MACHINE LEARNING
FOR DISEASE DETECTION AND PREDICTION
IN RETINOPATHY OF PREMATURITY

By

Aaron S. Coyner

A DISSERTATION

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
in
Bioinformatics and Computational Biomedicine

February 2021

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD dissertation of

Aaron S. Coyner

has been approved.

_____

Michael F. Chiang, MD, MS – Advisor

_____

Kemal Sönmez, PhD – Chair

_____

Jayashree Kalpathy-Cramer, PhD – Member

_____

J. Peter Campbell, MD, MPH – Member

_____

Ted Laderas, PhD – Member

TABLE OF CONTENTS

DEDICATION

I am nearly equal parts my mother and father.
I'm now old enough to understand how lucky that makes me.

To my extremely loving, supporting, and patient partner, Jessica:
Thank you so much for, y'know, being so loving, supporting, and patient :P

ACKNOWLEDGEMENTS

I would sincerely like to thank:

The Department of Medical Informatics and Clinical Epidemiology for all the knowledge they have bestowed upon me, and the skills they have taught me.

All members of my Dissertation Advisory Committee — Drs. Chiang, Sönmez, Kalpathy-Cramer, Campbell, and Laderas — for their guidance in planning and executing the research presented herein.

The other members of the Casey Eye Institute Ophthalmic Informatics Lab — Susan Ostmo, Michelle Hribar, Ryan Swan, Jimmy Chen, and Adam Rule — for letting me bounce ideas off them and helping me with various aspects of this work.

Finally, a huge debt of gratitude to the National Library of Medicine for providing me with a training grant that has funded me throughout the years and made all this possible.

INTRODUCTION

An estimated one in ten babies, 15 million worldwide, are born prematurely each year — prematurity being a significant risk factor for the development of retinopathy of prematurity (ROP), a potentially-blinding disorder of the retinal vasculature.[1–3] Despite the existence of known risk factors and many effective treatment options, ROP remains one of the world's leading causes of childhood blindness. This is primarily due to a combination of the scarcity of ROP experts and advances in neonatal intensive care units, which have increased the survival rate of younger, smaller infants and, consequently, the incidence and prevalence of ROP.[4,5]

To address the lack of ROP care in rural areas and developing countries, telemedicine pipelines and automated methods for the diagnosis of ROP have been proposed.[6–9] However, there are two primary issues that reduce their practicality. First, both require high quality retinal fundus images for diagnosis.[7,9] Imaging technicians are thoroughly trained, but image quality metrics must still be applied, as the cost of a missed diagnosis due to poor image quality or an incorrect field-of-view could be a lifetime of blindness. Unfortunately, basic image quality metrics do not capture the nuances of what constitutes a diagnosable retinal fundus image for ROP and remains an open area of research. Second, to ensure near 100% sensitivity to TR-ROP, these exams are performed weekly.[7,9] However, roughly 85–90% of those screened never develop TR-ROP, and weekly examinations not only increase the screening burden, but also the physiological stress placed on these extremely fragile infants. ROP risk models have been developed to reduce the screening burden, but often suffer from a lack of 100% sensitivity, or very low specificity.[1,2,10,11] This also remains an active area of research. Finally, even if a risk model predicted that a child was not going

to develop TR-ROP, they would still need to be screened once every two to three weeks. In areas without direct access to ROP experts, this could be accomplished by a pediatric ophthalmologist, so long as they knew what features of ROP were worrisome. Unfortunately, the current reference standard image for TR-ROP is highly-outdated — it is blurry and the field-of-view is no longer the standard.[1,3] Therefore, personalized reference standard images of TR-ROP for individual eyes would be ideal. Herein, we address these issues via the following specific aims:

Aim 1: Quality Control for Retinal Fundus Images — Ensure that retinal fundus images used for telemedicine and the automated diagnosis of ROP are of sufficient quality for an accurate and reliable diagnosis. This will be accomplished by training a convolutional neural network to detect which images are acceptable for the diagnosis of ROP and those which are not.

Aim 2: Prediction of Treatment-Requiring ROP Patients — Develop a risk model that, with near-perfect sensitivity, can predict which infants will develop treatment-requiring ROP, while simultaneously reducing the screening burden by maintaining high specificity with a 100% negative predictive value. This will be accomplished by using a novel deep learning-based vascular severity score.

Aim 3: Development of Personalized Reference Standard Images — Assist non-ROP experts with the identification of TR-ROP by synthesizing personalized reference standard retinal fundus images of TR-ROP. This will be accomplished using a series of generative adversarial networks that: (1) segment retinal fundus images into retinal vessel maps, (2) augment the vascular severity of said vessel maps to appear as TR-ROP, and then (3) convert augmented vessel maps back into retinal fundus images.

It is our goal to solve some of the issues associated with the accurate and timely diagnosis of TR-ROP. A robust image quality algorithm will be extremely useful for both telemedicine pipelines and the automated diagnosis of ROP. Such an algorithm could easily be implemented in retinal fundus cameras or the computers to which they are attached, so that imaging technicians could instantly be alerted as to whether captured images were of high enough quality. A risk model will theoretically predict all subjects who will develop TR-ROP, while correctly ruling out more than half of those who will not. Finally, those who are deemed low risk will still require follow-up examinations, albeit far less frequently. To reduce the screening burden further, these exams could be performed by non-experts using personalized reference standard images, which make it far easier for them to identify the features associated with TR-ROP.

BACKGROUND

RETINOPATHY OF PREMATURITY

Retinopathy of prematurity (ROP) is a potentially-blinding disorder of the retinal vasculature that affects premature infants, particularly those born prior to 31 weeks of gestation that weigh less than 1251 grams at birth.[1,3] For reference, a full-term pregnancy has a gestation period of 38 to 42 weeks, and the average birth weight, in the United States, hovers around 3500 grams.[12,13] Life-saving techniques for prematurely-born children often include some form of oxygen therapy.[12,14,15] However, oxygen (or lack thereof) is the main catalyst behind normal fetal retinal development — the retina being the layer of tissue on the back of the eye that supports photoreceptor cells, which are responsible for vision. At around 16 weeks of gestation, the blood vessels of the retina begin to form at the optic nerve. Because the womb is constantly in a hypoxic state, the retinal blood vessels slowly begin to grow outward in an attempt to supply oxygen and nutrients to the peripheral retina. Around 26 to 30 weeks of gestation, the eye begins to develop at an increased pace and, once born, the presence of oxygen signals to the retina to cease development. However, ROP patients, by definition, are born before or during this period of rapid vessel development. In order to preserve life, they are immediately placed in oxygen incubators upon birth; this completely inhibits the growth of the un- or under-developed retinal blood vessels. However, once these infants are deemed healthy enough to be removed from incubation, the lack of oxygen in the peripheral retina, due to the absence of peripheral retinal vasculature, signals vascular endothelial growth factor (VEGF) to promote the growth of new retinal blood vessels. In some cases, these vessels grow in an uncontrolled manner. These weak and fragile vessels eventually rupture and

bleed. The body naturally responds to stop the bleeding via scarring. When the scars shrink, they can pull on the retina and cause it to detach from the back of the eye, severing the neural connections that are required for vision, thus resulting in permanent visual impairment.

Features of Retinopathy of Prematurity

Fortunately, physicians have identified various clinical traits that are indicative of an eye that has, or is developing, moderate or severe ROP. There are three clinically-defined features of ROP: zone, stage, and plus disease.

Zone

Zone describes where and to what extent ROP is occurring (**Figure 1**).[1,3] This is accomplished by locating the ROP and its proximity to the optic nerve. There are three concentric zones in which ROP may occur. Zone I is described as a circle, centered on the optic nerve, that has a radius twice the distance of that from the optic nerve to the macula. Zone II is also centered on the optic nerve; however its radius extends to the nasal-ora serrata (the 3-o'clock position in the right eye and the 9-o'clock position in the left eye). Zone III is the residual crescent of retina unaccounted for by Zone I and Zone II. These zones are non-overlapping. For example, Zone II only includes portions of the retina not included in Zone I or Zone III. The extent of disease is recorded as the number 30° sectors, or clock-hours, that contain disease, and the boundaries between sectors lie on the clock hour positions. In general, the closer ROP occurs to the optic nerve and the macula, the more worrisome it is. This is because its potential to disturb the macula — the cone photoreceptor-dense region responsible for keen eyesight and color vision — and the optic nerve — responsible for transmitting all light gathered by photoreceptor cells to the brain to form an image — is far greater. Therefore, ROP found in Zone I is more alarming than ROP found in Zone II or Zone III. While

Zone III ROP is not ideal, its potential to disturb the optic nerve and macula is far less, and typically only rod photoreceptor cells — those responsible for both night and peripheral vision — are disturbed.

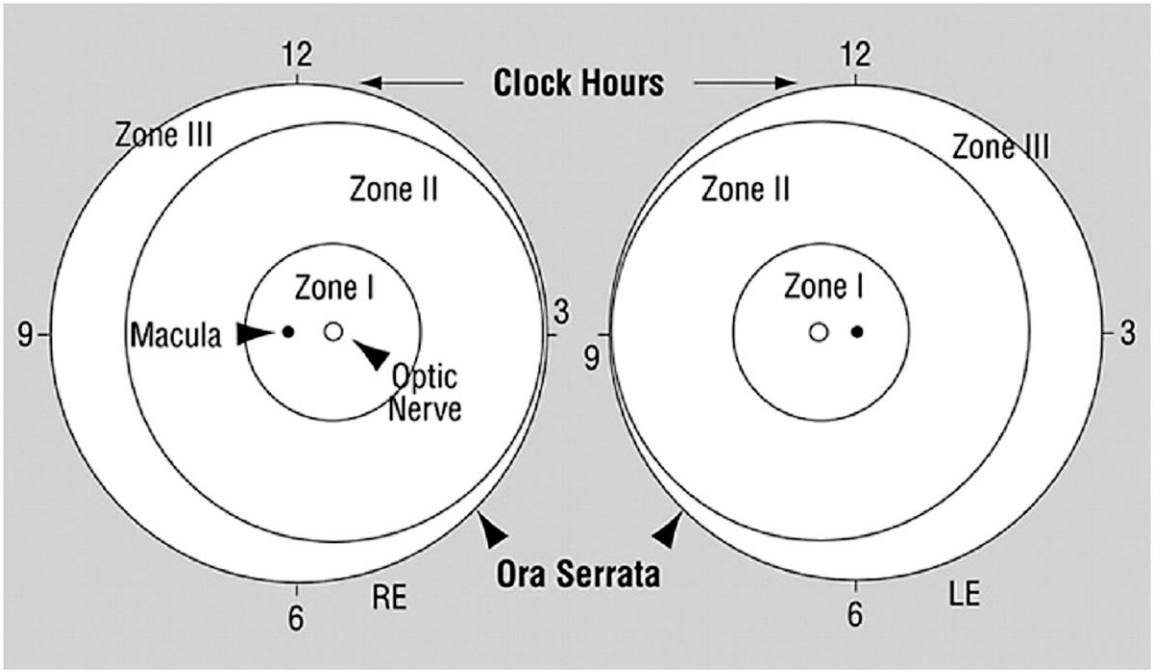

Figure 1: Scheme of retina of right eye (RE) and left eye (LE) showing zone borders and clock hours used to describe the location and extent of retinopathy of prematurity. Figure adapted from the The International Classification of Retinopathy of Prematurity Revisited [1].

Stage

While Zone describes where ROP is located, Stage describes to what extent it is occurring.[1,3] ROP is a result of incomplete vascular development; the appearance of the junction between the vascularized and avascular retina can be described as fitting into one of five different stages (**Figure 2**). Stage 1 ROP is described as a thin, but definite line of demarcation that separates the vascularized and avascular retina (**Figure 2A**). Stage 2 ROP is defined as a ridge (**Figure 2B**). It arises in the region of the demarcation line, but unlike the thin, flat line in Stage 1, it now has

height and width, and extends above the plane of the retina. Retinal blood vessels that meet the ridge may leave the plane of the retina posterior to the ridge to enter it. Small isolated tufts of neovascular tissue lying on the surface of the retina may be seen posterior to this ridge structure (see arrows in right image of **Figure 2B**). Stage 3 ROP begins to get more complicated. In stage 3, extraretinal fibrovascular proliferation or neovascularization extends from the ridge into the vitreous, the area of the eye anterior to the retina (**Figure 2C**). Typically, the ridge appears more ragged as this proliferation becomes more severe. The severity of Stage 3 is subdivided into mild, moderate, and severe depending on the extent of extraretinal fibrovascular tissue infiltrating the vitreous. Stage 4 ROP is defined as partial retinal detachment (**Figure 2D**). It can be subdivided into extrafoveal (Stage 4A) and foveal (Stage 4B) detachments. The extent of these detachments depends on how many clock hours of the retina the detachment is occurring and to what degree the scarring is pulling on the retina. Stage 5 ROP is defined as total retinal detachment. As the name suggests, this means that the entire retina has been torn away from the choroid (the tissue layer behind the retina responsible for supplying nutrients), and all neural connections severed.

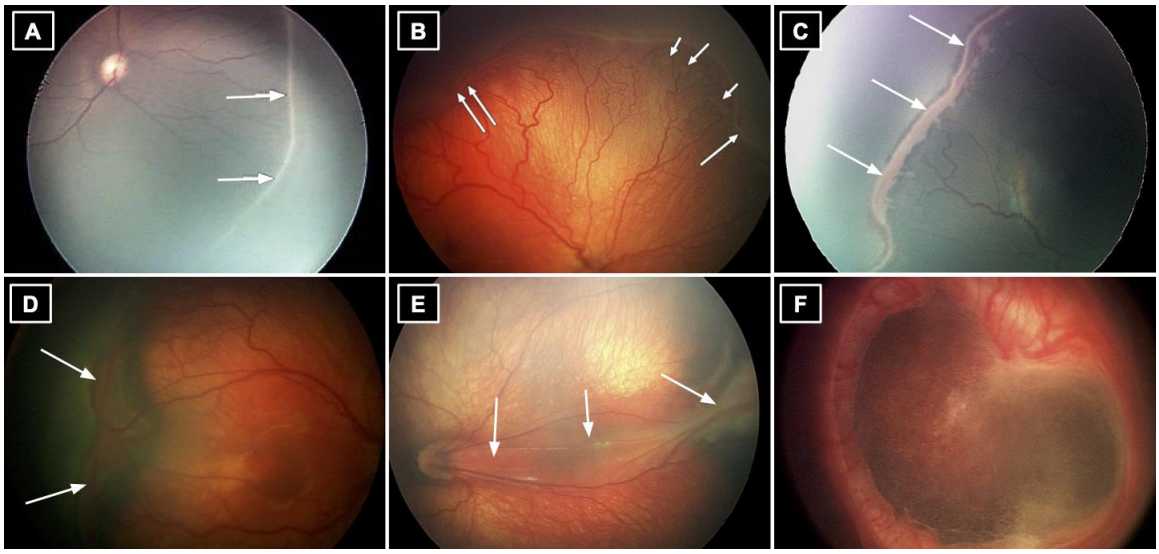

Figure 2: Examples of the various stages of retinopathy of prematurity. Arrows point to clinical features found in (A) Stage 1 ROP, (B) Stage 2 ROP, (C) Stage 3 ROP, (D) Stage 4A ROP, (E) Stage 4B ROP, and (F) Stage 5 ROP. Figure adapted from the The International Classification of Retinopathy of Prematurity Revisited.[1]

Plus Disease

Finally, the presence of plus disease, defined as venous dilation and arterial tortuosity, also helps determine the severity of ROP (**Figure 3C**).[1,3] Plus disease may later increase in severity to include iris vascular engorgement, poor pupillary dilatation, and/or vitreous haze. Plus disease need not be present in every portion of the retina for a diagnosis to be made — only two quadrants of the retina need to show plus disease for the entire eye to be diagnosed as such. As is common with many diseases, there is a spectrum of severity, and plus disease is no exception. Although physicians and researchers have attempted to binarize plus disease, there remains an in-between condition known as pre-plus disease (**Figure 3B**). This describes vessels that do not appear normal (**Figure 3A**), but do not have severe enough venous dilation and arterial tortuosity

to warrant a plus disease diagnosis. Over time, the vessel abnormalities associated with pre-plus disease can progress to plus disease as the vessels dilate and become more tortuous.

