Understanding the Biomarkers of Retinal Disease using Deep Learning

by Malika Shahrawat

[S.B., C.S. M.I.T., 2018]

Submitted to the
Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

Author: ____
Department of Electrical Engineering and Computer Science
May 24, 2019

Certified by: ____
Jayashree Kalpathy-Cramer, Associate Professor, Thesis Supervisor
May 24, 2019

Accepted by: ____
Katrina LaCurts, Chair, Master of Engineering Thesis Committee

Understanding the Biomarkers of Retinal Disease using Deep Learning

by
Malika Shahrawat

Submitted to the
Department of Electrical Engineering and Computer Science
May 24, 2019

in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

The retina has biomarkers not only for ophthalmic disease but also for diseases and conditions across the entire body. In this paper, I focus on retinopathy of prematurity (ROP), a proliferative vascular disease that can cause blindness in prematurely born infants. To prevent further disease progression and vision loss in infants with ROP, early and accurate detection of plus disease is crucial. In current practice, clinicians compare retinal fundus photographs to a reference standard image in order to detect plus disease for severe ROP. This process can be highly qualitative, subjective, and variable. Furthermore, some clinical environments may lack clinicians with the expertise to diagnose these diseases. I am to address these shortcomings in current clinical diagnosis of ROP by using deep learning methods to automatically extract biomarkers of disease without human intervention. Since ROP is primarily present in prematurely born infants, I also attempt to predict and analyze gestational and postmenstrual age, and how disease predictions vary from healthy to affected infants.

Thesis Supervisor: Jayashree Kalpathy-Cramer
Title: Associate Professor

Acknowledgements

I would first like to thank Dr. Jayashree Kalpathy-Cramer for giving me the opportunity to work on this project for my MEng. I would like to thank all the members of the QTIM lab at the MGH/HST Martinos Center and the iROP team for giving me guidance and support as well as the resources to make this project possible. I am also thankful to all the professors and classmates throughout my time at MIT that have given me the mens et manus to be successful. And finally, I am eternally grateful to my family and friends for their endless support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Images of the eye, particularly the retina, contain biomarkers both for ophthalmic disease and broader indicators of general health across the entire body [1]. Clinicians use retinal fundus photographs to detect such biomarkers, though their assessments can be highly qualitative, variable, and require expertise that is not always readily available in low-resource settings [2-6]. Automating this process of discovering and understanding retinal disorders can greatly benefit many patients and clinicians.

I focus on retinopathy of prematurity (ROP), a proliferative vascular disease that affects prematurely born infants. Although many cases are mild, 5% to 10% of cases can progress to a severe stage of ROP which, if left untreated, can cause retinal detachment and blindness. Therefore, to prevent further disease progression and vision loss, early and accurate detection of plus disease is crucial. Severe, treatment-requiring ROP is characterized by the presence of plus disease, which is arterial tortuosity and venous dilation at the posterior pole greater than or equal to the tortuosity and dilation in a reference standard retinal photograph [7]. This presents the challenge that clinical diagnosis of ROP can be highly qualitative, and thus subject to variation between individual clinicians [8, 9]. Since ROP is primarily present in prematurely born infants, it is also beneficial to analyze the prediction of age in relation to plus disease, and where and how these predictions fail. Additionally, predicting age can help prioritize diagnoses for plus disease so that the infants that are born extremely prematurely or have significant differences between their true and predicted age are treated first.

Recent studies have used machine learning to automatically analyze retinal photographs for biomarkers of health conditions. Previous work includes using deep learning in concert with classical and statistical methods to diagnose plus disease for ROP [10-14]. Other studies applied deep learning to retinal imaging to predict risk factors such as age, gender, and blood pressure [15]. These studies have contributed to accurate, automated disease diagnosis that lessens the impact of interrater variability by using repeatable and reproducible methods. Deep learning has become state-of-the-art for computer-based image analysis, and many solutions to these types of problems are moving toward data-driven deep learning algorithms. Most published techniques have utilized deep learning with supervised techniques that require large, well-annotated datasets. For example, at least 20,000 labeled images were needed for algorithm performance to plateau when classifying abnormalities on chest x-rays, showing that large datasets are favorable for best performance [16]. However, sufficiently large labeled datasets aren't always readily available for many diseases. Exploring unsupervised methods for the diagnosis of disease can be beneficial in the case of inadequate and/or subjectively labeled data. These methods involve unsupervised lower dimensional feature extraction, followed by supervised classification or clustering of those features for a chosen clinical task.

These previous methods have worked well and perform comparably to or better than expert analysis for diagnosing ROP [13], and I would like to continue exploring more methods to learn about health conditions in an automated fashion. There are two separate tasks in this analysis: 1) finding latent representations of retinal images using unsupervised methods in order to classify plus disease, and 2) predicting the risk factors of gestational and postmenstrual age from retinal images and understanding how the predictions vary from healthy infants to those with plus disease. Using various deep learning methods on a dataset of about 6,000 de-identified posterior view retinal images that are labeled for plus disease and other features such as age and weight, I am able to build predictive models that achieve these tasks. Both gestational and postmenstrual age can be predicted to a reasonable accuracy, and subsequent analyses show the effect of plus disease on predictability of age. The lower-dimensional latent representation captures important features of ROP to predict and diagnose plus disease.

By applying deep learning methods to predict health and risk factors of infants and diagnose plus disease, I contribute to more accurate and timely automatic diagnosis of plus disease and prediction of systemic features to help prevent serious disease progression.

Chapter 2

Related Work

Prior work in the analysis of retinal photographs includes both statistical/classical machine learning techniques as well as deep learning algorithms. Although retinal photographs have been used to diagnose diabetic retinopathy, glaucoma, and other ophthalmic disease [17-19], I focus on applications relevant to the goals of this study: diagnosis of plus disease in ROP and prediction of biomarkers of clinical variables, such as age.

One computer-based image analysis plus disease diagnosis system for ROP starts with manually segmented images that are then preprocessed to construct a vasculature tree. Features of the vasculature such as measures of tortuosity and dilation are extracted. A Gaussian mixture model (GMM) is then fit to the distribution of these features. Attributes of the GMM such as the means and standard deviations are then extracted. A support vector machine (SVM) classifier is then applied to classify the images as plus, pre-plus, or normal. The performance of the system was comparable to expert diagnosis against the reference standard diagnosis [10, 11]. Continuing on this pipeline, unsupervised techniques have also been used to obtain a continuous severity index as opposed to disease classification. With the distance matrix calculated from the probability distributions, six different manifold learning methods that mapped the images to a lower dimensional space were applied [12]. These methods result in a 1-dimensional output that represents the severity index as opposed to discrete classification of plus disease.

Deep learning techniques were used to classify plus disease in retinal photographs for ROP with the same dataset used in this analysis of nearly 6,000 images [13, 14]. Images were first preprocessed by segmenting for vessels with a patch-based U-Net CNN architecture to exclude variations in image features and nonvascular pathology [20]. Then disease classification was achieved using transfer learning with the Inception-v1 (GoogLeNet) CNN architecture [21] with weights initialized from training on the ImageNet database of about 1.2 million images [22]. A severity score and diagnosis were calculated based on probabilities outputted for normal, pre-plus, or plus. Once validated and evaluated on a test set, the network was able to diagnose images with 91% accuracy whereas the eight experts diagnosed with an average accuracy of 82% against the reference standard diagnosis [13, 14].

Due to the difficulty of obtaining large amounts of annotated data, the application of unsupervised learning to medical imaging is being explored. In addition to the manifold learning methods mentioned previously for plus disease classification, there has been some work in unsupervised medical image segmentation. One method involves learning deep representations of an image from a CNN and applying k-means clustering to obtain cluster labels that are then projected to the target image to get the segmented image [23]. Other methods involve autoencoders and generative adversarial networks (GANs) to obtain deep yet latent low-dimensional representations of image features that then can be used in classification methods [24].

The retina also offers plenty of information about one’s health, including risk factors for disease that can be obtained in a non-invasive way using anatomical features such as blood vessels and the optic disk. Researchers at Google used deep learning on retinal images to predict various risk factors for cardiovascular disease, in particular: age, gender, smoking status, body mass index, systolic blood pressure, and major cardiac events with areas under the receiver-operator curve (AUCs) above 0.70 and relatively low mean absolute errors [15]. They used pre-initialized weights from training on the ImageNet dataset, trained on nearly 285,000 patients, and performed validation on over 13,000 patients for both classification and regression tasks.

Chapter 3

Dataset

Training, validation, and test data sets are selected from a dataset of 5,561 de-identified retinal images from 971 infants taken from a commercially available camera called RetCam obtained between July 2011 and December 2016 as part of the Imaging and Informatics in Retinopathy of Prematurity (iROP) study [25]. The images were taken in the five standard fields of view (posterior, nasal, temporal, superior, and inferior), but this analysis uses only posterior view of both left and right eyes for each patient as it best captures the optic disk and vessels for ROP diagnosis. Many patients were photographed for 3-4 sessions. However, it is noted that the elapsed time between these sessions is not equal for all patients. Each retinal image, regardless of left and right, and session, was treated separately in this analysis, though multiple images of the same patient were processed at the same time.

