Diagnosing Diabetic Retinopathy

Fariha Tamboli*
Ben Chrepta †

February 11, 2024

Abstract

Diabetic Retinopathy (DR) disproportionately affects communities of color. Though DR cannot be reversed, early detection can prevent blindness. DR occurs when blood vessels in the retina are damaged by poor regulation of blood sugar levels. Since the retina can be imaged to identify DR, this project aims to create a machine learning model to increase successful DR diagnoses. The labeled image dataset from Kaggle has more than 35,000 images of left and right eyes with and without DR. Though reducing images results in loss of information, it is critical for higher performance. To address class imbalance, a binary classification algorithm was implemented on a dataset with an equal number of diseased and normal images. The models used are logisitic regression, ResNet, and CNN. The logistic regression gives a 12% test accuracy, the CNN gives a 63% test accuracy, and the ResNet gives a 80% test accuracy.

1 Introduction

For people between the ages of 20 and 74, Diabetic Retinopathy (DR) is the leading cause for new cases of blindness. Though early detection makes this disease preventable, DR still occurs in 75% of diabetic patients within a decade of diagnoses [9]. Diabetes occurs due to abnormal glucose levels in a patient's blood. Retinopathy occurs when there is blood vessel damage in the eye's retina [4]. When retinopathy occurs because of diabetes, it is referred to as diabetic retinopathy (DR).

DR occurs when high glucose levels in the blood cause a decrease the elasticity of the blood vessels, causing them to shrink. The shrinkage affects oxygen levels in organs, such as the eyes. A lack of oxygen increases blood pressure, and the small blood vessels swell up with blood. The increased pressure causes the blood vessels to swell, leak, and eventually erupt. As blood from the retina's vessels leak out, the patient's vision can be affected [2].

*ftamboli@seas.upenn.edu
†chreb@seas.upenn.edu

DR is a progressive condition with four main stages that track the deterioration of the retina's blood vessels. In the first stage called mild nonproliferative retinopathy, there are microaneurysms present in the retina. A common first sign, microaneurysms are dilated capillaries that rarely affect vision or cause other symptoms. This leads to DR being rarely diagnosed in the first stage [8].

The second stage is moderate nonproliferative retinopathy. Here the blood vessels in the retina swell and cause symptoms, such as blurred vision and floaters. The blood flow is essentially blocked, causing a lack of oxygen in the retina [8].

The third stage is severe nonproliferative retinopathy. This is essentially the same as the second stage, but the retina attempts to regrow the blood vessels to compensate for a lack of blood. Vision is still blurry, but this often causes abnormal blood vessel growth and can still harm the patient's vision [2]. By this stage, many symptoms include floaters, blurry vision, fluctuating vision, impaired color vision, and dark spots in vision [3].

In the fourth stage of DR, proliferative retinopathy, the abnormal and weak blood vessel growth can cause blood leakage, more vision obstruction and potentially complete blindness [8].

Currently, the Mayo Clinic recommends that diabetic patients conduct an eye exam at least once a year [3]. During pregnancy, a person can develop gestational diabetes which can also lead to vision loss [3].

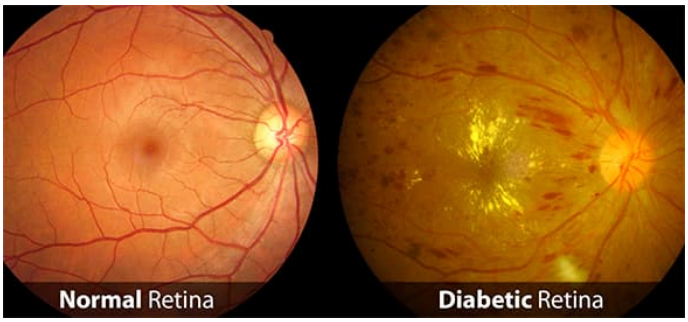


Figure 1: What to Look for When Diagnosing Diabetic Retinopathy. In a normal retina, the blood vessels are thin lines. When DR is present, there are red spots present, which is the leaking blood from the blood vessels [6].