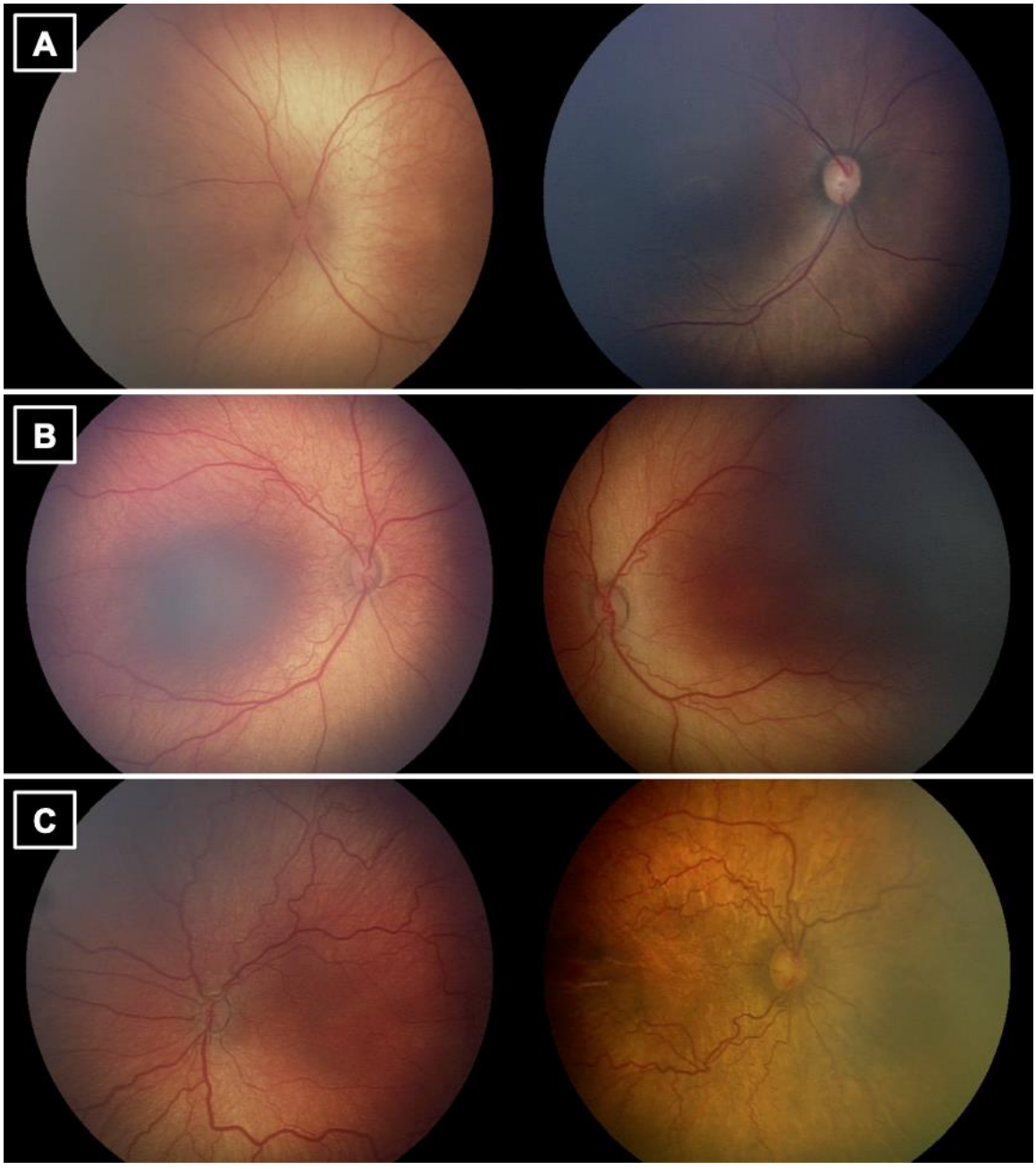

Figure 3: Examples worsening retinal vasculature. (A) Normal retinal blood vessels. They are relatively thin and straight. (B) Retinal blood vessels with pre-plus disease. These vessels are not

normal, but not severe enough to be diagnosed as plus. (C) Retinal blood vessels with plus disease. Note the extreme dilation and tortuosity.

Diagnosis of Retinopathy of Prematurity

Together, these features assist physicians in forming an ROP diagnosis.[1,3,13,16] Possible diagnoses, in order of severity, are: none, mild, Type-2, or Type-1 ROP. Although "none" implies that no ROP was found, it does not mean that a retina is healthy. It simply means that there is no visible demarcation line between the vascularized and avascular retina. Type-1 ROP, also known as treatment-requiring (TR-) ROP, is diagnosed given the following retinal finding:

- Zone I: any stage with plus disease
- Zone I: stage 3 without plus disease
- Zone II: stage 2 or 3 with plus disease

Upon a diagnosis of TR-ROP, treatment is typically initiated within 48–72 hours, if not immediately. Type-2 ROP, also known as moderate ROP, is diagnosed given the following retinal findings:

- Zone I: stage 1 or 2 without plus disease
- Zone II: stage 3 without plus disease

Conditions not described in Type-1 or Type-2 ROP where the appearance of the retinal vasculature can be diagnosed as pre-plus or plus are also typically diagnosed as moderate ROP. Infants diagnosed with Type-2 ROP are monitored very closely, with follow-up examinations occurring every one to two weeks. Mild ROP encompasses all other ROP-related retinal findings.

Treatment of Retinopathy of Prematurity

The presence of plus disease — abnormal dilation and tortuosity of the posterior retinal blood vessels in two or more quadrants of the retina — in Zones I or II suggests that treatment, rather than observation, is appropriate.[16] There is only a single situation where treatment is warranted in the absence of plus disease (Zone I Stage 3 ROP without plus disease). The most effective treatments for Type-1 ROP, prior to retinal detachments, are laser therapy, cryotherapy, and anti-VEGF therapy.[1,2,16] Both laser therapy and cryotherapy operate by destroying the peripheral retinal vasculature, which slows or completely eliminates the abnormal growth of retinal blood vessels. However, the consequences of this treatment are that patients' peripheral vision will inevitably be irreparably destroyed. Anti-VEGF therapy is a fairly new treatment option. In practice, anti-VEGF compounds are injected directly into the vitreous of the eye, where it acts upon the VEGF driving the abnormal retinal vessel growth. While there do not appear to be any immediate consequences of this treatment, future complications and consequences of the injection of an anti-growth factor into developing infant eyes have yet to be evaluated fully.

In Stages 4 and 5 ROP, a scleral buckle may be implanted.[1] This involves placing a tight silicone band around the eye. This method prevents the vitreous humor from pulling on the scar tissue and allows the retina to flatten and reattach to the choroid. For Stage 5 ROP only, a vitrectomy may be performed. This involves removing the vitreous humor, cutting away retinal scar tissue, reattaching the retina to the choroid, and then refilling the vitreous cavity with a saline solution.

Challenges in Retinopathy of Prematurity

While diagnosis of TR-ROP, and treatment thereof, appears straightforward, there are many underlying factors which have caused this disease to remain one of the world's leading causes of

childhood blindness. It should be noted that this is not just a disease found in developing countries. For instance, in the United States, approximately 30,000 children are born prematurely each year, and roughly half of them develop some form of ROP. As advances in neonatal care increase, allowing for the preservation of younger, smaller infants, so does the incidence of ROP.

Screening

To ensure that TR-ROP can be quickly and accurately diagnosed and treated, ROP screenings are performed frequently to ensure high sensitivity. Consequently, although the criteria that dictate which infants require screening, namely birthweight and gestational age, are sensitive, they are not very specific. Around 80% of those screened develop, at worst, mild ROP, which is self-regressing and does not affect long-term visual acuity.[2] This significantly increases the screening burden and the physiological stress placed on these already-fragile premature infants.[4,5] Additionally, these exams begin as early as 31 weeks PMA, and are carried out once every one to two weeks, until ROP is "unequivocally regressing."[16,17] However, in rural areas and developing countries, access to ROP experts is limited, and screenings are often not performed as often as recommended, if at all.[4,5] To combat this issue, telemedicine pipelines for ROP have been implemented.[7–9] However, examinations via telemedicine must be performed at least once per week. Thus, while telemedicine greatly expands the geographic area physicians can cover and increases the total number of infants screened, it increases the screening burden and frequency of exams, and fails to reduce the physiological stress placed on premature infants.

In an attempt to reduce the screening burden, various ROP risk models have been developed.[10,13,18,19] Many, in theory, have clinical relevance; in practice, they are not quite as practical. For instance, some infant risk factors, such as intraventricular hemorrhages or bronchopulmonary dysplasia, are associated with infants who develop TR-ROP. However, these

comorbidities are often rare, thereby rendering them difficult to associate definitively, or difficult to measure. Another risk model has attempted to use gestational age, birthweight, and weekly weight gain for TR-ROP.[10] While this model was able to achieve 100% sensitivity and 53% specificity on a held-out test dataset acquired from a single hospital, when a larger North American cohort, obtained from 30 hospitals, was retrospectively evaluated, specificity dropped to just 6.8% when 100% sensitivity was desired/[10,11] Unfortunately, this is not practical. While some may argue that reducing sensitivity slightly to increase specificity significantly, they fail to take into account the associated costs of a missed diagnosis — that is, a missed TR-ROP prediction will lead to life-long visual impairment. In the opinion of most physicians, this is simply not acceptable. Risk models for ROP, specifically TR-ROP, remain an open area of research.

In an effort to reduce physician burden, automated methods for the diagnosis of ROP have been implemented.[6,20] These models typically use retinal fundus images as input to a convolutional neural network that can scan for features of ROP. While these models have shown excellent performance, even when compared to ROP experts, they still require weekly ROP exams. Therefore, although they reduce the screening burden placed on ROP experts, babies must still be sedated, weekly, to undergo exams.

Finally, accurate and reliable image-based diagnosis, whether via telemedicine or automated methods, can suffer from poor image quality.[7,9] This is particularly challenging in ROP, as vitreous haze can be a feature of the disease, which can occlude portions of an image, but not necessarily render them undiagnosable.[1,9] Similarly, an image can appear free of haze or other obstructions, but an incorrect field-of-view can reduce the accuracy and reliability of an ROP diagnosis.[9] Unfortunately, this means that common image quality metrics, such as the peak signal to noise

ratio and the structural similarity index measure, do not work well for this specific application, and thus it remains an open area of research.[7,9,21]

Diagnosis

In-person and telemedical exams operate under two weak assumptions: (A) that there are enough ROP experts to screen all premature infants, and (B) that all experts agree on the clinical findings of ROP — specifically, on the diagnosis of plus disease. As previously mentioned, the diagnosis of plus disease was originally a binary decision.[3] Experts later recognized that there is a spectrum of venous dilation and arterial tortuosity, but rather than creating a continuous scale, added a third potential diagnosis, pre-plus disease.[1] Even with this addition, experts still have trouble agreeing on the level of vascular severity required to diagnose normal versus pre-plus versus plus disease vasculature.[22,23]

Campbell et al. demonstrated that agreement of plus disease diagnosis between ROP experts was imperfect, however images that were graded as normal, pre-plus, or plus disease, by a group of experts, tended to group together (**Figure 4**). Kalpathy-Cramer et al. further investigated this finding and showed that although the cut points between plus disease categories differed, the relative ranking of plus disease severity between experts was highly correlated (mean correlation coefficient, 0.97; range, 0.95–0.98; **Figure 5**). Ultimately, these results, together, suggest that (A) experts have their own cut-offs between transitions for plus disease classification, but (B) that experts rank vascular severity similarly. Put simply, although two experts may not agree on the specific diagnosis for a retinal fundus image (e.g., pre-plus versus plus), they can almost always agree that one image displays worse vascular severity than another. This could have potential applications in reducing ROP experts' screening burden by allowing pediatric ophthalmologists to perform ROP screenings, especially in developing countries, by comparing the severity of their

patients' retinal vasculature to reference standard images of plus disease. However, as mentioned, the current reference standard image of plus disease is outdated and extremely severe. Generation of personalized reference standard images, for this purpose, should be an active area of research.

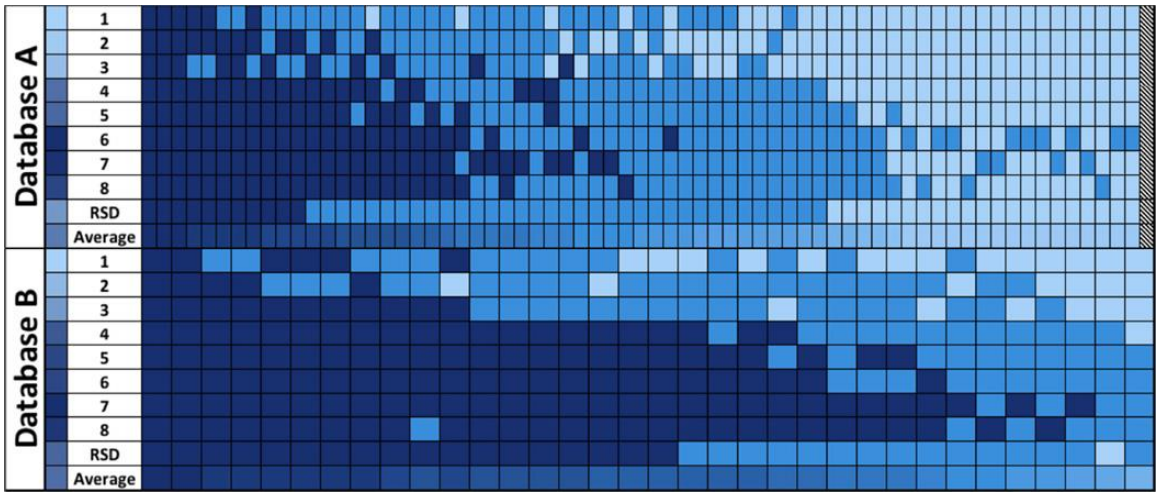

Figure 4: A representative range of images within each category of disease (plus, pre-plus, normal), and the range of expert diagnostic classifications for each image. This graphically depicts the continuous spectrum of severity of vascular abnormality within each discrete plus disease diagnosis (plus, pre-plus, or normal), from most severe (left) to least (right). In addition to demonstrating the spectrum of vascular abnormality within each ordinal classification, this shows that different experts appear to have different cut-offs for the transitions between diagnostic classifications. Figure adapted from Campbell et al., 2016.[22]

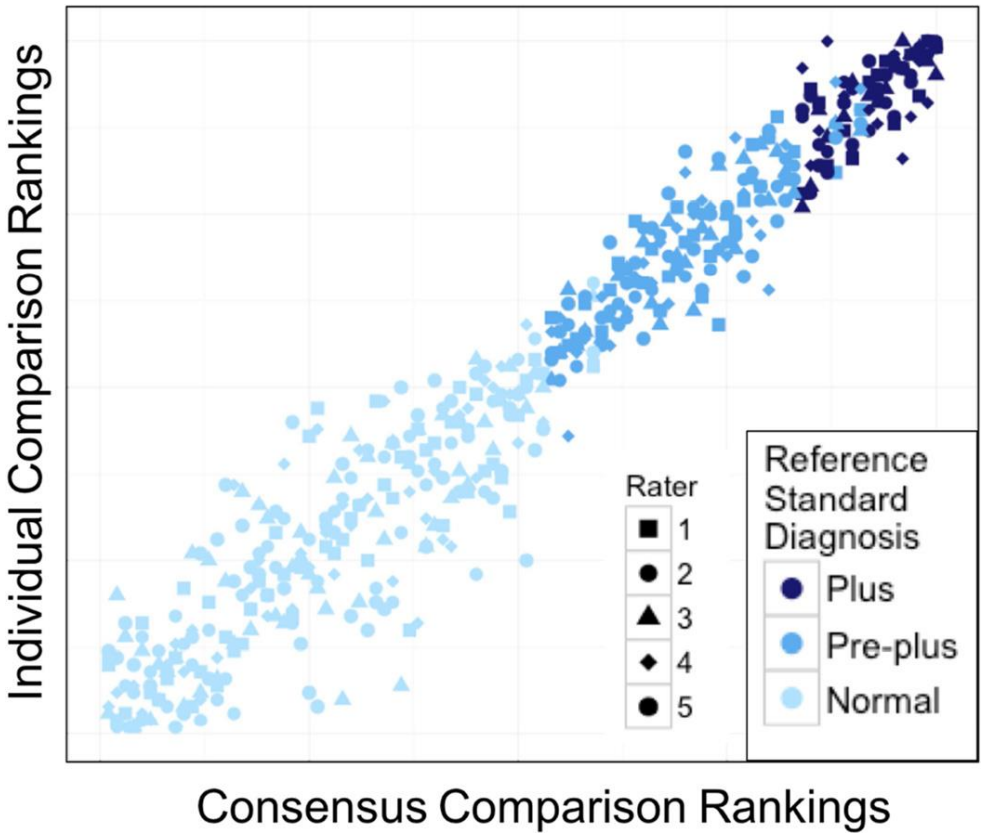

Figure 5: Scatterplot showing the comparison rankings of five expert graders. The relative vascular severity rankings between all five graders were highly consistent (mean correlation coefficient, 0.97; range, 0.95–0.98) between each expert and a consensus comparison ranking. Figure adapted from Kalpathy-Cramer et al., 2016.[23]

In conclusion, although ROP is highly diagnosable and treatable, it remains one of the world's leading causes of childhood blindness. The main challenges are: (1) a high screening burden, (2) a lack of quality assurance metrics for retinal fundus images that are used in telemedicine and automated diagnostic tools used to reduce said screening burden, and (3) a disagreement on the vascular severity required to diagnose plus disease and, subsequently, TR-ROP, which can lead to over- or under-treatment.

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Over the past decade, artificial intelligence (AI) and machine learning (ML) have become popular subjects both within and outside of the scientific community; an abundance of articles in technology and non-technology-based journals have covered the topics of machine learning, deep learning (DL), and AI.[6,24–28] Although these terms are highly associated, they are not interchangeable, and are discussed further herein.

In 1956, a group of computer scientists proposed that computers could be programmed to think and reason, "that every aspect of learning or any other feature of intelligence [could], in principle, be so precisely described that a machine [could] be made to simulate it."[29] They described this principle as "artificial intelligence." Simply put, AI is a field focused on automating intellectual tasks normally performed by humans, and ML and DL are specific methods of achieving this goal. That is, they are within the realm of AI (**Figure 1**). However, AI includes approaches that do not involve any form of "learning." For instance, the subfield known as symbolic AI focuses on hardcoding (i.e., explicitly writing) rules for every possible scenario in a particular domain of interest. These rules, written by humans, come from a priori knowledge of the particular subject and task to be completed. For example, if one were to program an algorithm to modulate room temperature of an office, he or she likely already knows what temperatures are comfortable for humans to work in and would program the room to cool if temperatures rise above a specific threshold and heat if they drop below a lower threshold. Although symbolic AI is proficient at solving clearly defined logical problems, it often fails for tasks that require higher-level pattern recognition, such as speech recognition or image classification. These more complicated tasks are where ML and DL methods perform well.