A quality measure is given to each image indicating ability to properly diagnosis plus disease. Images were excluded from this analysis if at least two of the three expert readers stated they were unacceptable for diagnosis or if the retina was diagnosed with stage 4 or 5 ROP, meaning partial or total retinal detachment which makes features, such as vessel tortuosity and dilation, difficult to visualize.

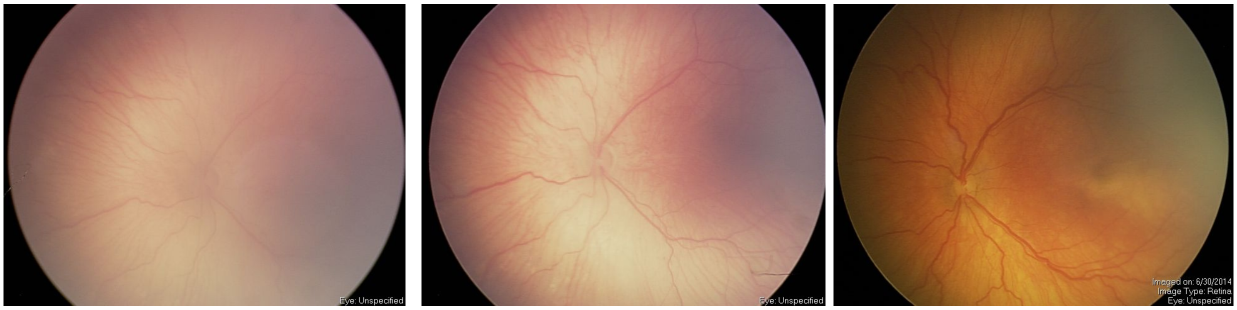


Figure 1. Retinal images exhibiting normal, pre-plus, and plus disease (left to right).

Each image was labeled with a reference standard diagnosis (RSD) by three trained graders (one study coordinator and two ophthalmologists) and the clinical diagnosis using the methods published previously [25]. The RSD was either normal, pre-plus disease, or plus disease, and these classifications were used in this analysis. Of the 5,665 retinal images included in the analysis, 4648 (82.05%) were labeled as normal, 825 (14.56%) as pre–plus disease, and 192 (3.39%) as plus disease based on the RSD. Examples of retinal images in each stage of the disease are shown in Figure 1.

This dataset also includes other important features that can be used for prediction and understanding the various biomarkers in the retina. These features include clinical features at a patient level such as gestational age, postmenstrual age, birth weight in addition to plus disease stage and classification by an expert at an image level. The gestational age (GA) is defined as the time between conception and birth of an infant. Postmenstrual age (PMA) is defined as the age of an infant from conception till the current time. The PMA is the gestational age plus chronological age. All infants in this dataset are born prematurely (less than 38 weeks of gestation). ROP primarily affects premature infants weighing about 1,250 grams or less that are born before 31 weeks of gestation. The average birth for infants with pre-plus and plus disease is a weight of 677 grams and gestational age of 24.91 weeks which is lower when compared to the statistics shown below on all 5,665 retinal images of all infants in the dataset.

	Gestational Age (weeks)	Postmenstrual Age (weeks)	Birth Weight (grams)	Session	ROP Stage (1-5)
mean (std)	26.36 (2.19)	36.41 (5.23)	877.31 (304.71)	3.37 (2.70)	0.92 (1.00)
minimum	22	30	320	1	0
25%	25	33	640	1	0
50%	26	35	830	3	1
75%	28	38	1060	4	2
maximum	34	98	2510	16	5

Table 1. Dataset statistics with respect to ROP in addition to other clinical variables.

Special care was applied when selecting images for analysis and splitting the dataset into train and test splits. For PMA prediction, only infants 42 weeks and younger were considered. Since gestational age is a fixed feature for an infant, it doesn't change with time. However in this dataset, infants are photographed at multiple sessions over time. So to ensure that the changing retinal images of a single infant does not affect the analysis, images of infants that are 30-33 weeks old PMA were used. Since patients had multiple sessions and eyes photographed, the dataset was split based on a patient level such that all retinal images of a patient were in the same split. This was done to data leakage and overfitting during the training and evaluation of the models. Eighty percent of patients were used in training the model, and the remaining 20% of the patients were used to evaluate.

Chapter 4

Methods

There are two separate tasks to this analysis: predicting clinical variables and classifying plus disease. To achieve these tasks, I use common approaches in machine learning: data preprocessing, augmentation, and deep learning. Various image processing and deep learning techniques were applied depending on the task. The following methods were achieved using the PyTorch deep learning library and scikit-learn in addition to other helpful Python packages.

4.1 Preprocessing

Both tasks involve taking the retinal photographs and analyzing them for systemic features and plus disease. These images are bitmap images mostly of size 640 x 480 pixels and color space RGB. There were taken using the commercially available RetCam camera (from Natus Medical Incorporated) as part of the iROP study [25]. Before these images are analyzed through their respective deep learning algorithms, they are properly preprocessed. The images are first randomly resized and cropped to a size of 300 x 300 pixels for age prediction and 256 x 256 pixels for plus disease prediction to capture retinal features and the optic disk. They are then flipped vertical with a probability of 0.5. All images were also normalized from ranges of values 0-255 per channel to values of -1 to 1. This is in accordance with PyTorch's recommended image preprocessing and data augmentation when applying transfer learning and other deep learning algorithms. For the unsupervised learning task, images were also segmented for vessels using a standard segmentation U-Net prior to image preprocessing step since arterial tortuosity and vessel dilation are key indicators of plus disease [20]. An example of a retina segmented for vessels using the U-Net is shown in Figure 2.

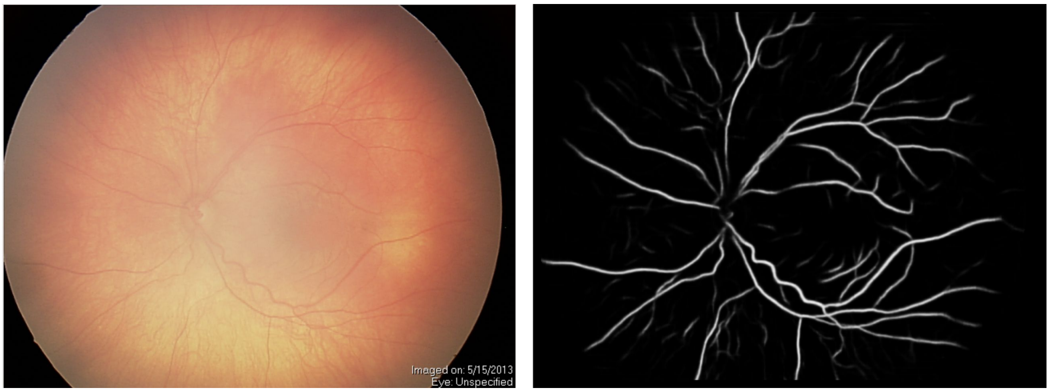


Figure 2. A retina with pre-plus disease before and after vessel segmentation using a U-Net.

4.2 Deep Learning Algorithms

Deep learning is the subfield of machine learning most commonly associated with deep neural networks to learn data representations. It has proven to be effective with state-of-the-art performance in many applications in computer vision, natural language processing, and more, and particularly advantageous in the area of medical imaging. Here I make use of deep neural networks and transfer learning to achieve the tasks of prediction of clinical variables and learning a latent representation of plus disease.

4.2.1 Prediction of Gestational and Postmenstrual Age

Prediction of postmenstrual age and gestational age is a supervised learning task where the retinal image is taken as input and the age is given as output. Two similar but slightly different approaches are used for this prediction. Postmenstrual age was predicted as a regression problem, so the age in weeks was directly computed. Gestational age was predicted as a binary classification problem. All infants in the dataset were born prematurely, so a binary classification was used to determine between extremely premature (< 26 weeks) and moderate to late premature (> 25 weeks). These predictions were then analyzed in relation to plus disease as it can provide insight to help prioritize diagnoses for patients.

This analysis makes use of a common machine learning approach called transfer learning for age prediction [26]. Transfer learning in deep learning is using knowledge gained for solving one problem and applying it to a different but related problem. In other words, a model trained for a task is reused as the starting point for a model to apply to another separate but related task.

