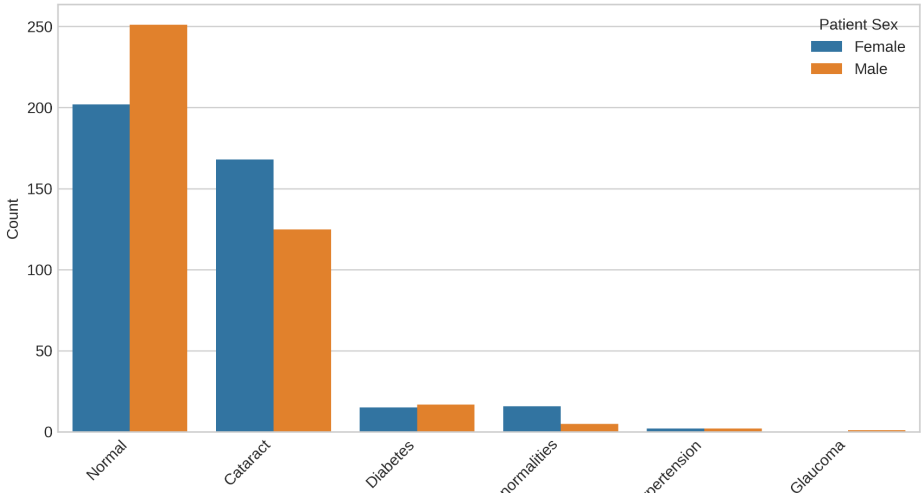
Inspecting the gender distribution of patients, there are considerably more male patients (1,342) than female patients (1,161).

The plot below depicting the age distribution indicates that the mean age is around 60 years. The male population has a stronger normal distribution, but this is likely due to the larger sample size for this population.



In order to determine how to frame our modeling problem, we looked into the counts of patients by the eight possible outcomes. The following two figures show the counts by outcome and the count by outcome grouped by gender.





There is a large imbalance between the patients having normal ocular conditions versus any other condition. The second most prevalent outcome is that of cataracts, and looking at the plot broken down by gender, it is roughly equally represented by both genders, with a slightly higher occurrence in the female population. For this reason, we chose to design a model that classifies between healthy patients (normal) and those exhibiting cataracts.

Modeling

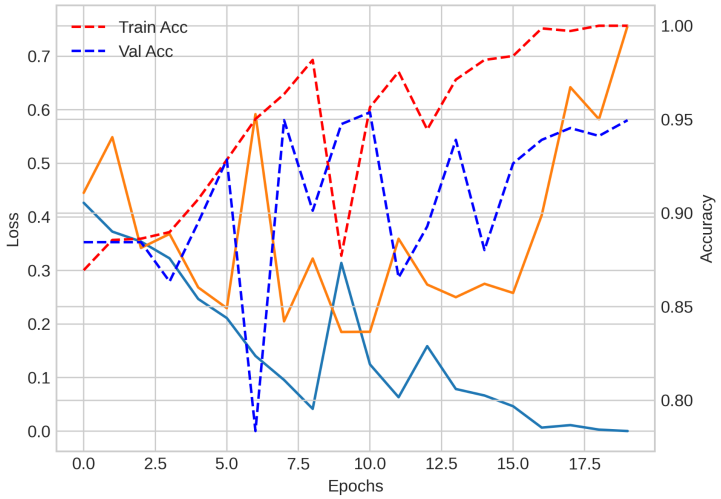
One challenge about this modeling task is that there inherently is a huge gap between the positive and negative examples we have in the data. In our case, while cataracts was the second most prevalent condition, after the normal condition, there are only 402 examples in the training data of such patients, versus 2,101 normal patients. For this reason, we performed a test-train split stratified by the target class, in order to ensure that the minority class is well-represented.

Although the data has patient age and gender available, we chose not to include these features in our model, as we wished to utilize a fully convolutional neural network (CNN) to predict between the two classes. Furthermore, as we wished to use saliency maps to interpret our prediction results, having these additional non-image-based features would not lend itself to our desired interpretation of the prediction results solely based on the images.

Our custom model consists of six convolutional layers, with filter sizes (16, 16, 32, 32, 64, 64), with a batch normalization layer and max pooling layer following the first two CNN layers, and another max pooling layer following the final two layers. This output is then flattened and passed through two, fully-connected layers of 1024 and 128 hidden units, respectively.

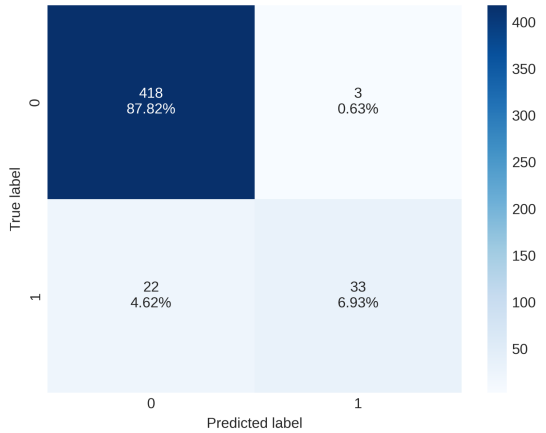
Results

The figure below shows the model's performance on the train and validation sets:



After training the model for 20 epochs, the validation accuracy (94.96%) is slightly below the train accuracy (100%), however, this amount is marginal. While the perfect train accuracy could indicate that the model is overfitting, the model's performance on the test set was 82.4%, which suggests that the model is perhaps doing a good job at discerning between the two classes. However, given the highly imbalanced nature of the target, it is necessary to look at the model's

precision and recall scores as well. The confusion matrix below shows the model's performance on the test set:



The test accuracy is 94.7%, the precision is 91.7%, the recall is 60%, and the F1 score is 72.5%. Given the lower recall of 60%, the model is yielding a higher number of false negatives, which is certainly due to the imbalance in the class labels.

Conclusions

In conclusion, we have built a model that is able to correctly detect the presence of cataracts in many patients, with an out-of-sample accuracy of 94.7%. However, the recall of our model is rather low, so it is incorrectly predicting a large number of patients as not having cataracts when in fact they do have the condition, which is of course problematic for a diagnostic model such as ours.

Future work would involve more involved methods for balancing the class label. Our current implementation splits the data stratifying by class label, but a more rigorous treatment would include additional data augmentation for the minority class and undersampling of the majority class. In summary, our model shows promising diagnostic ability, but its predictions should not be taken as a medical diagnosis, rather as a tool to aid physicians in their workflow and treatment of patients.