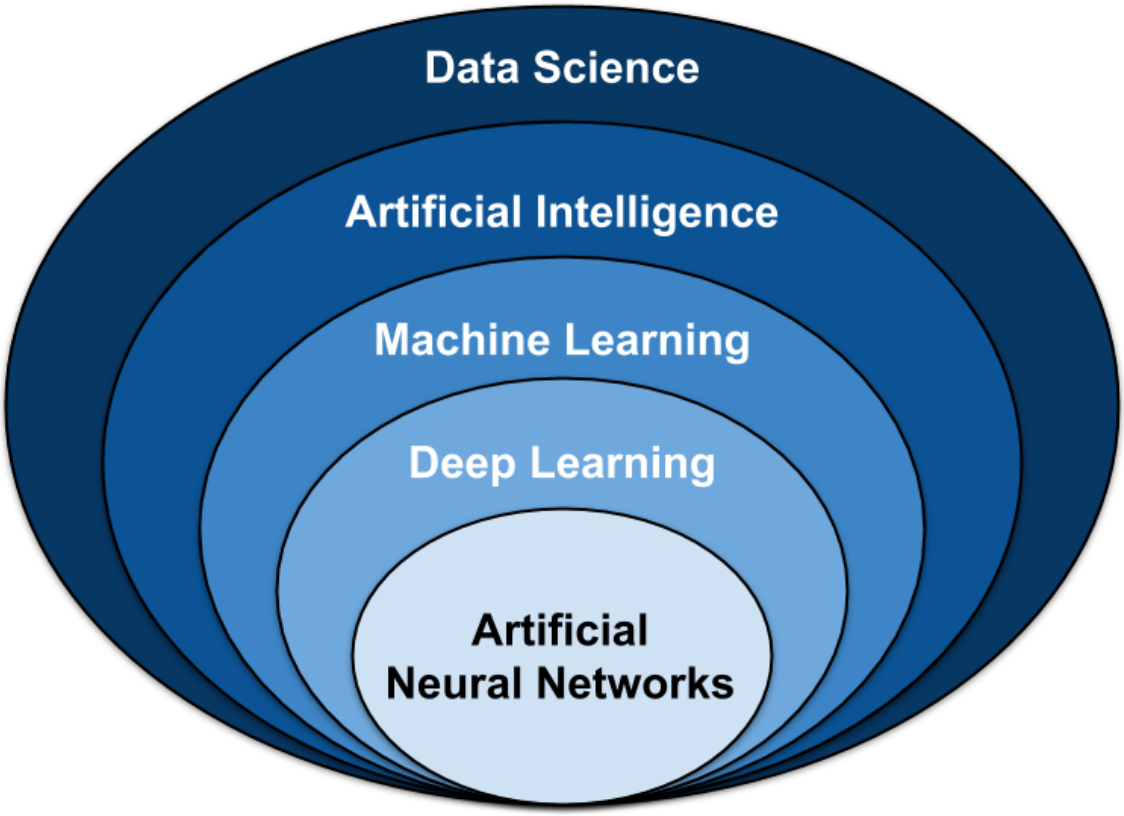

Figure 1: Umbrella of select data science techniques. Artificial intelligence (AI) falls within the realm of data science and includes classical programming and machine learning (ML). ML contains many models and methods, including deep learning (DL) and artificial neural networks (ANN).

ML is a field that focuses on the learning aspect of AI by developing algorithms that best represent a set of data. In contrast to classical programming (**Figure 2A**), in which an algorithm can be explicitly coded using known features, ML uses subsets of data to generate an algorithm that may use novel or different combinations of features and weights that can be derived from first principles **(Figure 2B**).[29–31] In ML, there are four commonly used learning methods, each useful for solving different tasks: supervised, unsupervised, semi-supervised, and reinforcement learning.[30–32] To

better understand these methods, they will be defined via an example of a hypothetical real estate company that specializes in predicting housing prices and features associated with those houses.

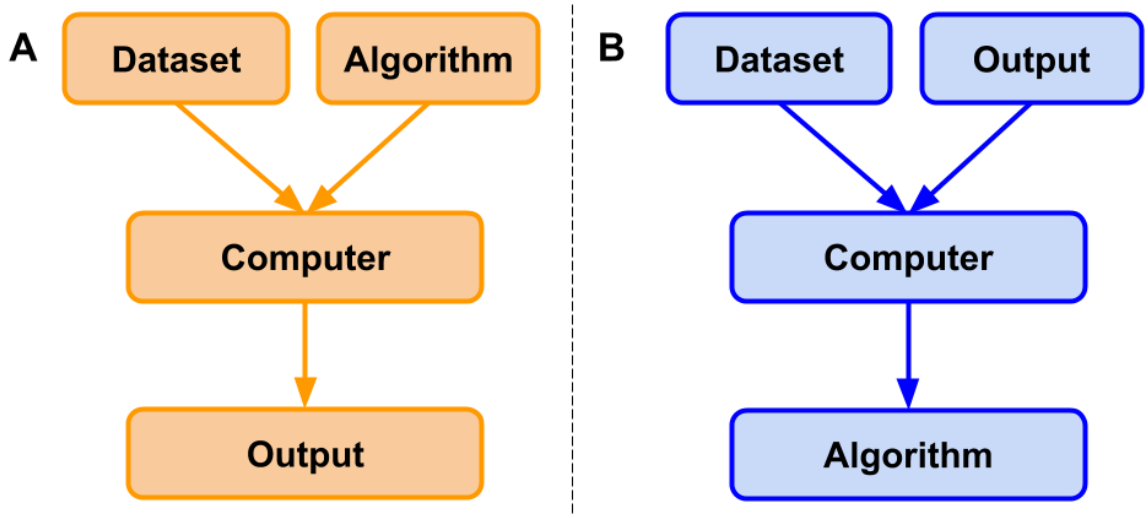

Figure 2: Classical programming versus machine learning paradigm. (A) In classical programming, a computer is supplied with a dataset and an algorithm. The algorithm informs the computer how to operate upon the dataset to create outputs. (B) In machine learning, a computer is supplied with a dataset and associated outputs. The computer learns and generates an algorithm that describes the relationship between the two. This algorithm can be used for inference on future datasets.

Supervised Learning

Suppose the real estate company would like to predict the price of a house based on specific features of the house. To begin, the company would first gather a dataset that contains many instances.[30,31] Each instance represents a singular observation of a house and associated features. Features are the recorded properties of a house that *might* be useful for predicting prices (e.g., total square-footage, number of floors, the presence of a yard) The target is the feature to be predicted, in this case the housing price.[30,31,33] Datasets are generally split into training, validation, and testing

datasets (models will always perform optimally on the data they are trained on).[30,31] Supervised learning uses patterns in the training dataset to map features to the target so that an algorithm can make housing price predictions on future datasets. This approach is supervised because the model infers an algorithm from feature-target pairs and is informed, by the target, whether it has predicted correctly.[30–32] That is, features, *x*, are mapped to the target, *Y*, by learning the mapping function, *f*, so that future housing prices may be approximated using the algorithm $Y = f(x)$. The performance of the algorithm is evaluated on the test dataset, data that the algorithm has never seen before.[30,31] The basic steps of supervised machine learning are (1) acquire a dataset and split it into separate training, validation, and test datasets; (2) use the training and validation datasets to inform a model of the relationship between features and target; and (3) evaluate the model via the test dataset to determine how well it predicts housing prices for unseen instances. In each iteration, the performance of the algorithm on the training data is compared with the performance on the validation dataset. In this way, the algorithm is tuned by the validation set. Insofar as the validation set may differ from the test set, the performance of the algorithm may or may not generalize. This concept will be discussed further in the section on performance evaluation.

The most common supervised learning tasks are regression and classification.[30–32] Regression involves predicting numeric data, such as test scores, laboratory values, or prices of an item, much like the housing price example.[30–32] Classification, on the other hand, entails predicting to which category an example belongs. Sticking with the previous example, imagine that rather than predicting exact housing prices in a fluctuating market, the real estate company would now like to predict a range of prices for which a house will likely sell, such as [0, 125K), [125K, 250K), [250K, 375K), and [375K, ∞). To accomplish this, data scientists would transform the numeric target variable into a categorical variable by binning housing prices into separate classes. These classes

would be ordinal, meaning that there is a natural order associated with the categories.[31] However, if their task was to determine whether houses had wood, plastic, or metal siding, classes would be nominal; they are independent of one another and have no natural order.

Unsupervised Learning

In contrast to supervised learning, unsupervised learning aims to detect patterns in a dataset and categorize individual instances in the dataset to said categories.[30–32] These algorithms are unsupervised because the patterns that may or may not exist in a dataset are not informed by a target and are left to be determined by the algorithm. Some of the most common unsupervised learning tasks are clustering, association, and anomaly detection.[30–32] Clustering, as the name suggests, groups instances in a dataset into separate clusters based upon specific combinations of their features.[30–32] Say the real estate company now uses a clustering algorithm on its dataset and it finds three distinct clusters. Upon further investigation, it might find that the clusters represent the three separate architects responsible for designing the homes in their dataset, which is a feature that was *not* present in the training dataset.

Semi-Supervised Learning

Semi-supervised learning can be thought of as the "happy medium" between supervised and unsupervised learning and is particularly useful for datasets that contain both labeled and unlabeled data (i.e., all features are present, but not all features have associated targets)[32] This situation typically arises when labeling images become time-intensive or cost-prohibitive. Semi-supervised learning is often used for medical images, where a physician might label a small subset of images and use them to train a model. This model is then used to classify the rest of the unlabeled images in the dataset. The resultant labeled dataset is then used to train a working model that should, in theory, outperform unsupervised models.

Reinforcement Learning

Finally, reinforcement learning is the technique of training an algorithm for a specific task where no single answer is correct, but an overall outcome is desired.[31,32] It is arguably the closest attempt at modeling the human learning experience because it also learns from trial and error rather than data alone. Although reinforcement learning is a powerful technique, its applications in medicine are currently limited and thus will be presented with a new example. Imagine one would like to train an algorithm to play the video game Super Mario Bros, where the purpose of the game is to move the character Mario from the left side of the screen to the right side in order to reach the flagpole at the end of each level while avoiding hazards such as enemies and pits. There is no correct sequence of controller inputs; there are sequences that lead to a win and those that do not. In reinforcement learning, an algorithm would be allowed to "play" on its own. It would attempt many different controller inputs and when it finally moves Mario forward (without receiving damage), the algorithm is "rewarded" (i.e., the behavior is reinforced). Through this process, the algorithm begins to learn what behavior is desired (e.g., moving forward is better than moving backward, jumping over enemies is better than running into them). Eventually, the algorithm learns how to move from start to finish. Although reinforcement has its place in the field of computer science and machine learning, it has yet to make a substantial impact in clinical medicine.

Performance Evaluation

To maximize the chance of generalizability to the performance of the algorithm on unseen data, the training dataset is usually split into a slightly smaller training dataset and a separate validation dataset.[30,31] Metrics used for evaluation of a model depend upon the model itself and whether it is in the training or testing phase. The validation dataset is meant to mimic the test dataset and helps data scientists tune an algorithm by identifying when a model may generalize well and work in a

new population. Because the validation dataset is a small sample of the true (larger) population, it may not accurately represent the population itself due to an unknown sampling bias. Therefore, model performance and generalizability should not be assessed via validation set performance. It is conceivable that a data scientist could create a validation dataset with an unknown bias and use it to tune a model. Although the model might perform well on the validation dataset, it would likely not perform well on the much larger test dataset (i.e., it would not be a generalizable model).

Typically, model performance is monitored via some form of accuracy on the training and validation datasets during this phase. So long as the accuracy of the model on the training set (*X*%) and validation set (*Y*%) are increasing and converging after each training iteration, the model is considered to be "learning." If both converge, but do not increase (e.g., *X* converges on *Y* at 50%), the model is not learning and may be underfit to the data, that is, it may not have learned enough of the relationship between features and targets in a way that it would be expected to work in another population. Finally, if training performance increases far more than validation set performance (e.g., the model has an accuracy of 99% on the data it was trained on, but only 80% on the validation data), the model is overfit. That is, it has learned features specific to the training dataset population at the expense of generalizability to another population. Although the validation dataset is not specifically used to train the algorithm, it is used to iteratively tune the algorithm. Therefore, the validation dataset is not necessarily a reliable indicator of model performance on unseen data.[30,31]

Upon completion of the training phase, a data scientist has, ideally, trained a highly generalizable model; however, this must be confirmed via a separate test dataset. In the case of supervised learning, which will be the focus of this review from here on, the performance of a learned model can be evaluated in a number of ways, but is most commonly evaluated based on prediction

accuracy (classification) or error and residuals (regression).[30,31] As previously mentioned, the test dataset contains instances of the original dataset that have not been seen by the algorithm during the training phase. If the predictive power of a model is strong on the training dataset, but poor on the test dataset, then the model is too specific to the patterns from the training data and is considered to be overfit to the training dataset.[30,31] That is, it has memorized patterns rather than learned a generalizable model. An underfit model, on the other hand, is one that performs poorly on both training and test datasets and has neither learned nor memorized the training dataset and still is not generalizable, An ideally fitted model is one that performs strongly on both datasets, suggesting it is generalizable (i.e., it will perform well on other similar datasets).

With regression models, the average mean squared error (MSE) can be an indicator of model performance.[30,31] MSE measures how close a predicted value is to the intended target value. MSE is calculated by summing the differences between predicted values and target values, squaring the results, and dividing by the total number of instances. There are many other measures of performance for regression models that are out of the scope of this review.

For binary classification, the output of the model is a class. However, before the class designation, the probability of an instance belonging to class A or class B is determined.[30,31] Normally, this probability threshold is set at 0.5. A receiver operating characteristic curve evaluates a model's true positive rate (TPR; i.e., sensitivity, recall), the number of samples correctly identified as positive divided by the total number of positive samples, versus its false-positive rate (FPR; i.e., 1 - specificity), the number of samples incorrectly identified as positive divided by the total number of negative samples (**Figure 3**, **Figure 4A**). Similarly, the precision-recall curve evaluates a model's positive predictive value (PPV; i.e., precision), the number of samples correctly identified as positive divided by the total number of samples identified as positive, versus its recall **Figure**

3, **Figure 4B**). Each curve is evaluated across the range of model probability thresholds from 1 to 0, left to right. A receiver operating characteristic curve starts at the point (FPR = 0, TPR = 0), which corresponds to a decision threshold of 1 (every sample is classified as negative, and thus there are no false or true positives). It ends at the point (FPR = 1, TPR = 1), which corresponds to a decision threshold of 0 (where every sample is classified as positive, and thus all points are either truly or falsely labeled positive). The points in between, which create the curve, are obtained by calculating the TPR and FPR for different decision thresholds between 1 and 0, trading off sensitivity (minimizing false negatives) with specificity (minimizing false positives). The area under the curve (AUC) of the receiver operating characteristics curve (AUROC) can be calculated and used as a metric for evaluating the overall performance of a classifier, assuming the classes of the dataset are balanced. If classes are not balanced, the area under the precision-recall curve (AUPR) may be a better metric of model performance because the threshold (set at 0.5 in Fig. 4B) may be adjusted. For example, if a dataset comprised 75% of class A and 25% of class B, the ratio between the two would be computed as the threshold (0.75). In practice, an AUROC value of 0.50 indicates a model that performs no better than chance, and an AUC of 1.00 indicates that the model performs perfectly; the higher the value of the AUC, the stronger the performance of the ML model. Similarly, an AUPR value at the preset threshold indicates a model that performs no better than chance, and an AUPR value of 1.00 indicates a perfect model.

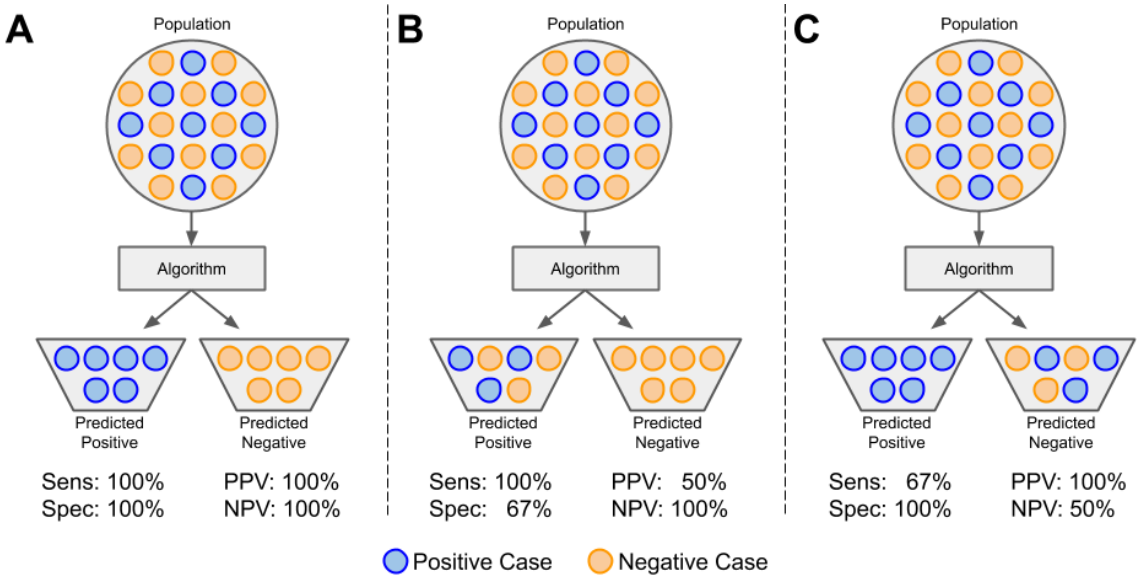

Figure 3: Sensitivity, specificity, positive predictive value, and negative predictive value. A population (dataset) is represented as circles colored blue if positive or orange if negative. The dataset is input to an algorithm that predicts each instance's class association. If an instance is correctly predicted as positive or negative, it is a true positive (TP) or true negative (TN), respectively. If an instance is incorrectly labeled positive or negative, it is a false positive (FP) or false negative (FN), respectively. (A) A model with perfect sensitivity ($\frac{TP}{TP + FN}$) and specificity ($\frac{TN}{TN + FP}$). (B) A model with perfect sensitivity (ability to correctly classify all positive cases), but poor specificity (ability to correctly classify all negative cases) and (C) a model with perfect specificity, but poor sensitivity. Although a model might have perfect sensitivity (B), it can have many false positives. Similarly, a model with perfect specificity (C) might have many false negatives. Therefore, it is also useful to evaluate the positive predictive value (PPV; $\frac{TP}{TP + FP}$) and the negative predictive value (NPV; $\frac{TN}{TN + FN}$). PPV and NPV are also thus dependent on the prevalence of disease in a population.