DR can have other side effects in addition to partial or complete vision loss [3]. Vitreous hemorrhage is when blood from the broken blood vessels leaks into the fluid in the center of the eye. Though this does not cause permanent vision loss, it can obstruct vision for up to a few months until the blood clears out [3]. The rapid growth of new blood vessels in stage 3 and stage 4 can cause retinal detachment. This abnormal blood vessel growth stimulates scar tissue growth and cause the retina to detach from the back of the eye [3]. Another side effect is glaucoma; the new blood vessel growth can occur in the front or the fluid part of the eye and cause pressure build up. This pressure can then affect the optic nerve, which connects to the brain [3].

The Mayo Clinic recommends prevention methods, including eye exams that test for DR. These prevention methods include controlling a patient's cholesterol, blood pressure, and glucose levels. Managing a patient's diabetes is the most essential prevention method [3]. When an ophthalmologist diagnoses DR, they look for blood leaking out of the blood vessels, as shown in Figure 1.

2 Motivation

Since early detection is crucial to avoiding complete blindness in patients with DR, it is necessary to have the resources to accurately and quickly predict DR. Lack of symptoms during the early stages of DR also make detection and DR screening difficult [1]. Within 20 years of diagnosis, almost all patients with type 1 diabetes develop some form of DR. In the same time frame, nearly 60% of patients with type 2 develop diabetes [4]. Furthermore, DR disproportionately affects communities of color. In addition to the prevalent medical racism against these communities, there are also other systemic factors at play such as lack of medical knowledge. For Hispanic diabetic patients, a language barrier and lack of health insurance is a leading cause of insufficient DR testing [1]. A study found that physicians had the most trouble finding sub-specialty care and diagnostic imaging appointments for their Black patients [1]. Furthermore, DR is on the rise for all racial groups in the US as seen in Figure 2, so it is essential that DR imaging and diagnoses be improved [7]. Using deep learning methods to

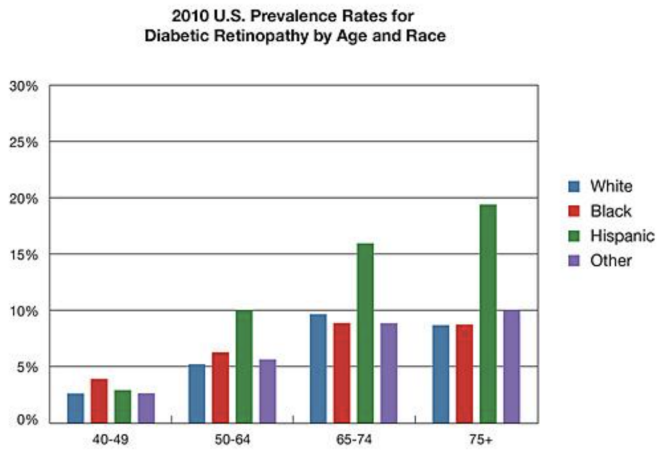


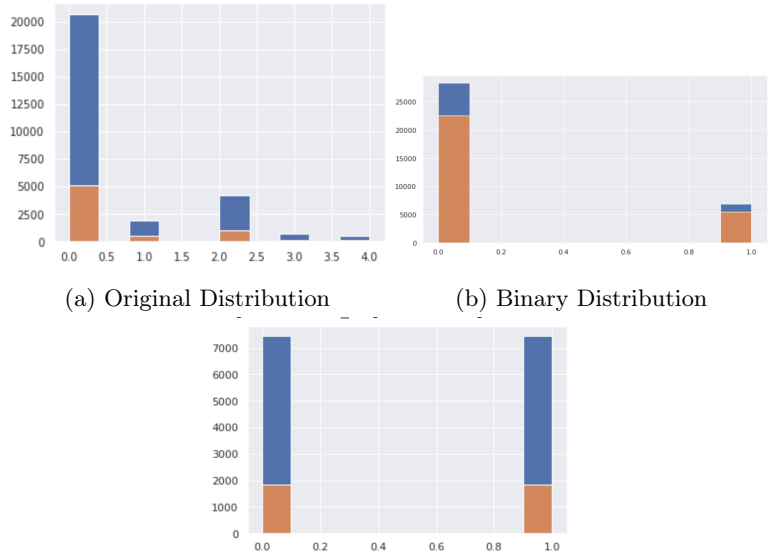
Figure 2: Diabetic Retinopathy Rates Projection [7].

detect and diagnose diabetic retinopathy can lead to an increase of accurate and accessible testing. Using a deep learning model can address accessibility in terms of cost and access to a stable power grid or internet connection. It can also reduce the time required for a physician to make a correct diagnosis.