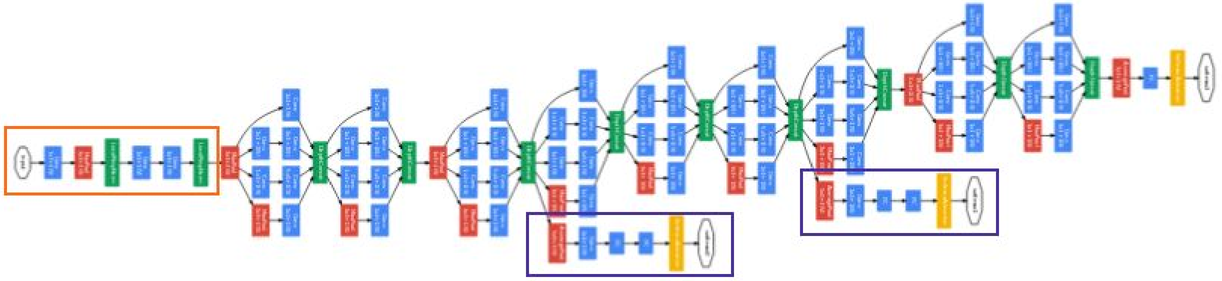


Figure 3. Model diagram of InceptionNet v1, very similar to InceptionNet v3 used in this analysis [21].

Here I use a model architecture called InceptionNet v3 (shown in Figure 3) pretrained on the ImageNet dataset of 1.2 million images in 1000 different classes [27, 22]. Despite there being no retinal images in the ImageNet dataset, using this pre-trained network enables faster convergence and better performance than if we trained from scratch on our relatively small dataset. The last fully connected layer is removed from the network, allowing it to serve as a feature extractor for the 4,068 features from InceptionNet and then two fully connected layers are added to collect the final 1,024 features and serve as the classifier for the retinal dataset.

This same network was used for both postmenstrual age regression and gestational age prediction with the only difference being the final output layer modified for the regression and classification task accordingly. Initially, freezing the pretrained ImageNet weights and only training the classifier was tried, but after multiple epochs, the network only did slightly better than random chance. Therefore, the pretrained InceptionNet v3 network was retrained with the additional classification layers with randomly initialized weights and was tuned for 20 epochs. The model used weighted cross entropy loss to address the class imbalance in the dataset and was optimized using the Adam optimizer with a learning rate of 1e-4 and weight decay of 1e-4.

4.2.2 Prediction of Plus Disease

The prediction of plus disease from retinal images involves learning a low dimensional latent representation of the images that captures the most distinguishing features, an unsupervised task. Two prominent approaches to unsupervised learning in deep learning are autoencoders and GANs. In this analysis, I make use of unsupervised neural network approaches to gain a low dimensional feature representation of the images and then use a simple classification network on the features to determine likely disease diagnosis. Ideally the algorithm would be fully unsupervised. However, the autoencoder used in this analysis maps the latent space to a Gaussian, so clustering methods based on multiple Gaussians will not be successful. Instead, the fully unsupervised pipeline is forgone and a small, simple convolutional neural network (CNN) is used to classify plus disease using the labels provided in the dataset to ensure that important features are still captured. This network was evaluated on its ability to classify plus disease using a further low dimensional representation of the segmented retinal images as opposed to the full vessel segmented or raw images.

As an attempt at a semi-supervised method, the 1,024 features extracted from the same InceptionNet v3 network architecture as described above for age prediction were clustered using the k-means algorithm. The clusters were then analyzed to see if they are able to distinguish between features of retinas with plus disease and without. However, no significant clusters were found as most samples tended toward one cluster regardless of plus disease classification.

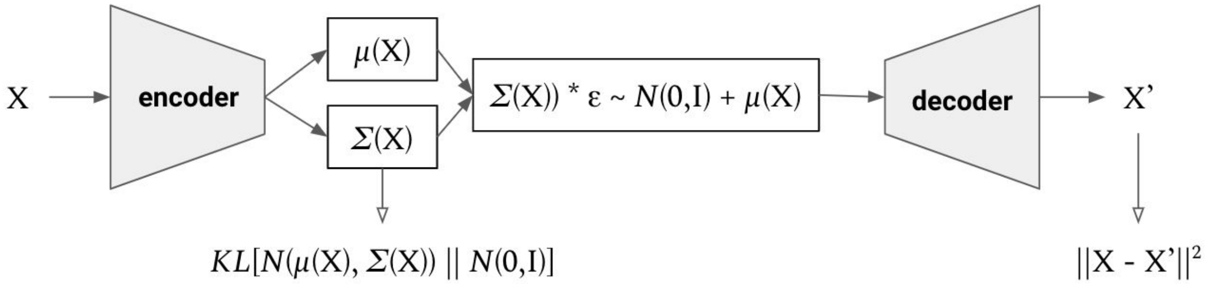


Figure 4. Variational Autoencoder (VAE) model diagram showing the key components of a VAE with the reparameterization trick and the calculations of KL divergence and pixelwise loss.

To find the latent space feature representation, I use a variational autoencoder (VAE) [28], trained normally to reconstruct images using pixelwise loss and additionally trained adversarially with a binary cross entropy loss for discrimination of fake generated images. The adversarial VAE takes important concepts from VAEs and GANs to learn a deep generative model that maps the input data distribution to a prior distribution [29, 30]. The particular prior distribution used is the multidimensional Gaussian with mean zero and diagonal identity covariance because this distribution is easy to sample from, thus allowing us to generate new samples easily and efficiently. This model was trained in two ways: with and without KL divergence in the calculation of the loss. Kullback–Leibler divergence (KL divergence) is an asymmetric measure of how different two probability distributions are from one another. Using KL divergence in the loss helps pull the input data distribution closer to the chosen prior as that prior allows us to easily sample over the whole distribution to generate new samples. Without this, the latent space can still learn, but there is no way of generating new samples from it since we don't know the distribution that the space lies in. Both ways to train the model aim to minimize the reconstruction error and improve the ability to discriminate between true and generated images, therefore being able to capture the most important features in the latent representation for plus disease. The adversarially trained model 1) learns a lower dimensional latent feature representation of the segmented retinal images, and 2) serves as a generative model that has the ability to reconstruct and generate meaningful images from the prior distribution. Model diagrams for a VAE and an adversarially trained VAE are shown in Figures 4 and 5.

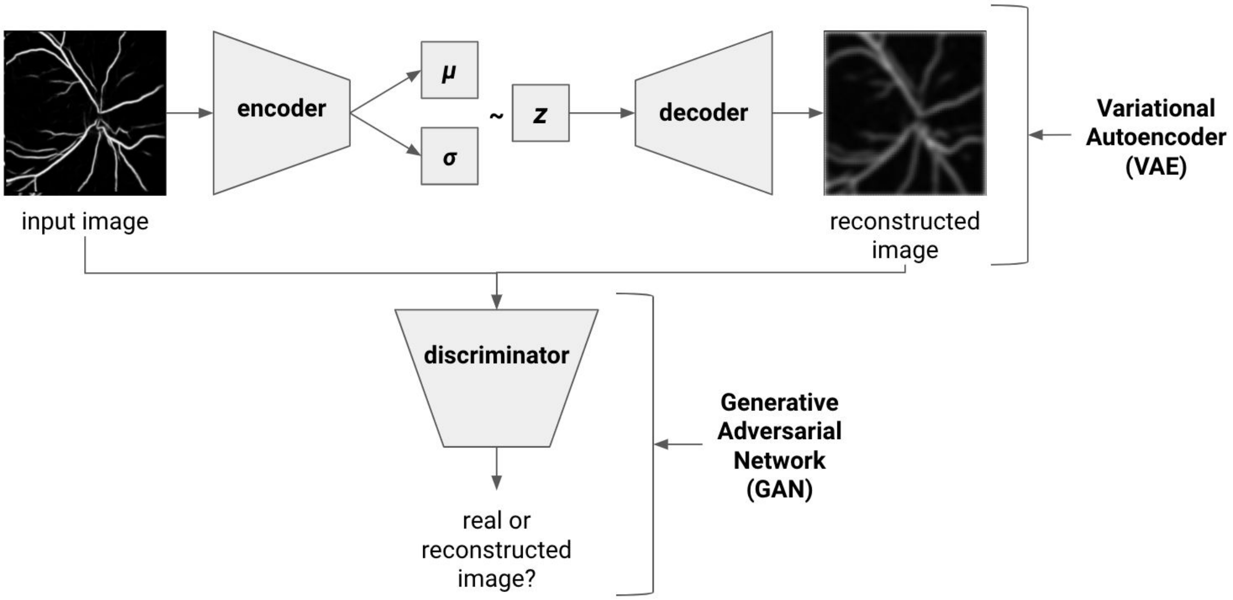

input image
encoder
μ
σ
~
z
decoder
reconstructed image
Variational Autoencoder (VAE)
discriminator
real or reconstructed image?
Generative Adversarial Network (GAN)

Figure 5. Model diagram of an adversarially trained VAE (or VAEGAN) used to find a lower dimensional latent space *z*.