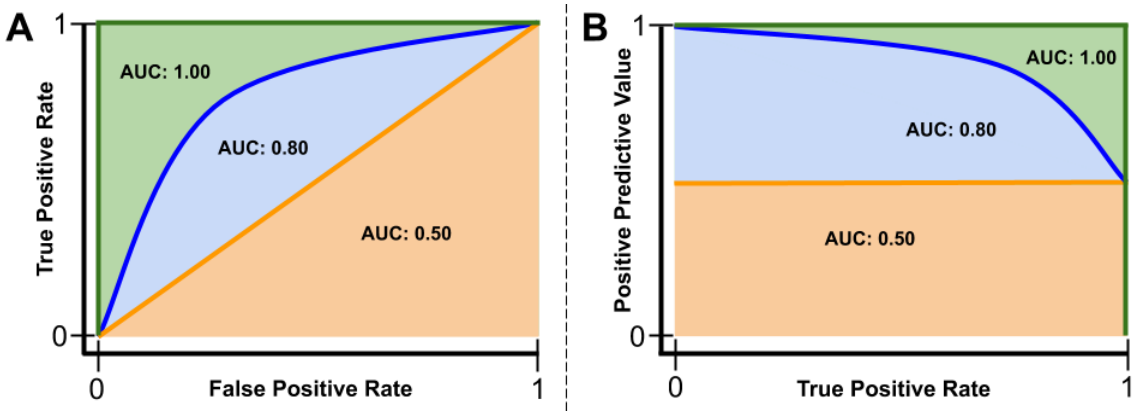

Figure 4: Example receiver operating characteristics and precision-recall curves. *Red line*: a model that performs no better than chance has an area under the curve (AUC) of the receiver operating characteristics curve (AUROC) of 0.50 or area under the precision-recall curve (AUPR) at the class ratio (*orange shaded area*). *Blue line*: a model that performs better than chance, but not perfectly, will have an AUC between 0.50 and 1.00 (*blue + orange shaded areas*). *Green line*: a model that performs perfectly has an AUC of 1.00 (*orange + blue + green shaded areas*).

Machine Learning Methods

There are many machine learning algorithms used in medicine. Described next are some of the most popular to date.

Linear Regression

Linear regression is arguably the simplest ML algorithm. The main idea behind regression analysis is to specify a relationship between one or more numeric features and a single numeric target.[30,31] Linear regression is an analysis technique used to solve a regression problem by using a straight line to describe a dataset. Univariate linear regression, a regression problem where only a single feature is used for predicting a target value, can be represented in a slope-intercept form: $y = ax + b$. Here, a is a weight describing the slope, which describes how much a line increases on the y-

axis for each increase in *x*. The intercept, *b*, describes the point where the line intercepts the y-axis. Linear regression models a dataset using this slope-intercept form, where the machine's task is to identify values of *a* and *b* such that the determined line is best able to relate the supplied values of *x* values to the values of *y*. Multivariate linear regression is similar; however, there are multiple weights in the algorithm, each describing to what degree each feature influences the target.

In practice, there is rarely a single function that fits a dataset perfectly. To measure the error associated with a fit, the residuals are measured. Conceptually, residuals are the vertical distances between predicted values, , and actual values, *y*. In machine learning, the cost function is a calculus derived term that aims to minimize errors associated with a model.[30,31,34] The process of minimizing the cost function involves an iterative optimization algorithm known as gradient descent, of which the mathematical calculations involved are outside the scope of this article.[30,31,35] In linear regression, the cost function is the previously described MSE. Minimizing this function often obtains estimates of *a* and *b* that best model a dataset. All model-based learning algorithms have a cost function, and the goal is to minimize this function to find the best-fit model.[30,31]

Logistic Regression

Logistic regression is a classification algorithm where the goal is to find a relationship between features and the probability of a particular outcome. Rather than using the straight line produced by linear regression to estimate class probability, logistic regression uses a sigmoidal curve to estimate class probability (**Figure 5**). This curve is determined by the sigmoid function, which produces an S-shaped curve that converts discrete or continuous numeric features (*x*) into a single numerical value (*y*) between 0 and 1.[30,31] The major advantage of this method is that probabilities are bounded between 0 and 1 (i.e., probabilities cannot be negative or greater than 1). It can be

either binomial, where there are only two possible outcomes, or multinomial, where there can be three or more possible outcomes.

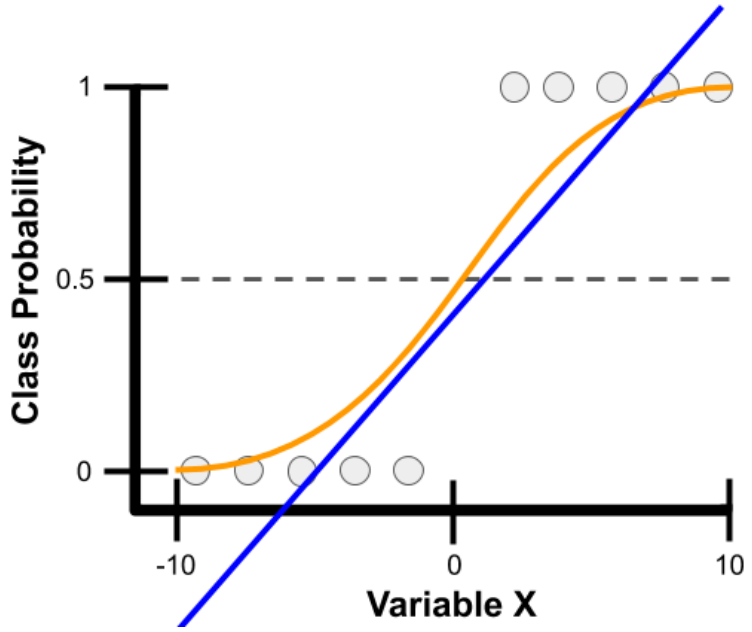

Figure 5: Example class probability prediction using linear and logistic regression. Presented are linear (blue line) and logistic (red line) regression models for predicting the probability of various samples (gray circles) as belonging to a particular class using a single variable, variable X, which ranges from -10 to 10. With logistic regression, variable X is transformed into class probabilities that are bounded between 0 and 1 using the sigmoid function. Simple linear regression attempts to estimate class probabilities, but is not bounded between 0 and 1; thus, it breaks a fundamental law of probability that does not allow for negative probabilities or those greater than 1.

Decision Trees and Random Forests

A decision tree is a supervised learning technique, primarily used for classification tasks, but can also be used for regression.[30,31] A decision tree begins with a root node, the first decision point for

splitting the dataset, and contains a single feature that best splits the data into their respective classes (**Figure 6**). Each split has an edge that connects either to a new decision node that contains another feature to further split the data into homogenous groups or to a terminal node that predicts the class. This process of separating data into two binary partitions is known as recursive partitioning. A random forest is an extension of this method, known as an ensemble method, that produces multiple decision trees. Rather than using every feature to create every decision tree in a random forest, a subsample of features are used to create each decision tree. Trees then predict a class outcome, and the majority vote among trees is used as the model's final class prediction.

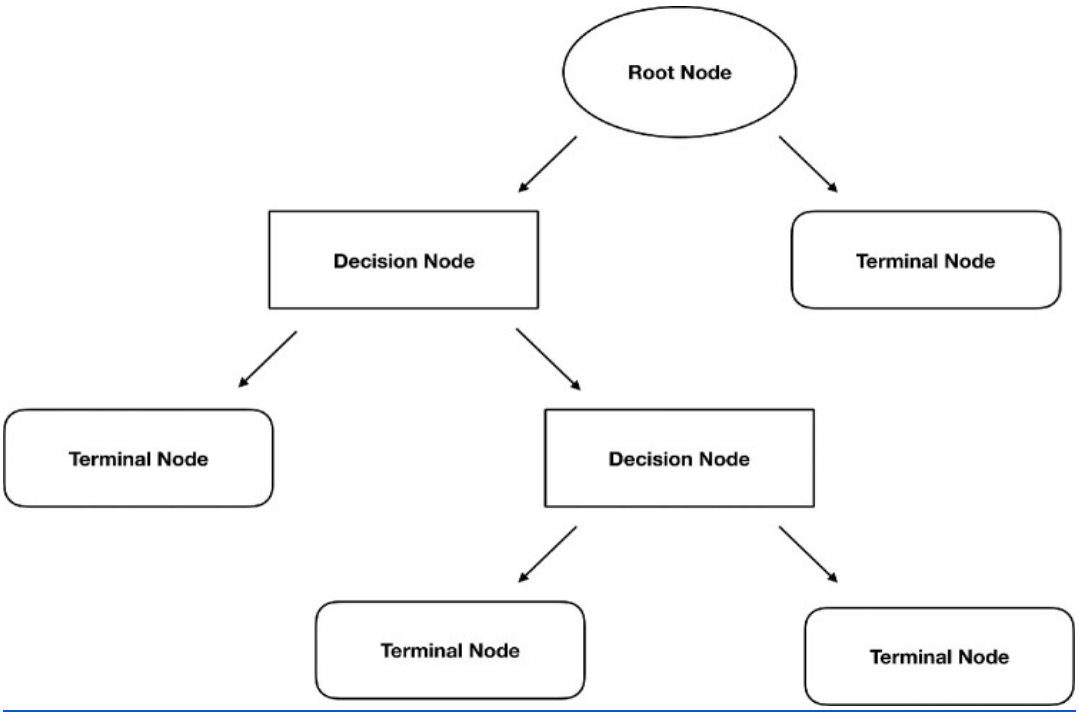

Figure 6: Structure of a decision tree. Splitting of the dataset begins at the root node. Each split connects to either another decision node, which results in further splitting of the data, or a terminal node that predicts the class of the data.

Feedforward Neural Networks

An artificial neural network (ANN) is a machine learning algorithm inspired by biological neural networks.[30,31,36] Each ANN contains nodes (analogous to cell bodies) that communicate with other nodes via connections (analogous to axons and dendrites). Much in the way synapses between neurons are strengthened when their neurons have correlated outputs in a biological neural network (the Hebbian theory postulates that "nerves that fire together, wire together"), connections between nodes in an ANN are weighted based upon their ability to provide a desired outcome.

A perceptron is a machine learning algorithm that takes in a series of features and their targets as input and attempts to find a line, plane, or hyperplane that separates the classes in a two-, three-, or hyper-dimensional space, respectively.[31,37,38] These features are transformed using the sigmoid function (**Figure 7A**). Thus, this method is similar to logistic regression; however, it only provides class associations, and not the probability of an instance belonging to a class.

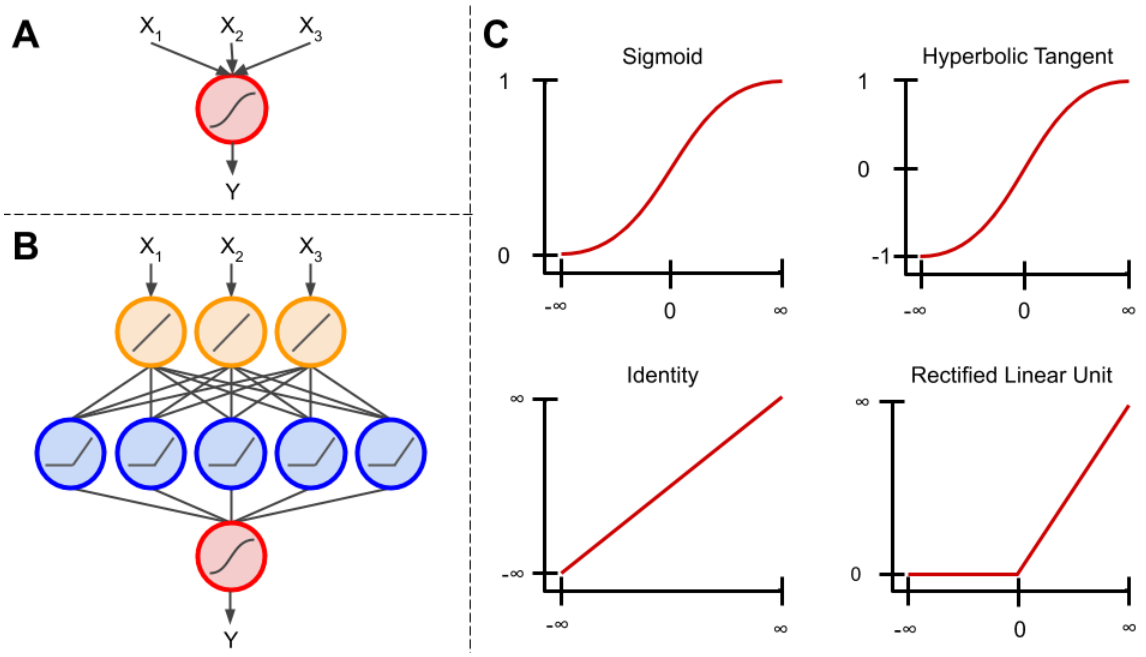

Figure 7: Components of a neural network. (A) The basis of an artificial neural network, the perceptron. This algorithm uses the sigmoid function to scale and transform multiple inputs into a single output ranging from 0 to 1. (B) An artificial neural network connects multiple perceptron units, so that the output of one unit is used as input to another. Additionally, these units are not limited to using the sigmoid activation function. (C) Examples of four different activation functions: sigmoid, hyperbolic tangent, identity, and rectified linear unit. The sigmoid scales inputs between 0 and 1 using an S-shaped curved. Similarly, the hyperbolic tangent function uses an S-shaped curve, but scales inputs between -1 and 1. The identity function can multiply its input by any number to produce a linear output. The rectified linear unit is similar to the identity function, however all inputs < 0 are given an output value of 0. There are other activation functions outside of these, but these are arguably.

When multiple perceptrons are connected, the model is referred to as a multilayer perceptron algorithm or an ANN. Commonly, ANNs contain a layer of input nodes, a layer of output nodes,

and a number of "hidden layers" between the two.[31,39] In simple ANNs, there exists an input layer between zero and three hidden layers and an output layer, whereas deep neural networks contain tens or even hundreds of hidden layers.[31] For most tasks, ANNs feed information forward. This is known as a feedforward neural network, meaning information from each node in the previous layer is passed to each node in the next layer, transformed, and passed forward to each node in the next layer (**Figure 7B**). In recurrent neural networks, which are out of the scope of this paper, information can be passed between nodes within a layer or to previous layers, where their output is operated on and fed forward once again.[31,38]

Each layer in an ANN can contain any number of nodes; however, the number of nodes in the output layer typically corresponds to the number of classes being predicted if the goal is multiclass classification, a single node with a sigmoidal activation for binary classification, or a linear activation function if the goal is regression.[31,39] These activation functions simply transform a node's input into a desired output (**Figure 7C**). Each node in an ANN contains an activation function (not just the output layer; **Figure 7B**). These activation functions, although not always linear, do not have to be complex. For instance, the rectified linear unit applies a linear transformation to inputs ≥ 0, and sets inputs < 0 to 0.[40] It follows that as inputs proceed through an ANN, they are progressively modified at each layer so that at the final layer they no longer resemble their original state. However, this final representation of the input is, in theory, more predictive of the specified outcome.

Convolutional Neural Networks

For image recognition tasks, each input into a feedforward ANN corresponds to a pixel in the image. However, this is not ideal because there are no connections between nodes in a layer. In practice, this means that the spatial context of features in the image are lost.[39,41,42] In other words,

pixels that are close to one another in an image are likely more correlated than pixels on opposite sides of the image, but a feedforward ANN does not take this into account.

A convolutional neural network (CNN) is a special case of the ANN that overcomes this issue by preserving the spatial relationship between pixels in an image.[31,39,41,42] Rather than using single pixels as input, a CNN feeds patches of an image to specific nodes in the next layer of nodes (rather than all nodes), thereby preserving the spatial context from which a feature was extracted.[39,41,42] These patches of nodes learn to extract specific features and are known as convolutional filters.

Convolutions are widely used in the realm of image processing, and are often used to blur or sharpen images, or for other tasks such as edge detection.[39,43] A visible-light digital image is simply a single matrix if the image is grayscale or three stacked matrices if the image is color (red, green, and blue color channels).[39,41,43] These matrices contain values, typically between 0 and 255, that represent pixels in the image and the intensity of each color channel at each pixel.[39,41–43] A convolutional filter is a much smaller matrix that is typically square and range in size from 2×2 to 9×9.[39,41–44] This filter is passed over the original image and, at each position, element-wise matrix multiplication is performed (**Figure 8**).[43] The output of this convolution is mapped to a new matrix (a feature map) that contains values corresponding to whether or not the convolutional filter detected a feature of interest [43].