3 Methodology

3.1 Dataset Description

The dataset obtained from Kaggle contains more than 35,000 unique images of both left and right eyes. This dataset is modified from a Kaggle competition. The retinal images are provided by EyePACS, a free platform for retinopathy screening. The images are 1024x1024 pixels, and are modified by the author since the original competition images were of varying sizes. Along with a folder of .jpeg image files is a CSV file of the image name and the diagnosis. The diagnosis is an integer value between 0 to 4, where 0 is no DR present in the image and 4 is severe retinopathy [5].



(a) Original Distribution

(b) Binary Distribution

(c) Reduce 0 Label Images

Figure 3: Class Distribution of Retina Images.

3.2 Data Cleaning

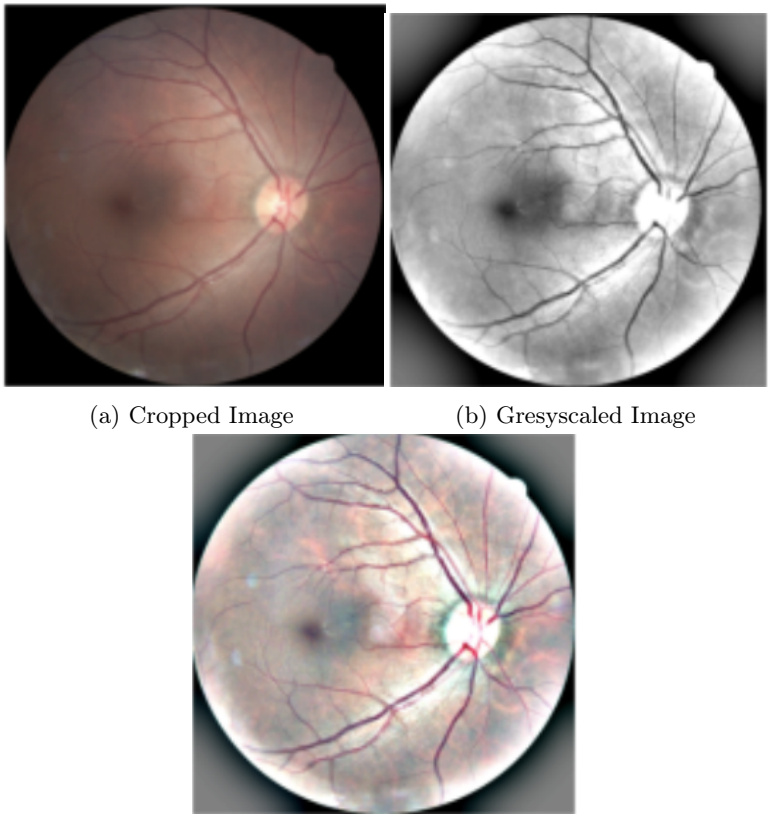
The distribution of the 35,126 images creates class imbalance as shown in Figure 3a. There are very few images of retinas with DR at any stage, and there are many images without DR. To reduce the imbalance, the labels were reconfigured to be binary. Instead of predicting the DR stage, the model predicts whether or not DR is present in the retina at all. This new distribution can be seen in Figure 3b. However, the class imbalance persists since there are very few images of DR at all. So, the number of 0 label, or healthy retinas, were reduced to

match the number of 1 label, or diseased retinas, images present. The new and final distribution of the image dataset can be seen in Figure 3c.

3.3 Image Pre-Processing

The original images of 1024x1024 were cropped to 512x512 in order to cut down on runtime and remove the black background in the image that offers no information about the retina. Since the images looked a bit dark, the images were greyscaled to extract more features. However, doing this removed some of the information that comes from the color of the image since the blood vessels are clearly red. Instead, images were brightened up to create more contrast.