Two latent space representations were experimented with: a 1D linear latent space of size 500 and 10,000 and a 3D space latent space of size (8, 64, 64). These provided low dimensional feature representations of the vessel map of the retina which is originally (3, 256, 256). The models were trained with Adam optimizer using default parameters for both the generator and discriminator. The models were trained for 20 epochs each.

After obtaining the latent space representation, the small classification network is trained on those features and evaluated for its ability to diagnose plus disease. This work will automate diagnosis and prediction in the case of inadequate or variable data as well as lessen the impact of variability by using objective methods.

4.3 Evaluation

Evaluation to understand the effectiveness of each algorithm is imperative. The evaluation metrics implementations used were primarily from the scikit-learn library of Python. Additionally, evaluation not only includes these quantitative metrics but also the ability to use these developed models in areas that diagnose and learn about ROP plus disease. It also shows what can and cannot be learned about a premature infant from a retinal image.

4.3.1 Age Prediction Evaluation Metrics

For assessing the classification of extremely premature versus infants born later in the pregnancy, the metrics of accuracy, precision, recall, area under the receiver operating characteristic curve (AUROC), and confusion matrix were calculated on the test set. For evaluating the test set results of the regression on postmenstrual age, mean squared error (MSE), explained variance, R^2 (coefficient of determination) score were the evaluation metrics.

4.3.2 Plus Disease Prediction Evaluation Metrics

Obtaining a low dimensional feature space from the VAE and adversarially trained VAE is an unsupervised method. The autoencoder provides a pixelwise loss between the original images and the decoded reconstructed images from the found latent representation. This loss, along with qualitative examination of the images, determines how well images can be reconstructed as well as generated using the decoder arm of the VAE model. The work of the discriminator in the GAN portion of the network also provides information as to how realistic a reconstructed or generated image is.

The classification of plus disease is a supervised algorithm between the latent feature representation and the provided plus disease labels as it uses a small CNN. The ability of the network using the latent representation to classify plus disease is evaluated by accuracy, precision, recall, f1 score, and a confusion matrix.

Chapter 5

Results

Automating the process of discovering and understanding retinal disorders and features can greatly benefit many patients and clinicians. In this work, I analyze three tasks: prediction of gestational age, prediction of postmenstrual age, and finding a low dimension representation for features to classify plus disease. The methods used are outlined above.

5.1 Prediction of Gestational Age

Gestational age was predicted as a binary classification problem for extremely premature versus moderate/late premature infants. In addition to proving that gestational age can be predicted from solely retinal images of an infant, it's valuable to understand how these predictions vary and fail among different disease classifications (healthy, pre-plus, and plus). To see how these predictions varied, two models were trained: one on only healthy eyes and one all eyes (both healthy and those with pre-plus and disease), and both were evaluated on a test set of both healthy and pre-plus and plus disease affected retinal images. If predictions are significantly different than the true age in infants with disease, this can be an indicator of disease. Both models have a similar accuracy of about 67% overall, but, as expected, the two models performed differently for each disease classification.

Results are shown in Tables 2 and 3, and Figures 6 and 7. Both models maintained about 67% accuracy in determining extremely prematurely born or not for all healthy retinas. However, regardless of the models being trained on pre-plus and plus disease affected retinas, both were able to tell with high accuracy that the plus disease retinas were extremely premature. Nearly all the pre-plus disease retinal images were also classified correctly as extremely premature. This indicates that retinas with disease appear younger than healthier retinas, possibly due to factors such as underdeveloped vessels.

	All Eyes	Healthy	Pre-Plus	Plus
Accuracy	0.6730	0.6723	0.6667	0.7500
Precision extremely (support) moderate/late	0.63 (129) 0.69 (186)	0.58 (111) 0.71 (185)	0.91 (14) 0.00 (1)	1.00 (4) (no samples)
Recall extremely moderate/late	0.49 (129) 0.80 (186)	0.45 (111) 0.81 (185)	0.71 (14) 0.00 (1)	0.75 (4) (no samples)
AUROC	0.6447	0.6279	0.3571	undefined

Table 2. Gestational age prediction evaluation results for model trained on only healthy retinal images and evaluated on all disease classifications of retinal images.

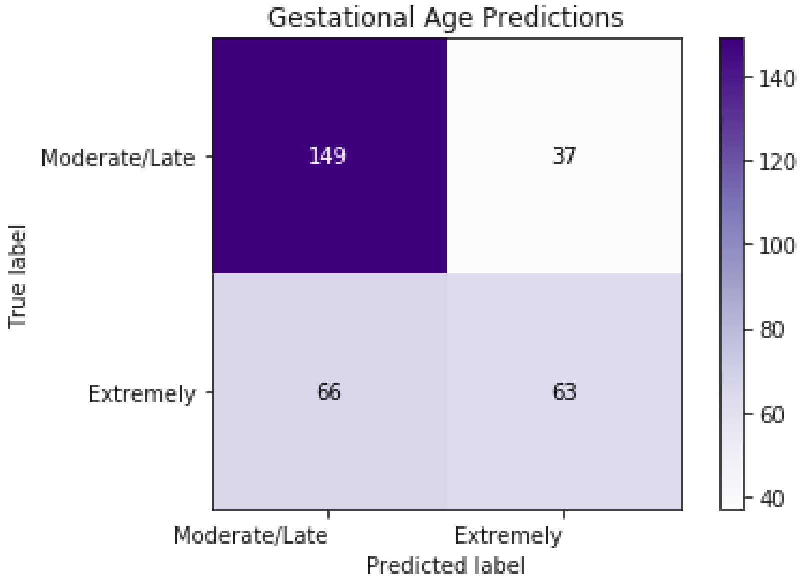


Figure 6. Gestational age prediction confusion matrix for model trained on only healthy retinal images and evaluated on all disease classifications of retinal images.

	All Eyes	Healthy	Pre-Plus	Plus
Accuracy	0.6794	0.6689	0.8000	1.00
Precision extremely (support) moderate/late	0.64 (129) 0.70 (186)	0.58 (111) 0.71 (185)	0.92 (14) 0.00 (1)	1.00 (4) (no samples)
Recall extremely moderate/late	0.50 (129) 0.80 (186)	0.44 (111) 0.81 (185)	0.86 (14) 0.00 (1)	1.00 (4) (no samples)
AUROC	0.6525	0.6234	0.4286	undefined

Table 3. Gestational age prediction evaluation results for model trained on and evaluated on all disease classifications of retinal images.

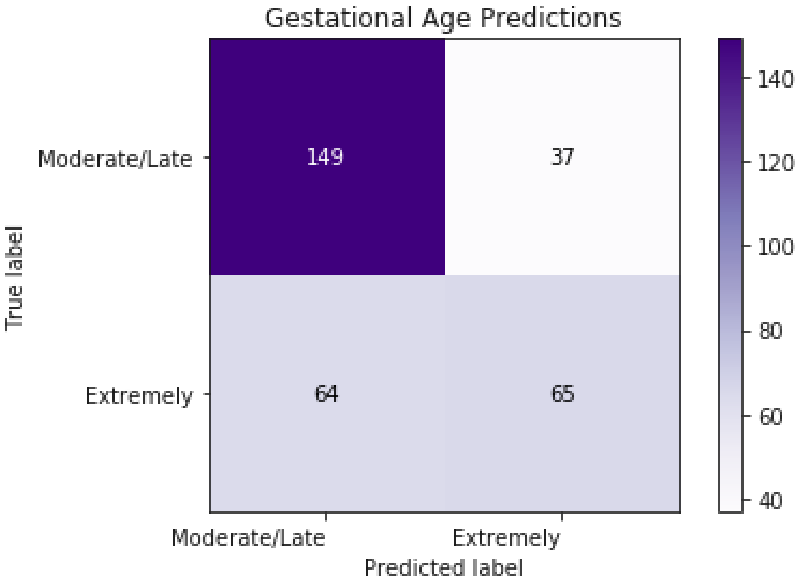


Figure 7. Gestational age prediction confusion matrix for model trained and evaluated on all disease classifications of retinal images.

5.2 Prediction of Postmenstrual Age

Postmenstrual age (PMA) for an infant was predicted directly with a regression neural network. Being able to predict the age for an infant solely based on retinal images shows that there are such indicators in the retina, even at a young age. It's also beneficial to interpret where predictions for age are over- and under-predicted for each disease classification. Again, to see how these predictions varied, two models were trained: one on only healthy eyes and one all disease classifications of eyes.

Results are shown in Tables 4 and 5, and Figures 8 and 9. There are correlations between the actual and predicted age as seen in the figures. For both models, infants with plus disease had retinas that indicated they were younger than they actually were. When trained on only healthy eyes, the model tended to predict that healthy eyes were usually a little older than their true age. Whereas when trained on all disease classifications of eyes, the model tended to predict that healthy retinal images were a little younger. For both models, retinas with pre-plus disease were predicted to be older than their true age.