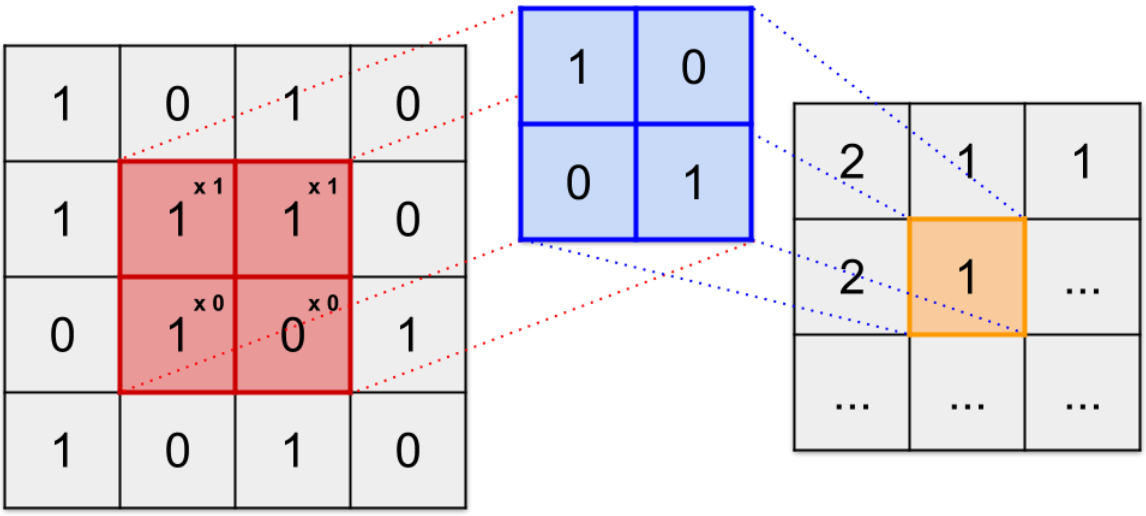

Figure 8: Example of a digital image convolved with a filter. The image (*left*) is transformed into the feature map (*right*) via a convolutional filter (*center*). The convolutional filter is designed to locate diagonal lines running from top left to bottom right of the image. The filter passes over the image in a specified manner and each element in the image (*red*) is multiplied by the corresponding element in the convolutional filter (*blue*). The summation of these elements (*orange*) is output into a new matrix that reports the presence of a diagonal line. The feature map indicates 2 when the specified diagonal line is found, 1 if a portion of it is found, and 0 if none of it is found.

In CNNs, filters are trained to extract specific features from images (e.g., vertical lines, U-shaped objects), and mark their location on the feature map.[39,41,42] A deep CNN then uses the feature map as input for the next layer, which uses new filters to create another new feature map. This can continue for many layers and, as it continues, the extracted features become abstract, but highly useful for prediction. The final features maps are then compressed from their square representations and input to a feedforward ANN, where classification of the image based on the extracted features and textures can occur.[41,42] This process is referred to as "deep learning" (DL).[39]

Aside from image classification tasks, DL has shown promise for image segmentation tasks.[6,45,46] Rather than classifying images as a whole, this method aims to identify objects within an image. To accomplish this task, DL classifies individual pixels given surrounding pixel information. For example, in diabetic retinopathy, a segmentation algorithm might segment (outline) the retinal vasculature by assigning probabilities to individual pixels as belonging to a retinal blood vessel or not belonging to a retinal blood vessel. A similar method for breast cancer detection could mark pixels as belonging to a mass or not belonging to a mass, and the output image could be provided to a radiologist for further review.

Generative Adversarial Networks

As opposed to discriminative algorithms, such as those previously discussed, generative algorithms aim to create new data instances that resemble the training data.[47] For instance, a generative model might attempt to produce images of dogs and cats, and a discriminative model could be trained to learn the difference between those images. To train a generative model, it must be pitted against a discriminative model. The generator output is connected directly to the discriminator input (**Figure 9**). The discriminator's classification of generated images provides a signal that the generator uses to update its weights and improve the images it synthesizes. When training begins, the generator produces data that is clearly not real, and the discriminator quickly learns to identify it. However, as training progresses and the generator begins to learn which features the discriminator uses to identify fake images, the generator begins producing images that appear real. The eventual goal is to train a generator that fools the discriminator into classifying generated images as real images. Because of the competing networks used to train the generator, the model is known as a generative adversarial network (GAN).

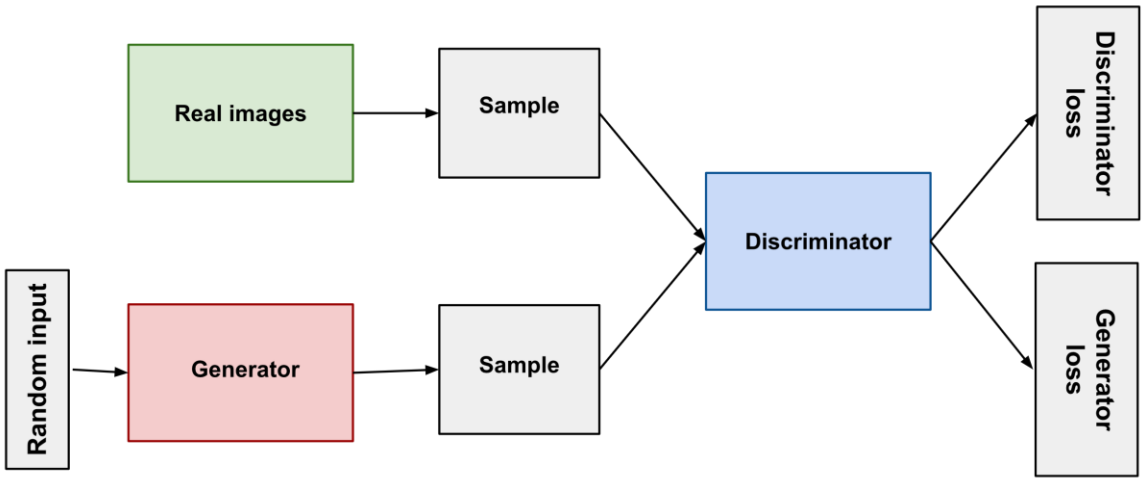

Figure 9: Overview of GAN structure.[48] A random input is provided to the generator, which attempts to create an output that is similar to the training data. The discriminator is provided with generated examples and real examples from the training data, and attempts to discern between real and fake images. The loss values associated with the discriminators outputs of generated images are recycled and used to update the weights of the generator so that it may begin to learn what it is doing correctly and what it is not.

Typically, neural networks need some form of input. A basic GAN takes random noise as its input and attempts to transform this noise into a meaningful output. This noise can get the GAN to produce a wide variety of data, sampling from different places in the target distribution. However, it is possible to provide non-random noise to guide the generated outputs. CycleGAN and pix2pix are two networks that have demonstrated this.[49,50] The combined goal of pix2pix and CycleGAN is image-to-image translation — to learn the mapping between an input image and an output image. pix2pix was proposed as a general-purpose solution to image-to-image translation problems.[49] The GAN would not only learn the mapping from input image to output image, but also learn a loss function to train said mapping. The major implication of this is that it makes it possible to apply

the same generic approach to problems that traditionally would require very different loss formulations (i.e. the same GAN structure can be applied for many different tasks). Typically, image-to-image translation is accomplished using a training set of aligned image pairs, but for many tasks paired training data is not available. CycleGAN is a GAN that learns to translate images from a source domain, X, to a target domain, Y, in the absence of paired examples.[50]

Machine Learning and Deep Learning in Ophthalmology

Although DL has become a highly popular technique in ophthalmology, there are a multitude of examples of classic ML algorithms being used in the field. Simple linear models have been used to predict patients who would develop advanced age-related macular degeneration and to discern which factors separate patients into who will respond to anti-vascular endothelial growth factor treatment versus those who will not.[34,51–53] Random forest algorithms have been used to discover features that are most predictive of progression to geographic atrophy in age-related macular degeneration and find prognostic features for visual acuity outcomes of intravitreal anti-vascular endothelial growth factor treatment.[54,55] Random forest classifiers have also been applied to diagnose and grade cataracts from ultrasound images, as well as identify patients with glaucoma based on retinal nerve fiber layer and visual field data.[56,57]

The popularity of DL has especially risen in the field of ophthalmology for image-based diagnostic systems. Gulshan et al. demonstrated that DL could classify diabetic retinopathy, in agreement with the Early Treatment for Diabetic Retinopathy Study scale, using only retinal fundus images as input and the consensus diagnoses of multiple clinicians as the “class labels.” The presence of features such as microaneurysms, intraretinal hemorrhages, or neovascularization were not supplied to the DL method as signs of diabetic retinopathy. Rather, the DL model either learned

these features or learned novel features that aid in the diagnosis of diabetic retinopathy. Further, Brown et al. trained a similar DL network for the diagnosis of plus disease in retinopathy of prematurity.[6] First, an algorithm was trained to segment retinal vasculature into binary vessel maps. Then another DL algorithm was trained to examine the vessel maps and conclude whether the vasculature appeared normal or abnormal.[6] This network, too, performs on par or better than most experts in the field. One of the most impressive examples of DL in ophthalmology was conducted by De Fauw et al. Using three-dimensional optical coherence tomography images, a DL framework was trained to not only detect a single disease, but more than 50 common retinal diseases.[28]

Challenges with Deep Learning

In recent years, DL has become a hot topic within the field of medicine given the digital availability of information; however, many challenges still exist. DL is limited by the quantity and quality of data used to train the model. It is difficult to estimate how much data are necessary to sufficiently and reliably train DL systems because it depends both on the quality of the input training data as well as the complexity of the task. Typically, thousands of training examples are required to create a model that is both accurate and generalizable. Thus, developing models for identification of rare diseases, where large datasets may not be readily available, is especially challenging. On the other hand, although one might assume that more data will always lead to better models, if the quality of the training data is imprecise, mislabeled, or somehow systematically different than the test population, training on very large datasets may result in models that do not perform well in real-world scenarios. Furthermore, there is an implicit assumption that datasets are accurately labeled by human graders. Unfortunately, this is often not the case, and noisy and/or missing labels are often a bane for data scientists.

DL methods also suffer from the "black box" problem: input is supplied to the algorithm and an output emerges, but it is not exactly clear what features were identified or how they informed the model output.[28,58,59] In contrast, simple linear algorithms, although not always as powerful as DL, are easily interpretable. The computed weights for each feature are supplied upon completion of the training process, which allow for one to interrogate exactly how the model works and possibly discover important predictors that may be useful for prevention of a disease. With deep learning, a complex series of matrix multiplication and abstract filters makes interpretability significantly more challenging.[28,58,59] Activation maps, or heatmaps, are methods that attempt to address the "black box" issue by highlighting areas of images that highlight regions of an image that "fire together" with the output classification label. Unfortunately, these methods still require human interpretation, as they are often not examined critically (examples are cherry picked for publication, highly subject to confirmation bias, etc.), and thus this remains an active area of research. For instance, if a DL model classifies a fundus image as having proliferative diabetic retinopathy, a heatmap will highlight feature areas on that fundus image that contributed to the decision of being classified as having proliferative diabetic retinopathy. It is up to the physician to interpret whether these DL model identified features are the same features the physician would use to diagnose the disease, and the implications of such findings.

AI methods have shown to be a promising tool in the field of medicine. Recent work has demonstrated that these methods can develop effective diagnostic and predictive tools to identify various diseases. In the future, AI-based programs may become an integral part of patients' clinic visits with their ability to assist in diagnosis and management of various diseases.

AIM 1: QUALITY CONTROL FOR RETINAL FUNDUS IMAGES

ABSTRACT

Accurate image-based ophthalmic diagnosis relies on fundus image clarity. This has important implications for the quality of ophthalmic diagnoses and for emerging methods such as telemedicine and computer-based image analysis. The purpose of this study was to implement a deep convolutional neural network (CNN) for automated assessment of fundus image quality in retinopathy of prematurity (ROP).

During routine ROP screenings, 6139 retinal fundus images were collected from preterm infants from nine academic institutions. Each image was graded for quality (acceptable quality [AQ], possibly acceptable quality [PAQ], or not acceptable quality [NAQ]) by three independent experts. Quality was defined as the ability to assess an image confidently for the presence of ROP. Of the 6139 images, NAQ, PAQ, and AQ images represented 5.6%, 43.6%, and 50.8% of the image set, respectively. Because of low representation of NAQ images in the data set, images labeled NAQ were grouped into the PAQ category, and a binary CNN classifier was trained using 5-fold cross-validation on 4000 images. A test set of 2109 images was held out for final model evaluation. Additionally, 30 images were ranked from worst to best quality by six experts via pairwise comparisons, and the CNN's ability to rank quality, regardless of quality classification, was assessed. CNN performance was evaluated using area under the receiver operating characteristic curve (AUC). A Spearman's rank correlation was calculated to evaluate the overall ability of the CNN to rank images from worst to best quality as compared with experts.

The mean AUC for 5-fold cross-validation was 0.958 (standard deviation, 0.005) for the diagnosis of AQ versus PAQ images. The AUC was 0.965 for the test set. The Spearman's rank correlation coefficient on the set of 30 images was 0.90 as compared with the overall expert consensus ranking.

This model accurately assessed retinal fundus image quality in a comparable manner with that of experts. This fully automated model has potential for application in clinical settings, telemedicine, and computer-based image analysis in ROP and for generalizability to other ophthalmic diseases.

INTRODUCTION

Technologies such as digital imaging, telemedicine, and artificial intelligence for image analysis are beginning to revolutionize the practice of ophthalmology.[6–8,60–66] A critical issue that plagues nearly all medical imaging applications is poor image quality.[9,21,67–81] In the best case scenario, poor image quality renders images useless for diagnosis and wastes time and resources due to required follow-up imaging sessions. In the worst case, it leads to incorrect diagnoses, resulting in either over- or under-treatment and the potential for life-altering consequences. To address this issue, we have focused on retinopathy of prematurity (ROP), a potentially-blinding childhood disease.

Advances in medical technology have also been witnessed in the neonatal intensive care unit (NICU).[14,15] The survival rate of premature infants has dramatically increased over the last few decades.[14,15] Unfortunately, this has not come without consequences. ROP, a vasoproliferative retinal disease, affects approximately two-thirds of premature infants weighing <1251 grams at birth.[16,18,82] While ROP has the potential to cause permanent blindness, it is treatable via laser photocoagulation or intravitreal injections of anti-vascular endothelial growth factor (anti-VEGF), if diagnosed promptly.[16] Treatment is initiated for the following retinal findings: Zone I ROP, stages 1, 2 or 3, plus disease present; Zone I ROP, stage 3, plus disease not present, and Zone II ROP, stages 2 or 3, plus disease present.[16] It is obvious that plus disease, defined as "abnormal dilation and tortuosity of the posterior retinal blood vessels in two or more quadrants of the retina," is a significant indicator of the need for treatment. It is therefore absolutely necessary to diagnose plus disease in an accurate and timely manner. The presence of plus disease in at least two quadrants of the retina is easier to diagnose when image quality is high (**Figure 1A**). However, as

image quality begins to deteriorate, visualization of the retina becomes difficult, if not impossible (**Figure 1B,C**).

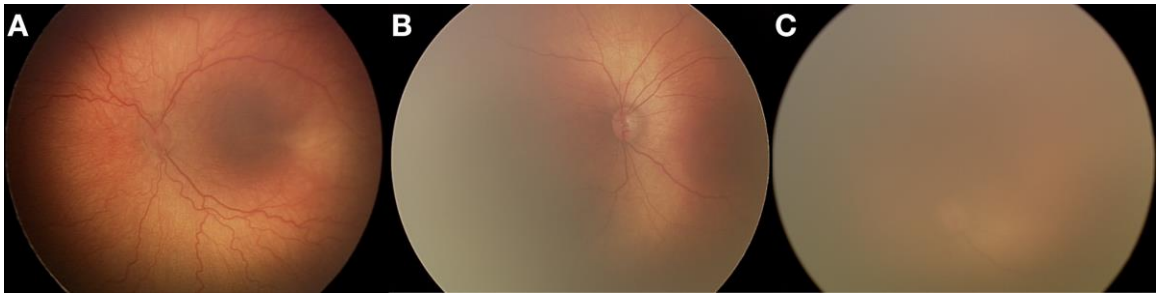

Figure 1: Varying qualities of retinal fundus images. Representative images from the (A) Acceptable Quality (AQ), (B) Possibly Acceptable Quality (PAQ), and (C) Not Acceptable Quality (NAQ) classes. Note that as image quality degrades, visualization of retinal vasculature becomes more complex, if not impossible. Because NAQ images were not highly represented in our data set (5.6%), they were grouped with the PAQ images into a single category. The final representation of AQ and PAQ images in our data set was 50.8% and 49.2%, respectively.