After this, the images and labels were transformed into a dataloader of tensors. The label is a tuple, where [0,1] means the image shows DR and [1,0] means the image is of a healthy eye.



(a) Cropped Image (b) Gresyscaled Image

(c) Bright Image

Figure 4: Retina Images Pre-Processing.

3.4 Non-Deep Benchmark: Logistic Regression

To implement a logistic regression, the model consists of one linear layer with an input size of 512*512*3 and an output size of 2 since this is a binary classification. Then, there is a dropout layer with a dropout rate of 0.3. The model optimizer is a stochastic gradient descent with a learning rate of 0.01. The loss is calculated using a binary cross entropy logits loss function from PyTorch. This loss function combines a sigmoid layer with a binary cross entropy. The model ran for 10 epochs in 2 hours.

3.4.1 Ablation Study: Logistic Regression

To do the ablation study on logistic regression, the dropout layer was removed. The other hyperparameters were kept the same. This model ran for 10 epochs in 1.5 hours.

3.5 Base Deep Learning Model: Convolutional Neural Network

Implementing the Convolutional Neural Network (CNN) was fairly straight forward. The model consists of 3 convolutional layers, 3 2-d max-pooling layers, a dropout layer with $p = 0.2$, and 3 linear layers with an output of a one-hot encoding. The model optimizer is a stochastic gradient descent with a learning rate of 0.01 and a momentum of 0.9. The loss is calculated using the same binary cross entropy logits loss function from PyTorch. The model ran for 10 epochs over 30 minutes.

3.5.1 Ablation Study: Convolutional Neural Network

Ablation was done by removing certain elements of the CNN and seeing what the resulting accuracies were. Dropout, MaxPooling, and linear layers were removed. Everything else was kept the same.

3.6 Advanced Deep Learning Model: ResNet

An important part of selecting this model was looking through the included models as a part of PyTorch. Some of these models included variations of ResNet, GoogLeNet, and AlexNet. The preferred model was decided based on performance and accuracy, which will be explained in the later results.

The last model implemented was a Residual Net (ResNet). The specific one used was ResNet-18, a pre-trained CNN that is 18 layers deep. A linear layer that outputs to 2 features is then added, which is an attempt to have it fit to the one-hot encoding desired. The model optimizer is a stochastic gradient descent with a learning rate of 0.01 and a momentum of 0.9. Lastly, a learning rate scheduler is added to decrease the learning rate by 0.001 every 7 epochs. The model is run over 25 epochs over 15 minutes.

Table 1: Logistic Regression Confusion Matrix.

	Actual Positive	Actual Negative
Predicted Positive	27%	20%
Predicted Negative	20%	29%

4 Results and Discussion

4.1 Logistic Regression

The logistic regression with and without the ablation study gave similar results. With the dropout layer, the logistic regression model had a training accuracy of 50% and a loss of 12.4. These values did not change for the 10 epochs as seen in Figure 5. The testing accuracy for the logistic regression model with a dropout layer was 12.4% and it had a loss of 3.2. The confusion matrix reported 29% true negatives, 27% true positives, 22% false negatives, and 22% false positives. Based on the confusion matrix in Table 1, the logistic model does not do well since minimizing false negatives is essential for diabetic retinopathy.

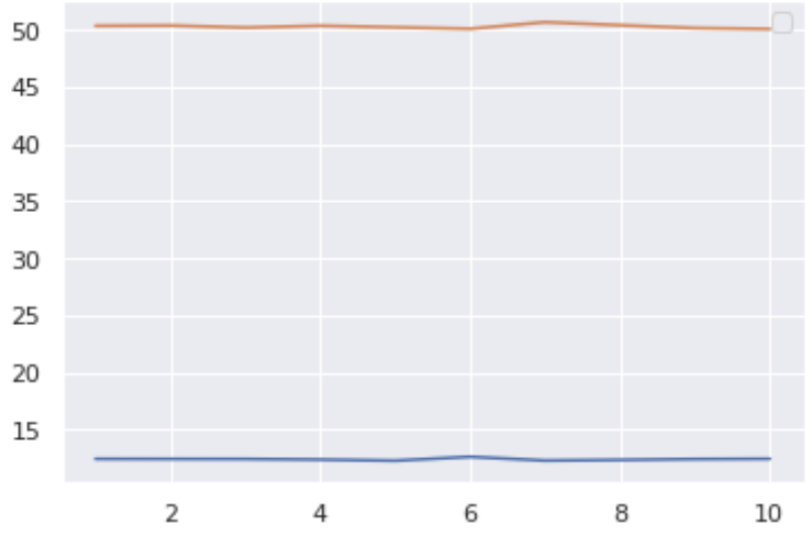


Figure 5: Logistic Regression Training Accuracy (orange) and Loss (blue).