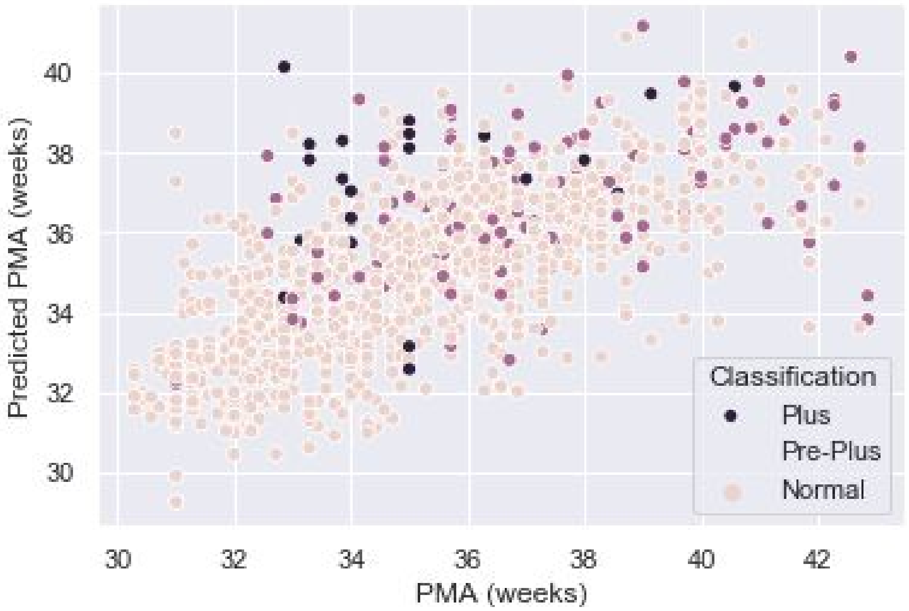


Figure 8. Prediction of postmenstrual age for model trained on healthy retinal images and evaluated on all disease classification retinal images.

	All Eyes	Healthy	Pre-Plus	Plus
MSE	4.4274	4.0808	6.1749	8.1076
R^2 Coefficient	0.4189	0.4440	0.1557	-0.8186
Explained Variance	0.4291	0.4578	0.1805	-0.2794
Mean / Median Error (true - pred)	-0.2789 / -0.2109 predicted older	-0.3184 / -0.2373 predicted older	-0.4263 / -0.1861 predicted older	+1.5505 /+ 1.7926 predicted younger

Table 4. Postmenstrual age evaluation results for model trained on healthy images and evaluated on all disease classifications of retinal images.

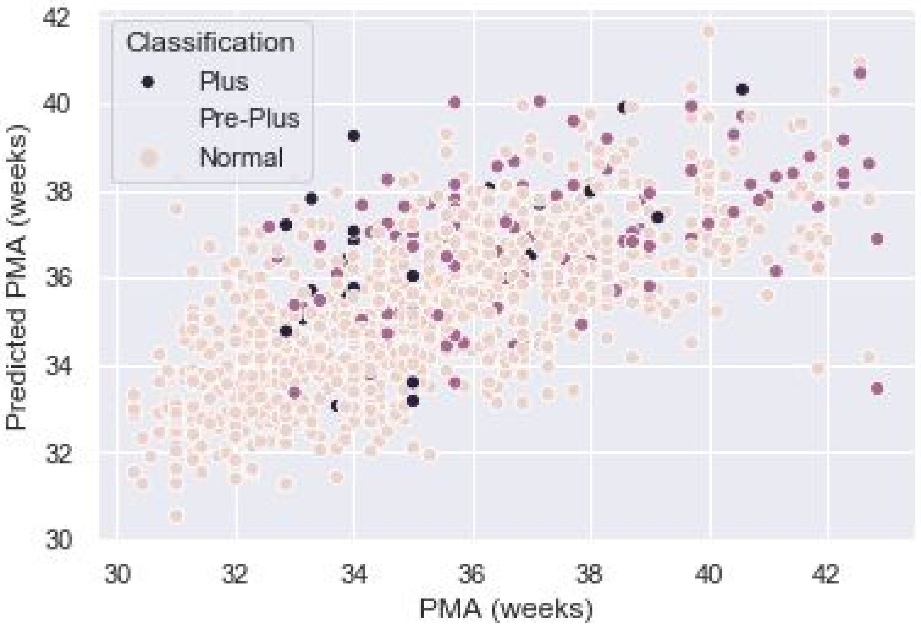


Figure 9. Prediction of postmenstrual age for model trained and evaluated on all disease classification retinal images.

	All Eyes	Healthy	Pre-Plus	Plus
MSE	4.1424	3.951	5.4866	4.6136
R^2 Coefficient	0.4563	0.4616	0.2498	-0.0349
Explained Variance	0.4564	0.4617	0.2612	0.2583
Mean / Median Error (true - pred)	+0.0100 / +0.1198 predicted younger	+0.0126 /+ 0.1163 predicted younger	-0.2892 / -0.0877 predicted older	+1.1432 /+ 1.1796 predicted younger

Table 5. Postmenstrual age evaluation results for model trained on and evaluated on all disease classifications of retinal images.

5.3 Prediction of Plus Disease

There are two steps to predicting plus disease using unsupervised feature extraction: 1) obtaining a low dimensional latent feature representation, and 2) finding likely disease classifications using a small classification network. The evaluation of the classification network shows how relevant the features are that are in the latent space. It was found that retinal image generation from a random Gaussian variable via the decoder was poor but image reconstruction was successful, meaning the latent feature representation carried important retinal information that could be translated by the decoder.

5.3.1 Effect of KL Divergence in the Loss Function

With both a linear and spatial latent representation of plus disease features, it was found that KL divergence penalized in the loss function was not able to properly reconstruct or generate images as the loss was continually too great. As the network trained, the progression of reconstructed images went from blurry white dots on to simply black with no indication of a retinal image, vessels, or important features as seen in Figure 10. Below is the progression of reconstructed images as the adversarial VAE trained on both pixelwise loss and KL divergence with a spatial latent space. From here I determine that including KL divergence is not beneficial to later diagnosing plus disease, so it is omitted in the remaining analysis.

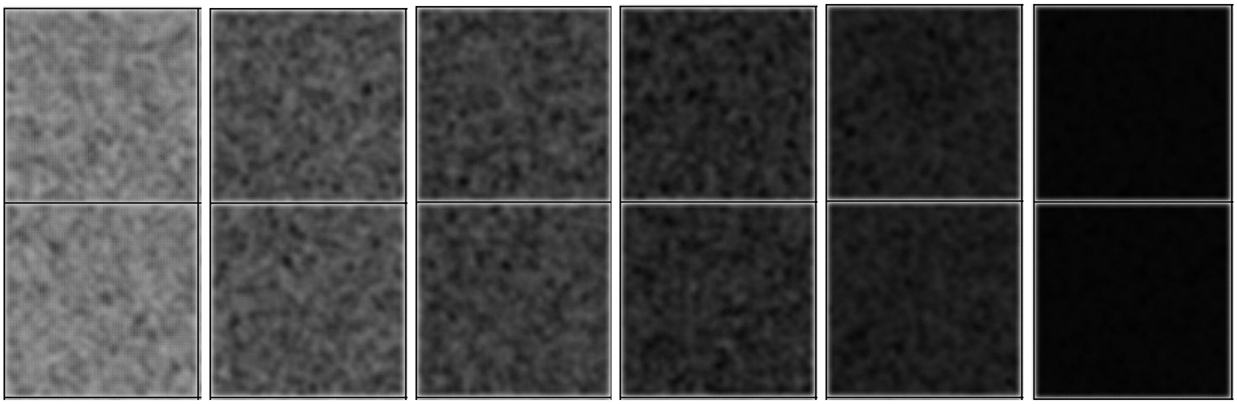


Figure 10. Progression of reconstructed images using the spatial adversarial VAE trained on pixelwise and KL divergence as the loss.

5.3.2 Effect of Linear versus Spatial Latent Space

The differences in retinal image reconstruction and generation from a linear latent space versus a spatial latent space for feature representation was also explored. A spatial latent space not only carries information about the features but more importantly their relative locations from the convolutions applied. Using a linear latent space does cause reconstructed retinal images to have some indicators of a retina, such as vessels which are imperative to diagnosing plus disease. However the majority of images had no clear vessels; they appeared shades of grey as shown in the first picture in Figure 11. Less than 5% of the reconstructed images displayed the rightmost reconstruction image, meaning that the linear latent space could carry some retinal information but it was not consistent. Figure 12 shows again that penalizing KL divergence was not able to produce vessels, only coloration, in recreating retinal images from their linear latent space feature representation.