A major barrier to timely ROP treatment is a lack of access to ROP experts in both developed and developing countries.[7,9,18,82] Therefore, the implementation of telemedicine and other computer-based image analysis applications that make use of high-quality fundus images is crucial.[8] Recently, we have developed DeepROP, a deep learning model for automated assessment of plus disease in ROP patients.[6] When this model is provided with high-quality images, it returns highly accurate diagnoses. However, it is reasonable to assume that images of lower quality will tend to be misclassified more often than images of higher quality. Herein we describe an extension of preliminary work that attempts to address this pitfall – a deep convolutional neural network (CNN) to automatically assess the quality of retinal fundus images.[83] A CNN is an artificial neural network trained to extract features from images. A deep CNN is an extension of this model, which creates

new images using the extracted features. Essentially, a deep CNN extracts features from features from features and so on. In the early layers of the network, the extracted features are typically straight lines of various rotations. In the deeper layers of a CNN, features become more abstract. Because there are typically tens of millions of parameters to train (e.g. weights of the edges connecting nodes), we take advantage of a method known as transfer learning. Here, we implement a pretrained CNN architecture, specifically Inception-V3, which has been trained to identify everyday objects, such as cats, cars, trees, dishwashers, etc., and we fine-tune its learned filters for this specific use case.[84,85] This has numerous potential applications, such as a pre-screening method for our ROP diagnostic tool, a quality metric for imaging technicians, or a workflow component for telemedicine-based applications.

METHODS

Institutional Review Board

All data for this study were obtained through the multi-center, NIH-funded, Imaging and Informatics in ROP (i-ROP) study centered at Oregon Health & Science University (OHSU). This study was approved by the institutional review board at the coordinating center (OHSU, Portland, Oregon) and at each of 8 study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Weill Cornell Medical Center, Asociación para Evitar la Ceguera en México) and was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from parents of all infants enrolled.

Retinal Fundus Image Data Sets

Using a RetCam (Natus; Pleasanton, CA), 6,139 wide-angle fundus images were collected from preterm infants during routine ROP screening examinations. Three masked graders evaluated images for ROP stage, zone, plus disease, and image quality (based upon acceptability for diagnosis of ROP). Labels for image quality were: Acceptable Quality (AQ), Possibly Acceptable Quality (PAQ), or Not Acceptable Quality (NAQ). Graders were not told what defined AQ versus PAQ versus NAQ images. The final classification represents a majority vote of the 3 independent assessments of the suitability of an image for the task of ROP classification (zone, stage, and plus). AQ, PAQ, and NAQ images represented 50.8%, 43.6%, and 5.6% of the final data set, respectively (Figure 1). Due to low representation of NAQ images in this data set, NAQ images were combined with images from the PAQ category. The final distribution of the data set was 50.8% AQ images and 49.2% PAQ images. It should be noted, however, that the PAQ label does not necessarily

imply that an image is useless for diagnosis, but that a higher-quality image would increase the confidence of the diagnosis being made. For example, it is possible that a diagnosis could be made from an image with half of the retinal image occluded, but that an image grader's confidence might be higher if the entire image were easily visualized.

To assemble the training set, 2,000 AQ images and 2,000 PAQ images were selected at random. These 4,000 images were randomly decomposed into five separate, equally-stratified sets to be used for 5-fold cross-validation. An independent test set was formed using 2,109 randomly selected images that represented the true underlying distribution of AQ to PAQ images. The remaining 30 images were used to create a ranked data set. Briefly, the six experts ranked the smaller set of 30 images from worst quality to best quality. A web-based interface was implemented, which presented each expert with two images and the prompt "Select the higher quality image for the diagnosis of plus disease." After multiple pairwise comparisons, individual expert rankings of worst to best quality images were developed. Using an Elo rating system, all expert rankings were aggregated to form an overall expert consensus ranking of the images.

Model Architecture

This model was built and trained using Keras, a deep learning library for the programming language Python, with the TensorFlow backend (an open source software library for numerical computation using data flow graphs). The convolutional portion of the model made use of a pretrained CNN, specifically Inception-V3.[84] The weights of the CNN were initialized using the values obtained after training the CNN on the ImageNet database, a collection of over 14 million hand-annotated images containing more than 20,000 classes.[85] This reduced training time, as it

allowed the CNN to learn basic features of everyday objects by developing filters to extract specific shapes and textures prior to fine-tuning on medical images. Two fully-connected layers were built on top of the convolutional layers. The first layer consisted of 4,096 nodes using a rectified linear unit (ReLU) activation function. Because we sought to discriminate between AQ and PAQ images, the second layer consisted only of a single binary output node. This final layer made use of the sigmoid activation function; images were not only assigned a classification of AQ or PAQ, but the associated probability of belonging to said class was reported. To prevent overfitting, a dropout function with a probability of 0.5 was inserted between the two layers. Inputs to the model were RetCam images of size 640×480×3 or 1024×768×3 scaled down to 150×150×3. All training and test set image pixel values were rescaled into the [0, 1] range. Training set images also had random zoom (± 20%), horizontal flips, and vertical flips applied to them to synthetically increase the size of the training data set and reduce the chance of overfitting.

Model Training and Evaluation

The five subsets of the training set were used to perform 5-fold cross-validation. Briefly, five versions of the CNN were trained and evaluated using unique validation sets and slightly different training sets. Each CNN was evaluated on subset 1, 2, 3, 4, or 5, and trained on the remaining four subsets. This method allows for close approximation of the test error and reduces the probability of overfitting. Training occurred for 100 epochs (iterations). However, the epoch with the lowest validation set error was selected for each of the five CNNs. Training was executed using the following hyperparameters: optimizer: mini-batch gradient descent, batch size: 20, learning rate:

0.001, momentum: 0.9, loss: binary cross-entropy, and validation metric: accuracy. All layers of the model were adjustable (i.e. the convolutional layers were not frozen).

Data Analysis

Following 5-fold cross-validation, the area under the receiver operating characteristics curve (AUC) was computed for each model. The CNN with the highest AUC was selected as the final model, on which test and ranked set predictions were made. Images were input to the CNN, which calculated the probability of an image belonging to the AQ category using the softmax function of the final layer. A score less than 0.5 placed the image into the PAQ category, and a score greater than or equal to 0.5 placed the image into the AQ category. The AUC of the model was evaluated.

As mentioned above, the output of the CNN for any given image was a probability from 0 to 1. These values were used to rank the set of 30 images from worst to best quality for diagnosis of ROP. The Spearman's rank correlation test was used to assess the similarity between the CNN and the consensus ranking of the images by the six experts, as well as the correlation between individual experts.

RESULTS

Classification Performance

The AUCs resulting from 5-fold cross-validation ranged from 0.953 to 0.965, with a mean (SD) of 0.958 (0.005) (**Figure 2A**). Model 1 was selected as the final model. On the test set, the AUC was 0.965 (**Figure 2B**), in line with the estimated test set AUC predicted by 5-fold cross-validation (**Figure 2A**), and the sensitivity and specificity were 93.9% and 83.6%, respectively. Depending upon the application for which the model was implemented, the classification cutoff probability could be increased or decreased to favor sensitivity or specificity (i.e. to avoid false negatives or avoid false positives).

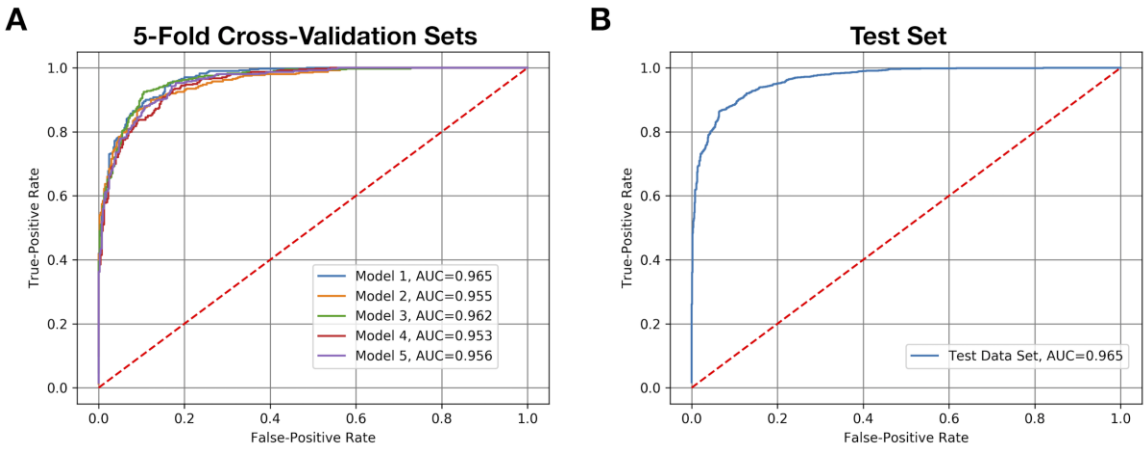

Figure 2: Areas under the receiver operating characteristics curves (AUC). (A) The AUCs for each convolutional neural network (CNN) produced by 5-fold cross-validation are shown, with mean (SD) equal to 0.958 (0.005). Model 1 demonstrated the highest level of discriminatory power between acceptable quality images and possibly acceptable quality images, as was indicated by the AUC. Therefore, it was selected for final evaluation on the independent test set (B), where it performed with an AUC equal to 0.965, a sensitivity of 93.9% and a specificity of 83.6%.

Ranked Set Performance

Figure 3 describes the Spearman's rank correlation coefficients for each individual expert grader's rank, the consensus rank, and the CNN rank. The Spearman's rank correlation test coefficients between experts ranged from 0.89 to 0.97, suggesting a very high correlation of agreement on relative image quality. Unsurprisingly, all experts were highly correlated to the consensus rank (0.94 - 0.97). The correlations between the CNN and individual experts ranged from 0.86 to 0.93, and the correlation between the CNN and the consensus ranking was 0.90, suggesting that the CNN not only has high inter-group discrimination, but high intra-group discrimination. In essence, given two images from the same class, the CNN can recognize which of the two images is of higher quality despite originating from the same class. This suggests that the model has not only learned the difference between an AQ image or a PAQ image, but that it has learned what features make any retinal fundus image of higher quality than another.

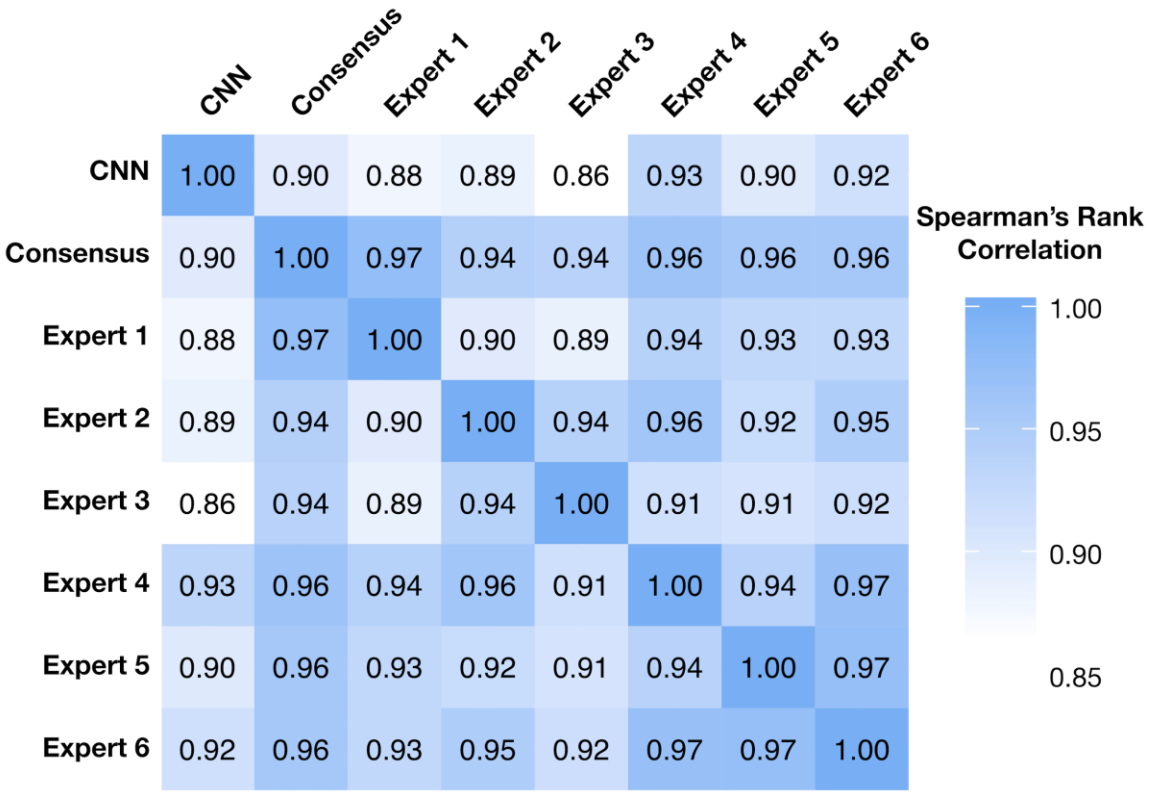

Figure 3. Correlation heatmap of expert image rankings versus the convolutional neural network (CNN). The correlation matrix shows Spearman's correlation coefficient values between the CNN image ranking, individual expert grader's image ranking, and the expert graders' consensus ranking. Experts were highly correlated with one another and the consensus ranking. The CNN performed nearly as well as individual experts on the ranked data set, as is demonstrated by the high correlation value to the expert consensus ranking.

DISCUSSION

We developed a model for the automated assessment of retinal fundus images in retinopathy of prematurity using a deep convolutional neural network. There are two key findings in this study: (1) with a high degree of confidence, the model can distinguish between images of acceptable quality and images of low or questionable quality, and (2) the model ranks image quality similarly to ROP experts, regardless of image quality classification, suggesting that the threshold at which images are classified as AQ or PAQ could be adjusted based upon the model's application.

The use of 5-fold cross-validation allowed us to train multiple models using all available training data while limiting the risk of overfitting. This finding is illustrated in **Figure 2A**, which shows that all models perform similarly to one another. The aim of cross-validation is to estimate test set performance. The mean (SD) of the five models was 0.958 (0.005). We used the best performing model to assess the independent test set (**Figure 2B**). The AUC was 0.965, similar to the mean (SD) predicted by 5-fold cross-validation. Taken together, we believe that this model has not overfit the data and that it is highly generalizable to RetCam-acquired ROP images.

An interesting result presented during model assessment on the ranked image data set. When training the CNNs, we cast our problem as a classification task. That is, we only cared to classify images as AQ or PAQ, and were never concerned about the intra-class ordering of images. However, to ensure applicability in use cases where the threshold at which AQ versus PAQ images may be different, it is important for the algorithm to be able rank image quality from worst to best, regardless to which quality class our experts believe an image belongs. In essence, we were testing the ability of the CNN to perform regression, even though it was only trained for classification. On a smaller data set of 30 images, six experts ranked images from worst to best quality for the

diagnosis of plus disease via an exhaustive pairwise comparisons process. This provided us with the individual rankings for the image set for each expert, which we were able to combine into an expert consensus ranking. All experts were highly correlated with one another (0.89-0.97) and, unsurprisingly, with the consensus ranking (0.94-0.97; **Figure 3**). Rather than have the CNN output class labels for each of the 30 images, we collected the probabilities of each image belonging to the AQ class and ordered them from smallest to largest, thereby establishing the CNN's ranking of the 30 images. The CNN was highly correlated to each individual expert (0.86-0.93) and to the consensus ranking (0.90; **Figure 3**). These results show that our model has a striking ability to rank images, further suggesting that the threshold at which our model classifies images as AQ or PAQ could be adjusted depending upon application.

Overall, these findings demonstrate the robustness of our model: it correctly identifies what our experts consider to be acceptable quality images vs. low and questionable quality images. This study also demonstrates that the threshold at which the model classifies images could be adjusted for other experts or applications. For example, in a telemedicine application where physicians manually review images, the model would likely remain unchanged since it was trained using the opinions of ROP experts. However, implementation as a prescreening method for a computer-based image analysis tool, such as DeepROP, may warrant some modifications. It is possible that a computer-based image analysis tool could still provide a reliable ROP diagnosis on a subset of PAQ images. Therefore, the threshold at which images were binned into the AQ versus PAQ category could be lowered until all images placed into the PAQ category could not be assessed via the computer-based method.[6,7,60,61,63,65,86]

While we are confident in the model we have trained, there are some limitations. First, only RetCam images were used for training and testing. We did not evaluate model performance on

images from other cameras. Differences in field-of-view and lighting could potentially affect the reliability of the model. Recently, ophthalmic lenses for smartphones have been created.[87,88] An interesting area of potential research involves training our model to accurately assess the quality of images acquired from these devices, thereby greatly enhancing the reliability of telemedicine applications. Second, the model was trained using images acquired from premature infants during routine ROP screenings. It is unclear whether this model can accurately classify images acquired from adults or older children with other ocular conditions, and further training of this model with images from those demographics would be beneficial. Third, the model was trained and validated on posterior pole images. In practice, nasal, temporal, superior, and inferior images may be used in addition to posterior pole images for diagnosis of ROP.[16] Further training of this model will include images from various regions of the retina to increase reliability and applicability in true clinical applications. The final limitation of this model is the lack of ability to distinguish a retinal fundus image from images of other items (i.e., non-ophthalmic images). This model was trained as a retinal fundus image quality classifier, not as a general image quality classifier. One could argue that users of this model will only be acquiring and assessing retinal fundus images. But to ensure conformity, a future direction of this work could involve training a CNN to classify images as retinal fundus images or not prior to images being assessed for quality.