For the ablation study, the dropout layer was removed from the logistic regression model. This new model gave a training accuracy of 50% and a much higher loss of 43.2. These values for the training dataset did not change, and this can be seen in Figure 6. The testing accuracy of this logisitc regression

model without a dropout layer was 12.3% and the testing loss was 11.2.

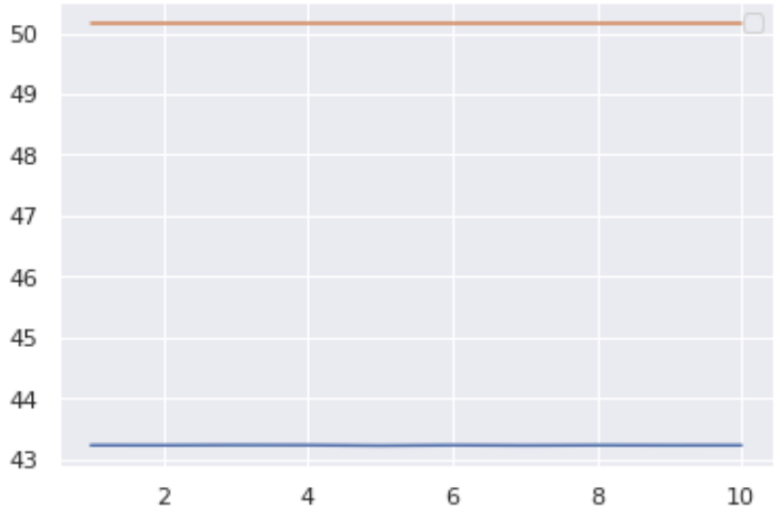


Figure 6: Logistic Regression Ablation Study Training Accuracy (orange) and Loss (blue).

The logistic regression is overall a terrible model. Since this is a binary classification problem, a random guess would have resulted in about a 50% accuracy; the logistic regression gives 12% which is significantly worse than a random guess. The low training accuracy in the logistic regression models can be attributed to the fact that the model is a fully connected network. This mean that the model does not have surrounding references to learn with, so it does not perform as well as a CNN might. Additionally, the data was downsized by more than 50% to 512x512 since the original 1024x1024 images are a resized sample of the original dataset. This causes a loss of information, especially since the blood vessels are a very small feature in the image; it is difficult to tell the difference between a stage 1 and stage 0 diabetic retinopathy patient by human eye, where stage 0 means the patient has no signs of diabetic retinopathy. For reference, see Figure 7.

4.2 CNN

Despite hopes of increasing the accuracy, the CNN only produced a peak accuracy of 74% on training and 63% on validation. The losses produced were fairly consistent, with some large variations on the way down. It generally converged down to a 0.6-0.7 loss. It was a bit better than logistic regression, as it showed that some learning of the features was done. The confusion matrix as seen in Table 2 produces 32% true positives, 31% true negatives, 19% false positives,

Table 2: CNN Confusion Matrix.

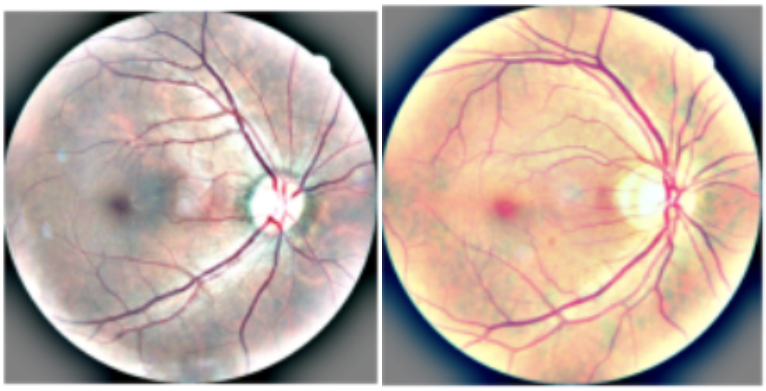
	Actual Positive	Actual Negative
Predicted Positive	32%	18%
Predicted Negative	19%	31%