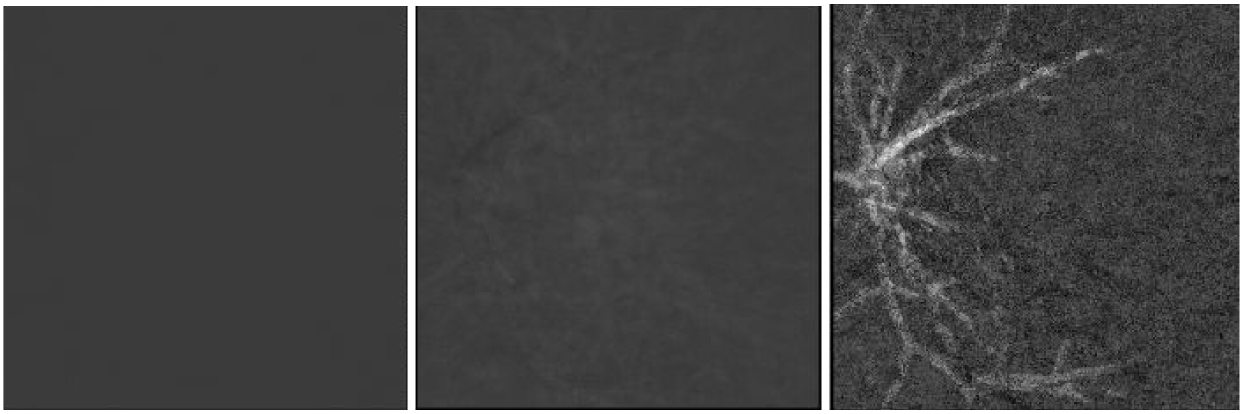


Figure 11. Examples of retinal image reconstruction with a linear latent space representation and penalizing only pixelwise loss.

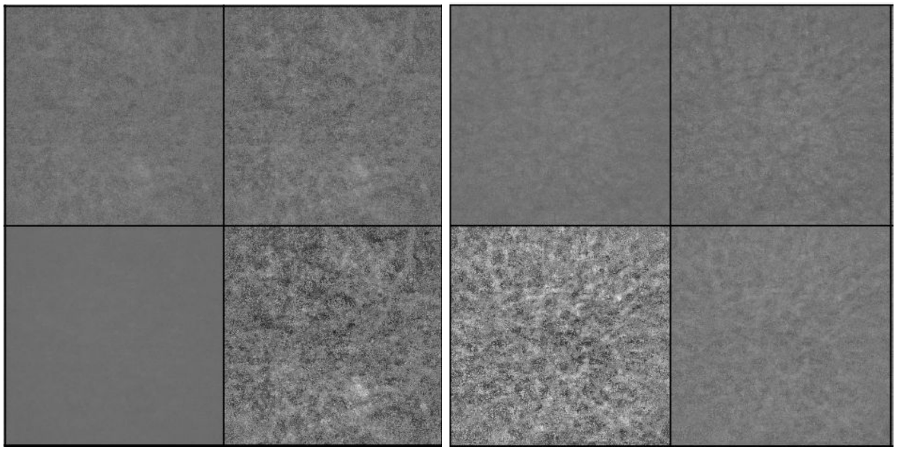
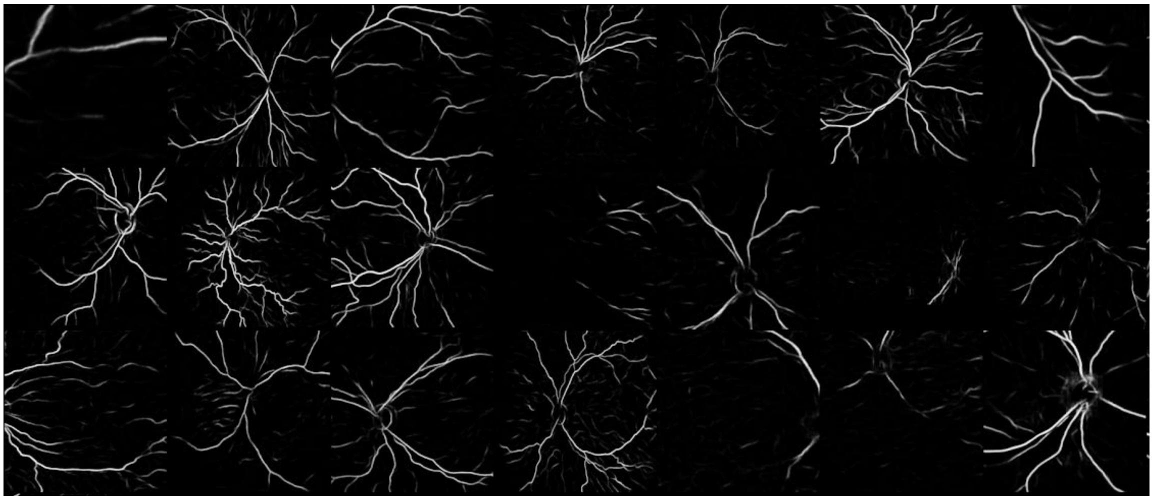


Figure 12. Retinal image reconstruction with a linear latent space representation and penalizing both pixelwise loss and KL divergence.

Using a 3D spatial latent space feature representation, however, had more promising results in image reconstruction. Since this spatial representation is higher dimensional than the linear representation and includes feature location information, the images were properly reconstructed, only losing some clarity and crispness from the original segmentations of vessels as shown in Figure 13. However, retinal image generation with a random Gaussian in this spatial feature representation did not create images with distinguishable vessels or any other features as shown in Figure 14.



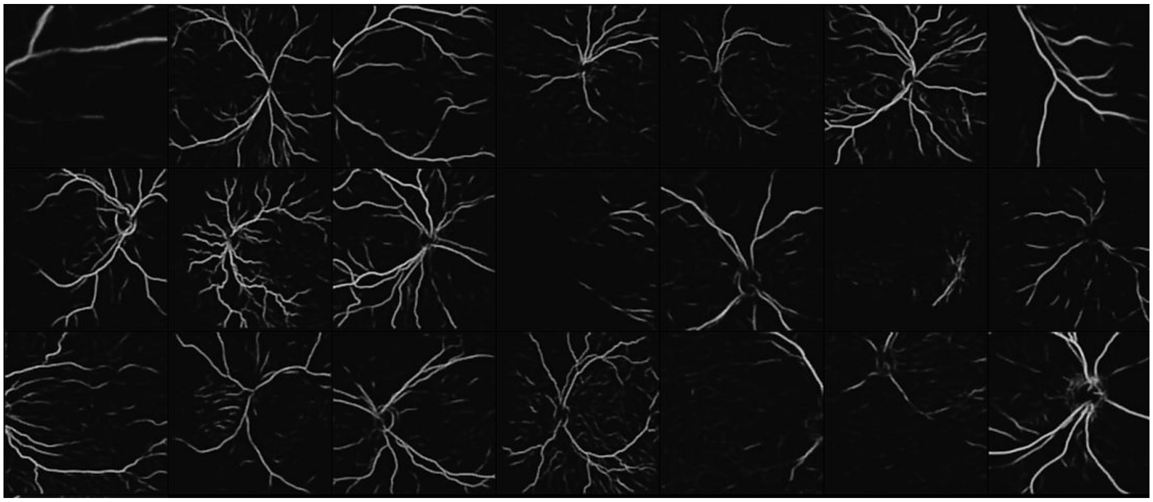


Figure 13. Retinal image reconstruction (bottom) of original image (top) from a spatial latent space using the VAE trained on pixelwise loss.

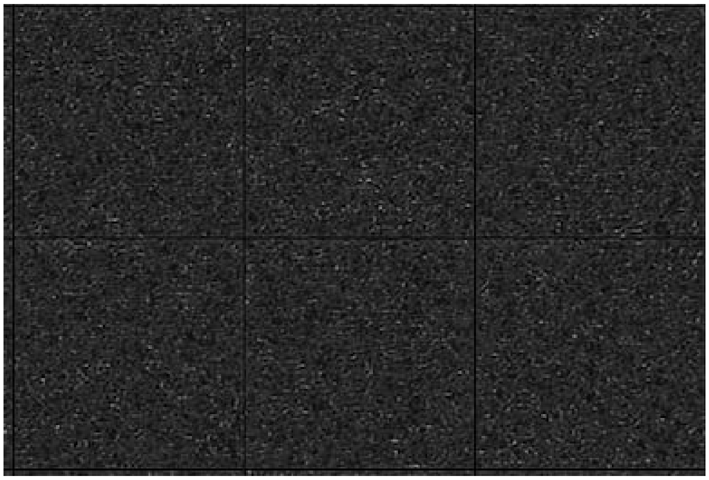


Figure 14. Generated retinal images from a random Gaussian decoded with a spatial latent space.

There is a significant difference in the ability of the autoencoder to perform image reconstruction versus generation. A retinal image is generated by sampling a random Gaussian and feeding it through the decoder. However, image generation performs poorly as the decoder is likely specialized to use selected retinal features via the encoder as opposed to a random Gaussian. Even with KL divergence adding to the penalty in training, the generation result remains nearly the same.