We are not the first group to produce a retinal image quality classifier; however, many other classifiers have severe limitations. To the best of our knowledge, Saha et al. have produced the only other retinal image classifier that takes advantage of a CNN.[89] They used AlexNet, an award-winning but older CNN, for assessing the quality of diabetic retinopathy images. Their model performed with an accuracy of 100% on a data set of 3,572 images. However, their image set only included images on which all graders agreed upon the quality of the images (i.e. images without

complete agreement were excluded from the test set) which could leave the data with a very bimodal distribution. Furthermore, their data set was severely imbalanced, as only 143 of the 3,572 images were of unacceptable quality. In theory, a naive model (one that only predicts AQ for every image) would be correct 96% of the time. Consequently, it is possible that their CNN would not generalize well in practice. Other groups have implemented linear algorithms for image quality assessment of retinal fundus photos, which have performed well, but all training and test data sets were small in comparison to the data set we used to train, validate, and test our CNN.[68,72,80,81] We believe that, because our CNN was rigorously trained on 4,000 images using cross-validation and tested on two separate test sets consisting of 2,109 images and 30 ranked images, it will better generalize and be more robust in practice.

CONCLUSION

In this study, we implemented a convolutional neural network for the assessment of retinal fundus image quality in retinopathy of prematurity. We have shown that a convolutional neural network is sufficient for providing a high degree of discrimination between acceptable quality and possibly acceptable quality images, and can rank a set of retinal fundus images from worst to best quality. Potential applications of this algorithm range from inclusion in computer-based image analysis pipelines to implementation in fundus cameras, where imaging technicians could be alerted as to whether their captured images were of acceptable quality for diagnosis of disease. More broadly, it should be noted that this methodology is not limited to retinopathy of prematurity or retinal fundus imaging, as it has potential application in different ocular diseases or for different imaging modalities altogether.

AIM 2: A RISK MODEL FOR TREATMENT-REQUIRING ROP

ABSTRACT

Retinopathy of prematurity (ROP) is a leading cause of blindness in children, although it is often preventable with accurate and timely diagnosis and treatment. ROP screening guidelines are designed to be highly sensitive to avoid missing cases of treatment-requiring (TR-) ROP; consequently, approximately 80% of exams in a screening population have no or mild disease. Current ROP risk models require multiple predictors and/or exams, and performance often decreases significantly when applied to more diverse populations. We aimed to develop a risk model that could reduce the screening burden without missing cases of TR-ROP by using demographic risk factors and a deep learning-derived vascular severity score (VSS, all of which can be evaluated during a single exam) using a large cohort of North American infants.

A multi-institutional ROP dataset consisting of retinal fundus images and clinical factors for 852 subjects was collected as part of the Imaging and Informatics in ROP (i-ROP) study. A reference standard ROP diagnosis was provided for each exam. Posterior pole images were assigned a vascular severity score ranging from 1.0 to 9.0. Considering that infants who develop TR-ROP often have increasing VSS prior to the diagnosis of TR-ROP, we developed a risk model based on demographic risk factors and the VSS at 32-33 weeks post-menstrual age. Using all combinations of birth weight, gestational age (GA), and VSS, seven ElasticNet logistic regression models were tuned via five-fold cross-validation. The best-performing model was evaluated using the held-out

i-ROP test dataset consisting of 121 infants, and an independent dataset of 30 infants screened as part of a telemedicine program in Salem, OR.

The best performing model used GA and VSS, based on the area under the precision-recall curve. On each independent test set, the model achieved sensitivity of 100% with a positive predictive value ranging from 12% to 18%, and specificity ranging from 55% to 68% with a negative predictive value of 100% (NPV).

This model, with just two predictors which can be collected during a single exam, can identify all subjects who will eventually develop TR-ROP, while correctly ruling out, with 100% NPV, more than half of those who will not.

INTRODUCTION

Retinopathy of prematurity (ROP) is a leading cause of childhood blindness, despite the fact that visual impairment is often preventable with appropriate screening and treatment.[1,3] In the United States, screening is indicated for any infant born prior to 31 weeks of gestation or with a birthweight less than 1501 grams; it is performed via dilated retinal examination, either in-person or by telemedicine.[1,2,16,18,82] Current screening guidelines are highly sensitive in order to identify all cases of treatment-requiring (TR-) ROP; however, they are not specific, with approximately 80% of examinations revealing no or mild ROP.[2,16,18,82,90] Since these exams can occupy a significant portion of physicians' schedules and can be physiologically stressful to infants, risk models that can reduce the screening burden without missing severe disease are desirable. Several approaches have demonstrated efficacy for improving the efficiency of physician time without misclassifying babies that develop severe disease. For instance, telemedicine for ROP screening has been used for over a decade.[7–9,16,91,92] Typically, telemedical exams are performed weekly, whereas in-person ophthalmic exams can often be performed bi-weekly.[7–9,16] Thus, while telemedicine expands the geographic area physicians can cover, with less time, it does not lower the overall number of exams, nor does it reduce the physiological stress placed on already fragile premature infants.

Although birth weight (BW) and gestational age (GA) are the two most indicative predictors of those who might develop TR-ROP, they are not specific.[1–3] In fact, roughly 90% of children born prior to 31 weeks of gestation and those that weight less than 1501 grams at birth do not develop TR-ROP.[13] Other risk factors, such as necrotizing enterocolitis (NEC) and intraventricular hemorrhages (IVH), have been associated with incident TR-ROP, but, due to their rarity and the

confounding risk of gestational age and birthweight, they are of limited predictive value for TR-ROP.[13,19]

The goal of developing risk models is to improve specificity without sacrificing sensitivity since the risk of a false negative is potentially a blind infant. Various ROP risk models have been developed with this goal in mind.[1,3,10,93] However, concerns around performance and practicality have often hindered their implementation. Previous work has suggested that there may be added value in using an artificial intelligence-derived vascular severity score (VSS) in a predictive model. Specifically, Bellsmith et al. and Taylor et al. demonstrated that eyes that eventually developed TR-ROP had, on average, increasingly worse vascular severity beginning as early as 32 weeks postmenstrual age (PMA).[94,95] This is consistent with previous clinical studies that have suggested that the presence of pre-plus disease is predictive of incident TR-ROP, with roughly 40% of infants who develop pre-plus disease eventually developing plus disease.[1,3,90] However, the clinical variability of the subjective diagnosis of pre-plus disease limits the effectiveness of this clinical observation.

Thus, there remains a gap in knowledge as to whether objective evaluation of vascular features might add specificity to current risk models. We hypothesize that a risk model that takes advantage of vascular severity can have relatively high specificity while maintaining extremely high sensitivity. In addition, a model such as this would, theoretically, only require a single examination to occur at 32–33 weeks PMA, thereby reducing the screening burden placed on physicians and the associated physiological stress on infants. Herein, we describe the development of such a model. We aimed to achieve 100% sensitivity and at least 50% specificity using easy-to-obtain clinical factors and VSS in a highly interpretable logistic regression model.

METHODS

i-ROP Study Details

As part of a multicenter ROP cohort study, 835 unique subjects weighing less than 1501 grams at birth that were born prior to 31 weeks of gestation were screened for ROP between January 2012 and Present. This study was approved by the Institutional Review Board at the coordinating center (Oregon Health & Science University) and at each of seven study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Weill Cornell Medical Center). This study was conducted in accordance with the Declaration of Helsinki. Written, informed consent for the study was obtained from parents of all enrolled infants.

For every exam, subjects were examined clinically at the bedside, while also receiving image-based ROP diagnoses (None, Mild, Type-II ROP, or Type-I ROP) by a consensus of three ROP experts using the full International Classification of ROP (ICROP) criteria. Clinical comorbidities and demographics were recorded for every subjects' exam, such as BW, GA, the presence of IVH, sepsis, periventricular leukomalacia (PVL), and NEC. During each screening exam, at least five different retinal fundus image views (nasal, temporal, inferior, superior, and posterior-pole) were captured via a RetCam (Natus; Pleasanton, CA).

Vascular Severity Score and Dataset Preparation

To be included in this study, subjects' posterior-pole images were required to have consensus agreement by experts that the quality of said images were acceptable for the diagnosis of ROP. Posterior-pole images were analyzed by i-ROP DL, an automated plus disease classifier currently under approval review by the Food and Drug Administration.[6] i-ROP DL provided a softmax

probability for each image as having normal, pre-plus, or plus disease vasculature. A softmax probability is used for multi-class classification. It outputs the probability of each class being the correct label; however, all three values must sum to 1.0. From these values, a VSS, ranging from 1.0 to 9.0, was developed:

$$Vascular\ Severity\ Score = P(normal) + 5 * P(preplus) + 9 * P(plus).$$

Based on prior work, we believed that the window in the 32–33 week PMA range could be predictive of future TR-ROP, and thus used the first eye examination to occur for each subject in this window.[94,95] Because our goal was to develop a model that predicted future TR-ROP, we excluded infants who were diagnosed with TR-ROP within this window from the training dataset — specifically, if they developed TR-ROP within seven days of the first exam to occur within the 32–33 week PMA window — since that would be less predictive and more diagnostic of TR-ROP, the efficacy of which has previously been documented.[6] However, the test dataset contained all infants who would be eligible for ROP screening, (e.g., those born prior to 31 weeks of gestation that weighed less than 1501 grams at birth), regardless of if and when they developed TR-ROP following prediction, as this better mimics real-world usage. There were 376 and 444 unique subjects in the training and test datasets, respectively. Eyes were considered independently, and were mutually exclusive to the train or test datasets. Thus, the training dataset contained 58 eyes that eventually developed TR-ROP and 660 eyes that did not; the test dataset contained 133 eyes that eventually developed TR-ROP and 729 eyes that did not. Some eyes did not have acceptable quality images for diagnosis via i-ROP DL, hence the slight discrepancy between the number of subjects and the number of eyes.

Risk Model Development

Correlations between all collected clinical factors (VSS included) and TR-ROP were evaluated. Clinical factors with low correlation coefficients or low representation in the dataset were eliminated as possible model features. The remaining features were evaluated via recursive feature elimination in multiple ElasticNet models, a type of logistic regression that uses a mixture of L1 and L2 regularization to reduce the potential for overfitting.[31] These models were trained using the Sci-Kit Learn package developed for the Python programming language.[96] The ElasticNet mixing parameter was tuned via five-fold cross-validation using the following values [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]. Values of 1.0 and 0.0 are equal to L1 and L2 regularization, respectively. Due to the significant class imbalance (i.e., those who eventually developed TR-ROP versus those who did not), area under the precision-recall curve (AUPRC) was the primary measure of overall model performance; however, area under the receiver operating characteristics curve (AUROC) was also evaluated.

The performance of the model with the highest AUPRC was assessed via the F_β score (also known as the F-score or F-measure) using five-fold cross-validation across 101 evenly distributed decision thresholds from 0.00 to 1.00. Whereas the F_1 score ($\beta = 1$) attempts to balance the proportion of false negatives to false positives, increasing β (e.g., F_2, F_3, etc.) prioritizes minimizing false negatives over minimizing false positives. The F_2 score is commonly used to slightly prioritize minimization of false negatives. To be sure that minimization of false negatives was the top priority, β was set equal to 4. The average decision threshold minus the standard deviation that maximized the F_4 score was selected and used to evaluate both test datasets.

The final tuned and thresholded model was evaluated on the held-out i-ROP test dataset. It was also evaluated on an independent dataset, collected between September 2015 and June 2018, from 30 unique subjects born at a hospital in Salem, OR. Data collection and exclusion criteria were similar to that of the i-ROP dataset. Retrospective evaluation of these data was performed under a waiver of consent from the Oregon Health & Science University Institutional Review Board. In total, there were four eyes that developed TR-ROP, and 56 eyes that did not. The main outcome measures were sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

RESULTS

Either due to low correlation ($correlation\ coefficient < 0.25$) or lack of sufficient data to develop a well-powered model, IVH, PVL, NEC, and sepsis were eliminated as possible features. ElasticNet was tuned via five-fold cross-validation for all possible combinations of the remaining features: BW, GA, and VSS. An ElasticNet model with an L1 ratio of 0.4 with the predictors GA and VSS evaluated at 32–33 weeks PMA had the highest AUPRC (0.35 ± 0.11, **Table 1, Figure 1**). It was, therefore, evaluated on both test datasets.

Table 1: Five-fold cross-validation results for every combination of birth weight, gestational age, and vascular severity score.

Variables	AUPRC	AUROC	L1 Ratio
BW	0.21 ± 0.14	0.77 ± 0.12	0.0
GA	0.23 ± 0.20	0.79 ± 0.09	1.0
VSS	0.29 ± 0.05	0.76 ± 0.03	0.0
BW + GA	0.23 ± 0.20	0.78 ± 0.10	0.0
BW + VSS	0.32 ± 0.13	0.82 ± 0.11	0.0
GA + VSS	0.35 ± 0.11	0.82 ± 0.07	0.4
BW + GA + VSS	0.31 ± 0.11	0.81 ± 0.11	0.0

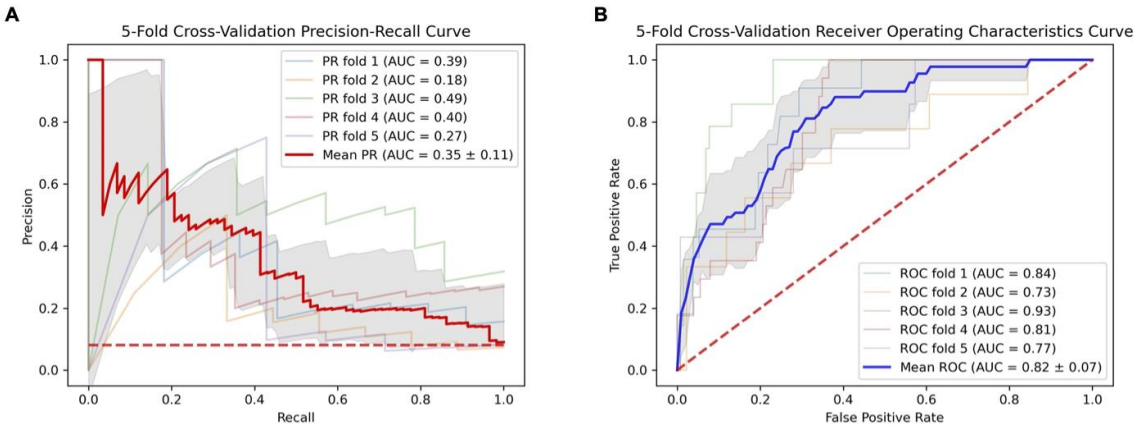

Figure 1: Areas under the precision-recall and receiver operating characteristics curves for the GA + VSS model. For the variables gestational age and vascular severity score, the mean ± standard deviation of the (A) AUPR and (B) AUROC, respectively, were 0.35 ± 0.11 and 0.82 ± 0.07.

The maximum F_4 score and corresponding decision threshold was 0.74 ± 0.12 and 0.33 ± 0.08, respectively. To be sure that all cases of TR-ROP were correctly predicted, the average threshold (0.33) was lowered to 0.25, which was the lowest bound suggested by the standard deviation (0.08). Although this threshold has the consequence of increasing false positives, it significantly minimizes false negatives, which is essential for any actionable ROP screening model. This model and decision threshold were then evaluated on the held-out test dataset from the i-ROP database (**Table 2**). It correctly identified all eyes that would eventually require treatment (sensitivity: 100.0%, PPV: 27.0%) while correctly ruling out more than half of the eyes that would never would (specificity: 51.0%, NPV: 100.0%). For children who developed TR-ROP after the prediction (i.e., they were not identified as having TR-ROP at the time of exam), the average number of weeks ± standard deviation to TR-ROP diagnosis was 3.6 ± 2.6 weeks, range [0.14, 11.0] weeks. This model and decision threshold were then evaluated on an independent test dataset collected from a

hospital located in Salem, OR (**Table 2**). Again, the model correctly identified all eyes that would eventually require treatment (sensitivity: 100.0%), however PPV dropped to 15.5%. On the other hand, specificity increased to 59.2%, with NPV still 100.0%. The average time to TR-ROP diagnosis following prediction, for this dataset, was 4.5 ± 0.6 weeks, range [4.0, 5.0] weeks.

Table 2: Confusion matrix of the GA + VSS model evaluated on both the i-ROP and Salem test datasets at the statistically optimized decision threshold of 0.25.

		True Label			
		i-ROP Test Dataset		Salem Test Dataset	
		Not Treated	Treated	Not Treated	Treated
Predicted Label	Not Treated	372	0	71	0
	Treated	357	133	49	9

A post-hoc analysis was performed to examine the VSS at the first exam to occur at 34–35, 36–37, and 38–39 weeks PMA for all children who were predicted (whether correctly or incorrectly) to develop TR-ROP. It was found that VSS was significantly more severe ($P < 0.05$) at 34–35, 36–37, and 38–39 weeks PMA for children who required treatment versus those who did not (**Figure 2**).