and 18% false negatives. It is worth assuming that the model was not expressive enough, as after very large features were captured, the smaller ones could not be learned (represented by the false positives and false negatives). It could suggest that the false positive and false negative images are too similar to each other. Possibly more filters could be applied to help accentuate these differences.

Within the Ablation, all removals decreased the accuracy for both training and testing drastically. Referring to the validation accuracy specifically, removing the specific elements not helpful for the goal of classification. Removing any one of the convolution layers reduced the accuracy to as low as 50%. Removing any more would reduce it to 47%. At its worse, all 3 would give 45%. The convolution, broadly speaking, is a way to extract a feature map from an image of a certain size (as it relates to the mathematical function of a convolution). Removing that would then remove that advantage of that acquired feature map.

Removing the dropout layer was also detrimental to the model's performance, leading to a 57% accuracy. Given that this network was going to be wide, a drop out had to be applied in order to prevent overfitting. Such overfitting was found through the confusion matrix: while it was able to identify 17% of the true positives and 40% of the true negatives, it ended up creating 35% false negatives and 8% false positives as seen in Table 3. The heavy weightage on the false positives suggested that it was fitted more towards the negative weights.

Lastly, the max pooling was removed. Under any condition, removing a



(a) Stage 0 DR (b) Stage 1 DR

Figure 7: Comparing Stage 0 and Stage 1 Diabetic Retinopathy.

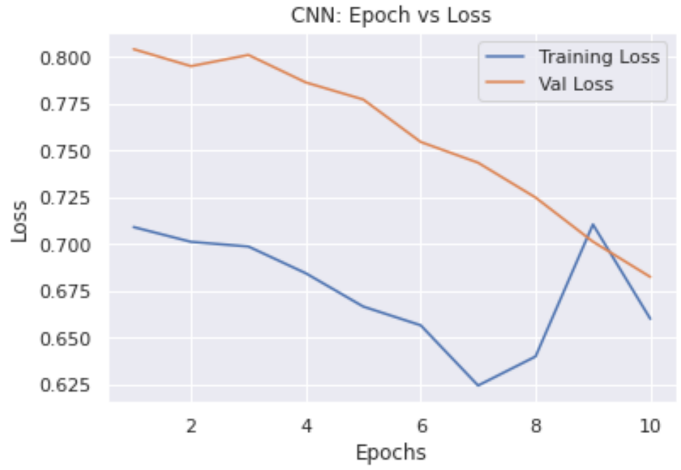


Figure 8: CNN Training and Validation Loss.

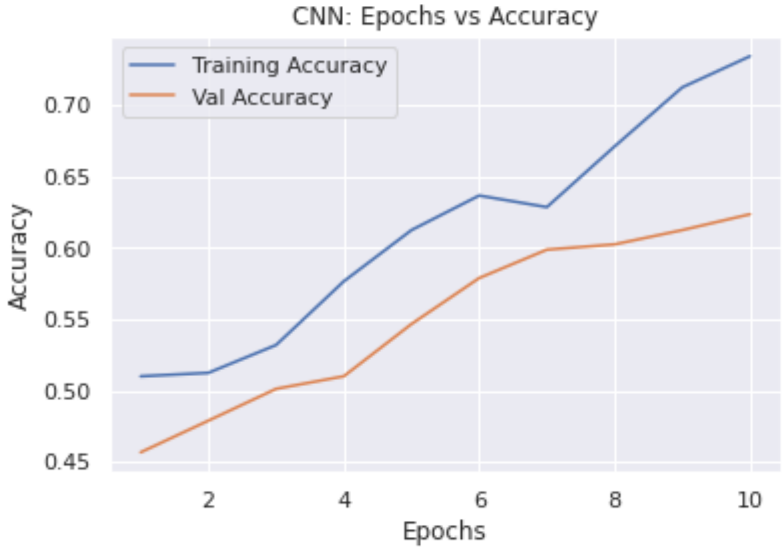


Figure 9: CNN Training and Validation Accuracy.