5.3.3 Effect of Adversarial Training on Image Reconstruction and Generation

The goal of adding the adversarial training is to teach the model which generated images are reasonable and close to the actual retinal images. However, training the VAE adversarially in this application seemed to take away from the clarity and crispness of the image and did not fully capture the tortuosity and dilation of vessels required for plus disease diagnosis. The image reconstruction and generation outputs using a spatial latent space without KL divergence are shown in Figures 15 and 16. Again, the generation of segmented images is poor and requires more exploration and experimentation with model architectures, loss functions, and other algorithms.

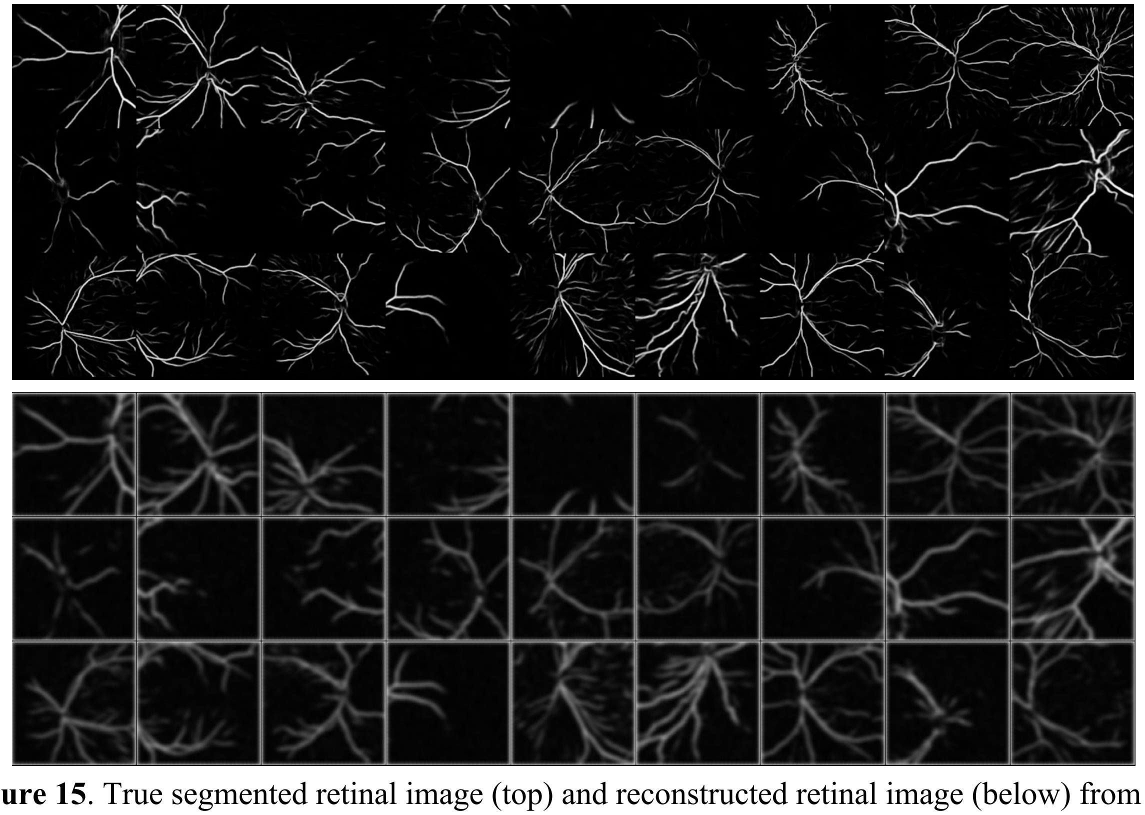


Figure 15. True segmented retinal image (top) and reconstructed retinal image (below) from the adversarially trained VAE.

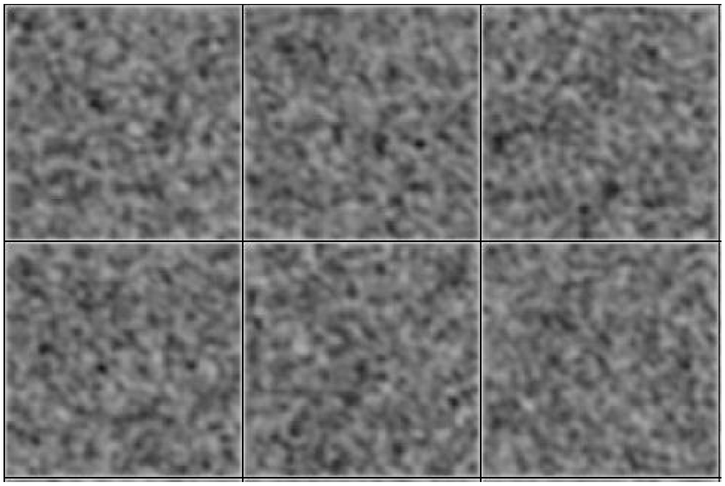


Figure 16. Generated images from a spatial latent space sampled from a Gaussian distribution using an adversarially trained VAE.

5.3.4 Plus Disease Classification Network Results

Once reasonable retinal image reconstruction can be achieved, it is assumed that the encoder and decoder of the VAE has learned a good latent lower dimensional feature representation that captures the most important features that can be used to diagnose plus disease. Now we look at the ability of a simple classification network to distinguish between different disease classifications using these feature representations. Based on findings in the sections above, Ia use images reconstructed from a VAE using a spatial feature latent space representation.

	precision	**recall**	**f1 score**	**support**
normal	0.86	0.79	0.82	892
pre-plus	0.27	0.47	0.34	182
plus	0.00	0.00	0.00	59
class weighted average	0.72	0.69	0.70	1133

Table 6. Results for plus disease classification on the low dimensional latent space.

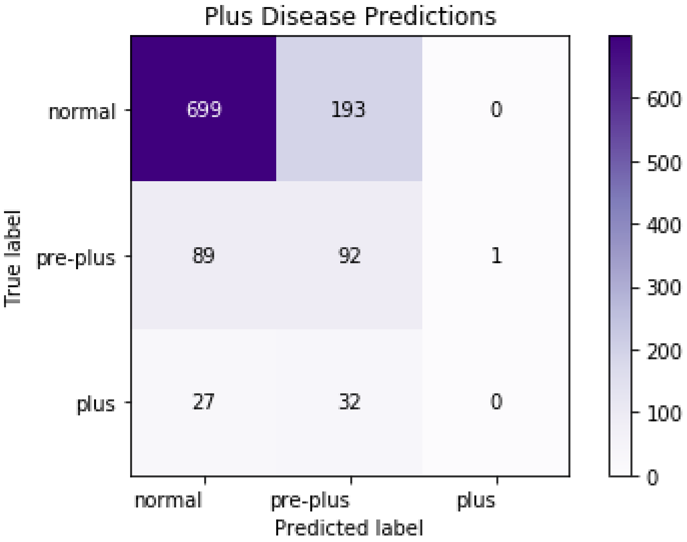


Figure 17. Plus disease classification confusion matrix on the test set.

The results in Table 6 and Figure 17 show that the classifier is good at determining if retinal images appear normal but poor at diagnosing pre-plus disease and isn't at all able to distinguish plus disease. The overall accuracy is 67.73%. This indicates that either the latent space lacks enough information about the tortuosity and dilation or that the simple classification network isn't powerful enough to capture these subtleties. During training, the network was able to pick out some plus disease affected retinas; however, this was not reflected in the test environment.

Chapter 6

Discussion

6.1 Impact of Age Prediction

Even in infants, there are biomarkers in the retina that express systemic features such as age. Results found through this analysis show that it is possible to predict an infant's gestational and postmenstrual age solely from retinal images. Despite having an accuracy of 67%, results show that there are, albeit weak, indicators in the retina that show this, and these indicators change with plus disease diagnosis. If an infant is found to be born extremely prematurely, diagnoses can be prioritized to inspect for plus disease. Being born extremely prematurely lends to biomarkers of an underdeveloped retina and therefore a younger appearing postmenstrual age and the possibility for plus disease.

It's also interesting to note that in this task, predicting age for pre-plus disease is more difficult than for a healthy or plus disease affected retina. This stems from the fact that the onset of plus disease and ROP in stages 1-3 is a spectrum, not a clear distinction. It's relatively easy to distinguish between a healthy retina and a plus disease affected retina, but diagnosing pre-plus disease is more difficult and variable, for both experienced clinicians and deep learning algorithms. The possible disagreement and variability among graders and clinicians can cause these types of results seen in age prediction and other tasks.

6.2 Unsupervised Feature Extraction for Plus Disease

Significant arterial tortuosity and venous dilation at the posterior pole of a retina characterize plus disease. It's important that any feature extraction captures the extent of this tortuosity and dilation. Whether or not a full segmented retinal image is required to automatically diagnose plus disease is discovered in this lower dimensional feature representation analysis. When the algorithm reconstructs (or decodes) a segmented retinal image from its (encoded) latent representation, some of these important features appear to be captured. Although currently the algorithm is not completely unsupervised, achieving 67% accuracy and 0.72 precision using the simple classification network shows that this feature extraction is conveys that it is possible to work toward more supervised methods in the lower dimension and other unsupervised methods. However, more work needs to be done to improve the classification ability as it's more beneficial to clinicians and patients to diagnose plus disease than correctly determine a normal retina.

Ideally a fully unsupervised algorithm would include an additional step or algorithm after obtaining the latent representation, such as clustering for disease classifications. However, since this latent space in this work is sampled from a single Gaussian as opposed to three for the three disease classifications, this isn't currently possible. Overall, using unsupervised methods to encode an image into a low dimensional representation allows for fewer labeled samples in a dataset and also lessens the effect of possible variability in the labeling.

6.3 Future Work

This analysis was conducted with the goal of providing more accurate and timely automatic disease diagnosis and information about infants at risk for ROP. These models were trained and evaluated on one dataset from the United States, so it would be valuable to demonstrate the algorithm's effectiveness across datasets with diverse populations as well as potentially incorporating more patients' retinal images and information into the training process. Applying this algorithm in a real world setting with a clinician to assess its effectiveness in providing diagnoses and information is also important to show its capability to be beneficial in medical settings.

As always with deep learning models, further work and enhancement can be conducted on the architecture, tuning, and feature extraction to further improve results for age prediction. In particular, feature extraction may be able to reveal not only if biomarkers of age and other features exist but what and where they are in the retina itself.

In addition, further exploration of unsupervised learning methods is also beneficial in the complete absence of labeled data. To examine with what and where a retina goes from healthy to pre-plus and plus disease, change detection algorithms are also an algorithm that can help to better understand retinal diseases. To further the current analysis, examining why KL divergence causes the creation of the latent space and retinal image reconstruction and generation to perform so poorly is important as this is a key part of VAEs. Experimenting further with adversarial training would also be beneficial to reconstruct and generate even better images with a smaller dimension. GANs are intended to generate images that fool the network on their realisticness, but in this analysis, this was not the case. Testing different loss functions can also provide insight into the reason behind the quality of the reconstructed images [31].

Exploring other feature extraction methods is also valuable in understanding and diagnosing plus disease. For example, principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP), are popular extraction methods. These methods can also capture multiple axes of the data distribution that could capture the separation between the features of the healthy and diseased retinas.

Chapter 7

Conclusion

The goal of this work is to provide more accurate and timely automatic plus disease diagnosis and prediction of systemic features for premature infants at risk for ROP. With nearly 6,000 labeled retinal images and deep learning algorithms, this thesis shows and understands that there are clear biomarkers in the retinas of infants that indicate various clinical variables. By being able to predict gestational age, we can prioritize diagnosis for plus disease, ensuring the infants most at risk are helped first. Understanding the differences in prediction of postmenstrual age in infants with and without plus disease help clinicians better understand the effects of disease on the retina. Being able to predict plus disease using unsupervised machine learning algorithms not only provides diagnoses quickly and accurately but also lessens the effect of possible variability and subjectivity in the way clinicians diagnose and label retinal images for plus disease in ROP. This work contributes to preventing serious disease progression and further understanding the biomarkers of the disease to provide better health as well as lays a framework to use similar methods for other medical conditions.

References

[1] Abràmoff MD, Garvin MK, Sonka M. Retinal imaging and image analysis. IEEE Rev Biomed Eng. 2010;3:169-208.

[2] Braverman RS, Enzenauer RW. Socioeconomics of retinopathy of prematurity in-hospital care. Arch Ophthalmol. 2010;128(8):1055-1058.

[3] Wallace DK. Fellowship training in retinopathy of prematurity.J AAPOS. 2012;16(1):1.

[4] Wong RK, Ventura CV, Espiritu MJ, et al. Training fellows for retinopathy of prematurity care: a web-based survey.J AAPOS. 2012;16(2):177-181.

[5] Nagiel A, Espiritu MJ, Wong RK, et al. Retinopathy of prematurity residency training. Ophthalmology. 2012;119(12):2644-2645.e1, 2.

[6] Krause J, Gulshan V, Rahimy E, et al. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. Ophthalmology. 2018;125(8):1264-1272.

[7] Revised Indications for the Treatment of Retinopathy of Prematurity Results of the Early Treatment for Retinopathy of Prematurity Randomized Trial. Arch Ophthalmol. 2003;121(12):1684–1694. doi:10.1001/archopht.121.12.1684.

[8] Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. Arch Ophthalmol. 2007; 125(7):875-880.

[9] Kalpathy-Cramer J, Campbell JP, Erdogmus D, et al; Imaging and Informatics in Retinopathy of Prematurity Research Consortium. Plus disease in retinopathy of prematurity: improving

diagnosis by ranking disease severity and using quantitative image analysis. Ophthalmology. 2016;123(11):2345- 2351.

[10] Ataer-Cansizoglu E, Bolon-canedo V, Campbell JP, et al. Computer-Based Image Analysis for Plus Disease Diagnosis in Retinopathy of Prematurity: Performance of the "i-ROP" System and Image Features Associated With Expert Diagnosis. Transl Vis Sci Technol. 2015;4(6):5.

[11] Campbell JP, Ataer-Cansizoglu E, Bolon-canedo V, et al. Expert Diagnosis of Plus Disease in Retinopathy of Prematurity From Computer-Based Image Analysis. JAMA Ophthalmol. 2016;134(6):651-7.

[12] Peng tian, Ataer-Cansizoglu E, Kalpathy-cramer J, et al. Toward a severity index for ROP: An unsupervised approach. Conf Proc IEEE Eng Med Biol Soc. 2016;2016:1312-1315.

[13] Brown JM, Campbell JP, Beers A, et al. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. JAMA Ophthalmol. 2018;136(7):803–810. doi:10.1001/jamaophthalmol.2018.1934

[14] Brown JM, Campbell JP, Beers A, Chang K, Donohue K, Ostmo S, Chan RVP, Dy J, Erdogmus D, Ioannidis S, Chiang MF, Kalpathy-Cramer J. Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning. In Proc. SPIE 10579, Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, 2018;105790Q. doi: 10.1117/12.2295942.

[15] Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2(3):158-164. doi:10.1038/s41551-018-0195-0.

[16] Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of Convolutional Neural Networks for Automated Classification of Chest Radiographs. Radiology. 2019;290(2):537-544.

[17] Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316(22):2402–2410. doi:10.1001/jama.2016.17216.

[18] Ting DSW, Cheung CY, Lim G, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA. 2017;318(22):2211–2223. doi:10.1001/jama.2017.18152.

[19] Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. Ophthalmology. 2018;125(8):1199-1206. doi:10.1016/j.ophtha.2018.01.023.

[20] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Cham, Switzerland: Springer International Publishing; 2015;234-241.

[21] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Paper presented at: 2015 Institute of Electrical and Electronics Engineers Conference on Computer Vision and Pattern Recognition; June 7-12, 2015; Boston, MA.

[22] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. Paper presented at: 2009 Institute of Electrical and Electronics Engineers Conference on Computer Vision and Pattern Recognition; June 20-25, 2009; Miami, FL.

[23] Moriya, T., Roth, H. R., Nakamura, S., Oda, H., Nagara, K., Oda, M., & Mori, K. Unsupervised segmentation of 3D medical images based on clustering and deep representation learning. Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging. 2018;161-169. doi:10.1117/12.2293414.

[24] Chen, M., Shi, X., Zhang, Y., Wu, D., & Guizani, M. Deep Features Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network. IEEE Transactions on Big Data, 1-1. 2017; doi:10.1109/tbdata.2017.2717439.

[25] Ryan MC, Ostmo S, Jonas K, et al. Development and Evaluation of Reference Standards for Image-based Telemedicine Diagnosis and Clinical Research Studies in Ophthalmology. AMIA Annu Symp Proc. 2014;2014:1902-10.

[26] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? Paper presented at: 27th International Conference on Neural Information Processing Systems; December 8-13, 2014; Montreal, Québec, Canada.

[27] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision, 2016. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2818-2826.

[28] Kingma, Diederik P and Welling, Max. Auto-Encoding Variational Bayes. In The 2nd International Conference on Learning Representations (ICLR), 2013.

[29] Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial autoencoders. International Conference on Learning Representations (ICLR), arXiv:1511.05644, San Juan, 2016.

[30] Baur C, Wiestler B, Albarqouni S, Navab N. Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images. BrainLesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes, 2018. doi:10.1007/978-3-030-11723-8_16.

[31] Isola P, Zhu J, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks, 2016. in arXiv:1611.07004.