Table 3: CNN Ablation Confusion Matrix.

	Actual Positive	Actual Negative
Predicted Positive	17%	35%
Predicted Negative	8%	40%

Table 4: ResNet Confusion Matrix.

	Actual Positive	Actual Negative
Predicted Positive	40%	10%
Predicted Negative	10%	40%

max pooling layer was detrimental to the models performance. Removing any one layer caused it to decrease to 59%, while removing all 3 led to 53%, as well as much worse time performance due to a larger linear layer needing to be created. Max pooling is a good way to down-sample an image, which allows for the reduction of dimensionality (good for intensive processing) and can lead to assumptions being made about the features within those regions. This in essence also helps to combat overfitting and reduces computational cost. The worst case scenario of 53%, however, did not show any cases of overfitting, as the confusion matrix showed a true positive of 20%, a true negative of 33%, a false positive of 23%, and a false negative of 24%. It does serve as a general trend that max-pooling prevents overfitting, and thus increases accuracy.

4.3 ResNet

ResNet-18 was able to provide an implementation that had high training and validation accuracy, at 99% and 80% respectively. The losses were variable (they ranged between 0.0025 and 0.034), but were much lower than the CNN case. The confusion matrix in Table 4 shows 40% true positives, 40% true negatives, 10% false positives, and 10% false negatives.

All of these results over all models suggested that even with a powerful model, there is a significant classes of images that need more processing done in order to differentiate them. It is at this point that either a more powerful model must be used, or a powerful filter must be applied, in order for these false positives and false negatives to become minimized.

Other models tested included ResNet-50, ResNet-101, ResNet-152, and GoogLeNet. All models had similar validation and testing accuracies, but had a much longer runtime. ResNet-50, 101, and 152 had runtimes of 30 minutes, 45 minutes, and 1 hr 23 minutes respectively. GoogLeNet took 27 minutes. With this information in mind, it seemed like ResNet-18 was the best choice. Given that these models were pretrained, it was easier for features of images to be fitted to certain layers and neurons. Given that there aren't too many features that need to be identified, a smaller amount of layers sufficed. Even if more features needed to be identified, the other ResNet's with more layers did not solve this problem, as the accuracies stayed the same.

Figure 10: ResNet-18 Training and Validation Loss.

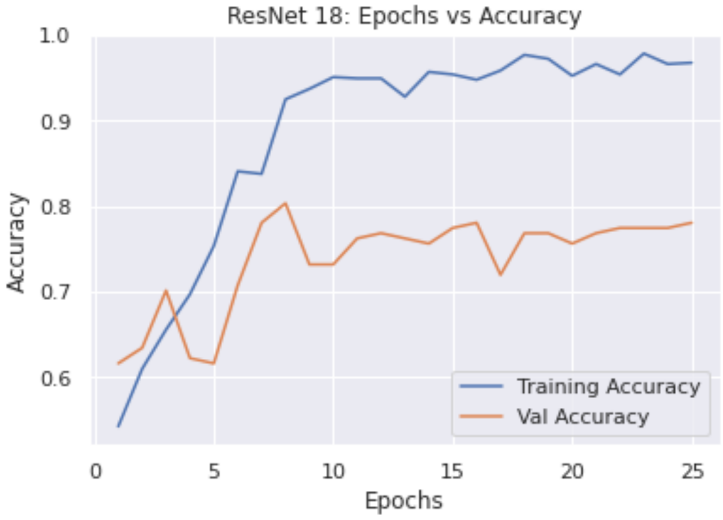


Figure 11: ResNet-18 Training and Validation Accuracy.

5 Future Work

In order to maximize the number correct of DR diagnoses made, it is essential to make the model easy to use, quick, and efficient. This means that, along with the loss, the model's runtime should be minimized. Furthermore, since diagnoses rates are higher in medically underserved communities, it is essential to make a model that can run quickly without a GPU, a stable powergrid, or internet connection. This would allow remote areas to also have access to DR diagnosing tools.

Additionally, since this project minimized the size of the image, a lot of information was lost. In the future, it would be essential to maintain the highest resolution of the image to obtain the best accuracy. One solution to this would be to create a model based on subsections of the retina images rather than processing the image of the entire eye. Also, this model was converted into a binary classification problem. A way to improve this project would be to find other way to eliminate class imbalance between the four stages of DR and classify a patient's DR based on the stages. This would result in a multi classification problem. One method to address the problem of class imbalance would be resampling of the data.

Beyond the scope of this project, some future work would involve using these models to detect other eye disorders and conditions using the same retina images.

6 Summary

Diabetic retinopathy (DR) disproportionately affects minorities at high rates. In the US, DR cases are expected to rise for the next 3 decades for all racial groups [7]. In order to prevent complete blindness in DR patients, early detection is essential. This project focuses on creating three different models to detect DR on resized retinal images. The models are logistic regression, convolutional neural network, and residual network. The logistic regression model is a baseline non-deep learning model that performs terribly; this can be attributed to the fact that the retinal images are greatly downsized and have lost the important features, since they are very small in the images. The baseline deep learning model used was a CNN. The CNN performed better than the logistic regression due to it not being fully connected and being able to learn from its surrounding. The advanced model used was a pre-trained ResNet. This model did pretty well and gave an overall accuracy of 80%.

Broader Impact

Diabetic retinopathy disproportionately affects marginalized groups. In the paper "Diabetic Retinopathy: Focus on Minority Populations" by Barsegian et al., studies find that 17% of Black patients with DR face total vision loss from diabetic retinopathy. Latinx DR patients face closer to 8%, while some Native American tribes have rates as high as 45.3%. When the study looked at the

Latinx population, 42% of the Latinx people with diabetes for more than 15 years developed diabetic retinopathy at some point. Other studies show that the Latinx community's diabetic retinopathy rates are twice that of the white community across the US. Furthermore, studies show that Black communities are already at a high risk of diabetes as compared to other races. The drastic difference of DR rates between white and POC communities can be attributed to medical racism, lack of access to healthcare, and other discrimination factors [1]. Since complete vision loss from DR can be prevented by early detection, it is important to develop tools to better identify diabetic retinopathy. Furthermore, DR affects populations in other countries that also do not have great healthcare systems. Quickly and accurately identifying DR at early stages gives patients the ability to retain as much of their vision as possible. Some negatives of introducing deep learning models into healthcare is ethics [10]. There should be fairness in representation in the training data for the model. If the data is not representative of all types of patients, then the model is not accurate and can hurt minorities. Beyond representation, Vayena et al. argue that there also needs to be transparency; the doctors using these models should know how they were created and how the models reach decisions [10]. Ultimately, the models do not present a 100% accuracy and may also need human supervision.

References

[1] Arpine Barsegian, Boleslav Kotlyar, Justin Lee, Moro O Salifu, and Samy I McFarlane. Diabetic retinopathy: focus on minority populations. *International journal of clinical endocrinology and metabolism*, 3(1):034, 2017.

[2] Kierstan Boyd. What is diabetic retinopathy?, 2020.

[3] Mayo Clinic. Diabetic retinopathy, 2018.

[4] Donald S Fong, Lloyd Aiello, Thomas W Gardner, George L King, George Blankenship, Jerry D Cavallerano, Fredrick L Ferris, and Ronald Klein. Retinopathy in diabetes. *Diabetes care*, 27(suppl 1):s84–s87, 2004.

[5] ilovescience. Diabetic retinopathy (resized), 2019.

[6] Premier Eye Care of Eastern Idaho. Diabetic eye disease, 2021.

[7] National Institute of Health. Diabetic retinopathy data and statistics, 2020.

[8] Eye Center of Texas. What are the four stages of diabetic retinopathy?, 2019.

[9] Research to Prevent Blindness. Diabetic retinopathy, 2021.

[10] Effy Vayena, Alessandro Blasimme, and I Glenn Cohen. Machine learning in medicine: addressing ethical challenges. *PLoS medicine*, 15(11):e1002689, 2018.