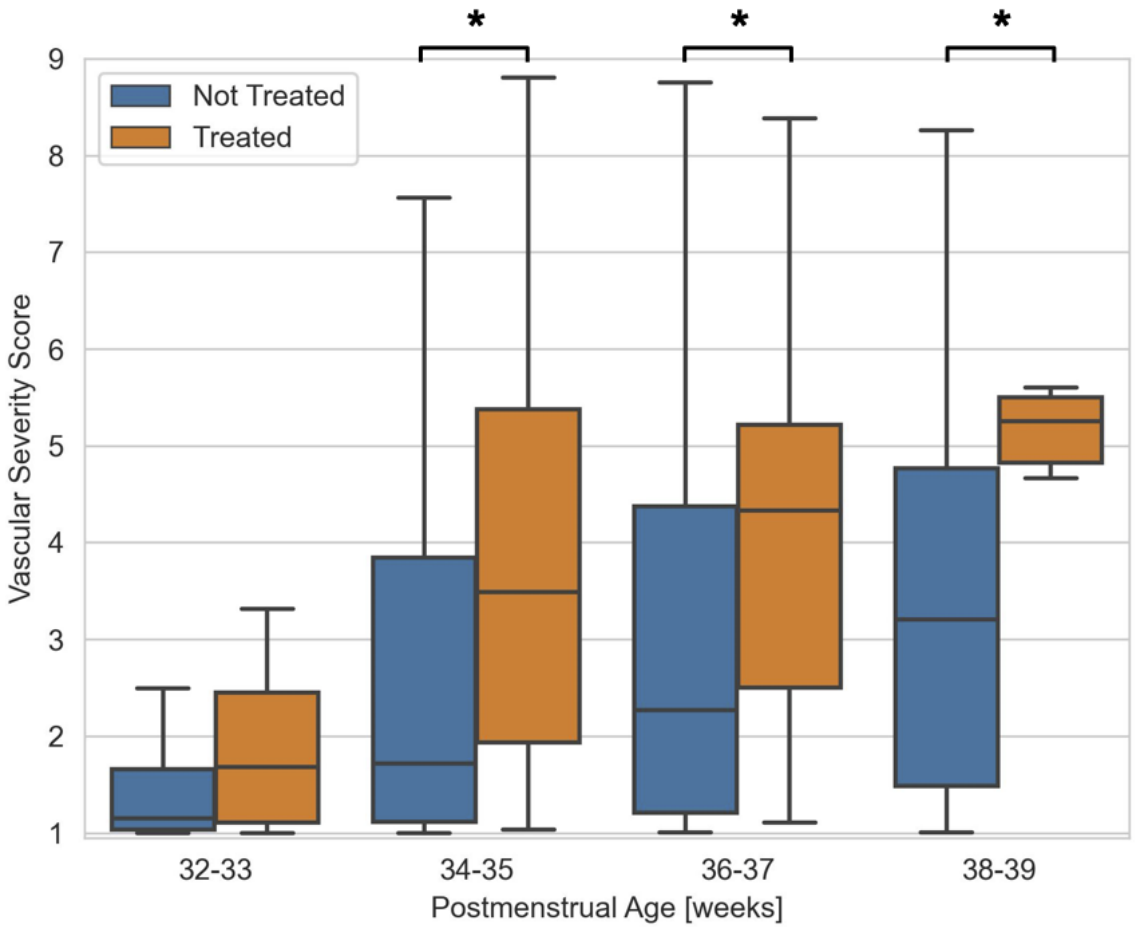

Figure 2: Post-hoc analysis of VSS from subjects predicted to develop TR-ROP. The VSS for subjects who developed TR-ROP trends up and away from those who did not. Asterisks indicate significantly different groups ($P < 0.05$).

DISCUSSION

These results demonstrate that our risk model, which uses just two features, can identify all infants that will require treatment for ROP more than one month prior to a TR-ROP diagnosis, while correctly reducing the screening pool by more than half. Using five-fold cross-validation, we trained and tuned multiple ElasticNet logistic regression models on all possible combinations of BW, GA, and a deep learning-derived VSS, which was evaluated at 32–33 weeks PMA. We found that a combination of GA and VSS produced the model with the highest AUPRC (**Table 1**). We tuned this model's decision threshold, via five-fold cross-validation, using the F_4 score. On a held-out test dataset and on an independent test dataset, the model had 100.0% sensitivity, and specificity greater than or equal to 51.0% (**Table 2**). There were two key findings: (1) a VSS, evaluated at 32–33 weeks PMA, when coupled with GA, can predict all subjects who will eventually develop TR-ROP while maintaining a high degree of specificity, and (2) following the VSS of those predicted to develop TR-ROP beyond 32–33 weeks PMA may provide further specificity.

We hypothesize that VSS captures subtle vascular abnormalities that are not normally cause for concern, nor warrant weekly screenings. This was evidenced by the univariate VSS model, which had an AUPRC 0.07 points higher than the BW or GA univariate models, or the combination thereof (**Table 1**). Furthermore, of those who developed TR-ROP, all had a VSS less than 4 at 32–33 weeks PMA, which equates to a clinical plus disease diagnosis of "normal."[6,20] However, of those predicted to develop TR-ROP, those who actually developed TR-ROP appeared to have an increased, albeit not statistically increased, VSS at 32–33 weeks PMA as compared to those who did not (**Figure 2**). Therefore, we believe that the minute differences between VSS 1, 2, and 3,

when evaluated at 32–33 weeks PMA, have significant power for the prediction of future development TR-ROP.

Performance-wise, the GA + VSS model is comparable to that of the initial performance calculations achieved by the CHOP ROP model, which uses a combination of BW + GA + weight gain to predict future occurrences of Type-II and Type-I (TR-) ROP.[10,11] Both models achieved a sensitivity of 100.0% in predicting TR-ROP. PPVs were comparable (CHOP ROP: 17%, GA + VSS: 27.0% and 16.0% for i-ROP and Salem test datasets, respectively). The CHOP ROP model had a specificity of 53%, whereas the GA + VSS model had specificities equal to 51.0% and 62.5% on the held-out i-ROP test set and the independent Salem, OR test set, respectively. NPV was 100.0% for both models. However, when the CHOP ROP model was applied to a larger and more diverse cohort collected from infants admitted to 30 hospitals spread across North America — similar to the i-ROP dataset — the decision threshold had to be significantly lowered to achieve 100.0% sensitivity, which resulted in a specificity of just 6.8%.[11] This finding suggests that the CHOP ROP model does not generalize well, which is likely due to the fact that the CHOP ROP model was only trained on infants that were admitted to a single Philadelphia, PA hospital.[10] In contrast, the GA + VSS model was trained using the i-ROP database, which contains exam-level information for infants admitted to eight different hospitals spread across the United States. It is, therefore, more likely to better generalize, as was evidenced by both test datasets.

Furthermore, the GA + VSS model requires data that can be collected during a single ROP screening to make an accurate prediction. GA, even in developing countries, is generally trivial to calculate, and a VSS can be easily provided by uploading a digital retinal fundus image to a secure web server.[97] Admittedly, acquiring retinal fundus images during routine ROP screenings is not part of the current standard of care in many neonatal care units, whether in developed or developing

countries.[16] Furthermore, expensive retinal fundus cameras are, presumably, not commonplace in most developing countries. However, as neonatal care and documentation advance, and retinal fundus camera attachments for smartphones become more prevalent, this may soon become a nonissue.[98–100] Regardless, we maintain that the simplicity of this model is an advantage over other risk models, which often require multiple examinations and detailed records of weight gain, comorbidities, etc., and will become more practical as digital fundus photography becomes more widespread.[10,19]

Still, there is one major limitation regarding generalization. Although this model appears robust and able to generalize well to the North American populace, it will likely need to be retrained for populations in other regions of the world. The GA (and BW) of children who develop TR-ROP in developing countries is often higher than those in developed countries.[101,102] Therefore, if this model were to be used to screen for future TR-ROP patients in these regions, it would need to be retrained on a sizable dataset collected from subjects in those regions. However, the VSS is determined by first segmenting the retinal vasculature of retinal fundus images into grayscale retinal vessel maps so that race and pigmentation do not affect the overall diagnosis and has been shown to perform well on infants from other countries and regions. Therefore, retraining the model would be trivial assuming a sizable training dataset could be acquired.

In the future, we plan to further validate this model on a larger cohort of ROP patients. We would also like to incorporate other features to further increase specificity of subjects who are predicted to develop TR-ROP. Although our model outperformed the CHOP ROP validation study on a large North American cohort, and has some advantages over it regarding implementation, there are potential performance increases to be had if elements of the two models were combined. We also plan to investigate the role of oxygen exposure and saturation, and how it pertains to TR-ROP.

CONCLUSION

In conclusion, we have trained, tuned, and thresholded a highly interpretable, parsimonious model for the prediction of infants who will eventually develop treatment-requiring retinopathy of prematurity. With prospective validation, we have demonstrated that this model can identify all infant eyes that will develop TR-ROP and reduce the screening burden by more than 50%, thereby prioritizing care to those who are most at-risk of developing TR-ROP, while simultaneously reducing the physiological stress placed on those who will never require treatment.

AIM 3A: CONVERTING RETINAL VESSEL MAPS INTO RETINAL FUNDUS IMAGES

ABSTRACT

Advances in generative adversarial networks have allowed for engineering of highly realistic images. Many studies have applied these techniques to medical images. However, evaluation of generated medical images often relies upon image quality and reconstruction metrics, and subjective evaluation by laypersons. This is acceptable for generation of images depicting everyday objects, but not for medical images, where there may be subtle features experts rely upon for diagnosis.

We implemented the pix2pix generative adversarial network for retinal fundus image generation and evaluated the ability of experts to identify generated images as such and to form accurate diagnoses of plus disease in retinopathy of prematurity. We later implemented pix2pixHD, and also evaluated whether experts could identify generated images from real images.

We found that, while experts could discern between real and generated images produced by pix2pix, the diagnoses between image sets were similar. Images generated by pix2pixHD could not be identified by experts. By directly evaluating and confirming physicians' abilities to diagnose generated retinal fundus images, this work supports conclusions that generated images may be viable for dataset augmentation and physician training.

INTRODUCTION

Advances in graphics processing units have allowed for development of complex models, such as deep neural networks and variants thereof.[39,103] Generative adversarial networks (GAN) are one such variant. These models contain both discriminative and generative networks that are trained to deceive one another.[47,49,50] A discriminative network attempts to estimate an output, y, given a set of inputs, x.[47,49,50] In contrast, a generative network attempts to model the distribution of x given y. To train these networks, data are supplied in pairs – inputs and their corresponding output(s). These two models are pitted against one another, and as training progresses, the ability of each model improves (i.e., as the discriminator better learns to discern between real and generated images, the generator must also learn how to better simulate generated data). Ideally, this results in a generator that consistently fools a well-trained discriminator into classifying its outputs as real. These models can be used for many types of data, but are primarily used for image synthesis. For images, both the discriminative and generative networks attempt to learn the overall style and pattern of output images (i.e., the relevant features of output images). However, the generative network also tries to learn how to map the original input image to the style of the output image. These models have begun to gain traction for synthesis of medical images.[104–107]

The model presented in Image-to-Image Translation with Conditional Adversarial Networks, pix2pix, allows for one to employ style transfer without having to hardcode the style mapping (**Figure 1, 2**).[49,50] pix2pix has been used to map real images to labels, labels to images, convert black-and-white images to color, and convert images captured during the day to representations of the same images at night. A few studies have even used this method to map retinal blood vessel maps to retinal fundus images for research in diabetic retinopathy.[104,105,107] However, while these generated images have been evaluated using subjective visual quality inspections and various

image quality/reconstruction metrics, the diagnosability of said images — arguably the most important factor — has never formally been evaluated.

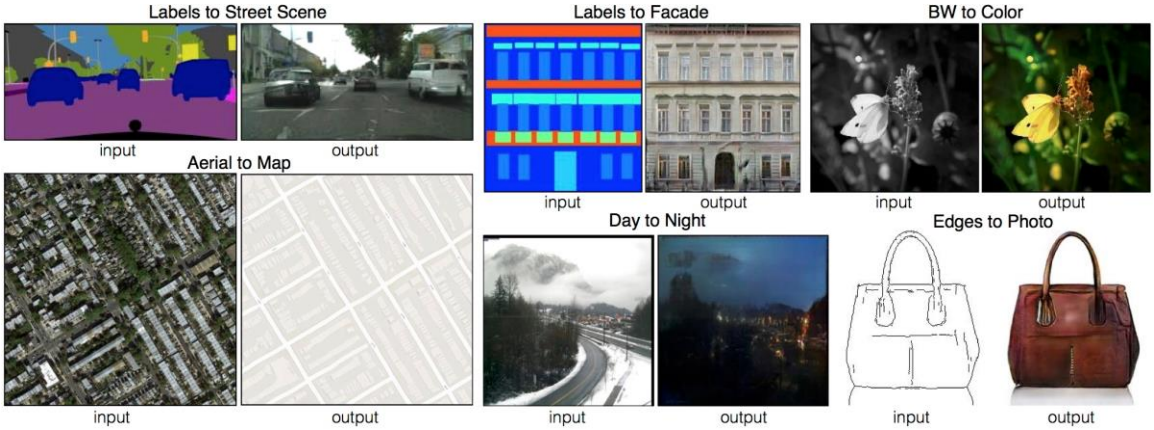

Figure 1: Example implementations of the pix2pix generative adversarial network. This model demonstrates excellent ability to convert feature maps to real images (Labels to Street Scene, Labels to Facade, Edges to Photo), and real images to feature maps (Aerial to Map). The results are realistic and of relatively high resolution. Figure adapted from Image-to-Image Translation with Conditional Adversarial Networks.[49]

In this study, we have deployed the pix2pix and pix2pixHD pipelines for retinal fundus image synthesis from retinal vessel maps.[49,108] The application of this work is in retinopathy of prematurity (ROP), a potentially-blinding disorder that affects premature infants [1,49]. A significant predictor of treatment-requiring ROP is the presence of plus disease, described as venous dilation and arterial tortuosity (**Figure 2**).[1] It stands to reason that, according to the definition of plus disease, the only information required to diagnose plus disease is the appearance of the major retinal blood vessels.[1] To generate synthetic ROP retinal fundus images, we first generate retinal vessel maps from retinal fundus images using a previously-reported U-Net model, then create new retinal fundus images of varying pigmentations from original images using the raw retinal vessel maps, and also create new retinal fundus images of varying pigmentations that lack choroidal blood

vessel patterns and/or any unique/abnormal features of the retina (e.g., haemorrhages, discolorations, etc.) using filtered retinal vessel maps.[6,45] Data for this study were obtained through the multi-center, NIH-funded, Imaging and Informatics in ROP (i-ROP) study centered at Oregon Health & Science University (OHSU).

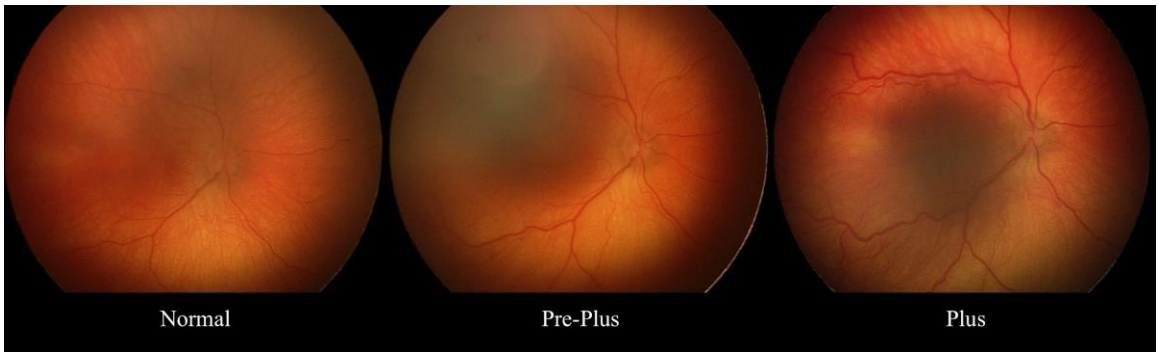

Figure 2: Example retinal fundus images. From left to right, retinal fundus images of an eye that was originally diagnosed normal, developed pre-plus disease, and then plus disease. In plus disease images, retinal blood vessels are dilated and tortuous as compared to normal images. The degree of dilation and tortuosity of pre-plus blood vessels is less than that of plus disease blood vessels, but greater than normal.

METHODS

Institutional Review Board

This study was approved by the Institutional Review Board at the coordinating center (OHSU) and at each of 8 study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Weill Cornell Medical Center, Asociación para Evitar la Ceguera en México [APEC]). This study was conducted in accordance with the Declaration of Helsinki. Written informed consent for the study was obtained from parents of all infants enrolled.

Retinal Fundus Image Dataset

As part of the multicenter ROP cohort study, i-ROP, over 30,000 nasal, temporal, inferior, superior, and posterior-pole retinal fundus images were collected from 970 preterm infants during routine ROP screening examinations. Between three and eight independent experts labeled each image set as normal, pre-plus, or plus, and an expert consensus diagnosis was formed and established as the ground truth diagnosis. Experts were all ophthalmologists with extensive experience in both ophthalmoscopic and image-based diagnosis of ROP. A subset of fundus images were selected; exclusion criteria were: images not centered on the posterior pole, images of stage 4 ROP (partial retinal detachment), and images of stage 5 ROP (total retinal detachment). The remaining images were downsampled to create a final dataset consisting of 6058 wide-angle fundus images centered on the posterior pole. This dataset was randomly split (retaining the even distribution), 80/10/10, into train, validation, and test datasets, respectively. Because a subject may be represented in the dataset more than once (multiple imaging sessions), it was ensured that subjects were unique to each dataset.

Model Setup and Training

Models were built and trained in Python using PyTorch on an Nvidia V100 GPU (Santa Clara, CA).[109] For each image in the training, validation, and test datasets, vessel maps were generated using a previously-trained U-Net.[6,45] The open-source pix2pix and pix2pixHD codes were forked from Github repositories hosted by their respective authors.[108,110] We applied a modified pix2pix GAN, using ResNet9 blocks, to the i-ROP training dataset (**Figure 3**). The value of λ was set at 10; this weights the L1 loss 10 times greater than the adversarial loss of the generator during training, resulting in objectively higher fidelity images. Additionally, the original model was designed to generate color images of size 256x256x3, but retinal fundus images are generally color images of size 640x480x3. Rather than upsampling generated 256x256x3 images, which resulted in slightly blurred images that could (A) affect diagnosability and/or (B) be more easily discerned as a generated image, the model was modified to produce images of size 640x480x3. During training, each image in an image set (vessel map and corresponding retinal image) was scaled to size 572x572x3, and a random 512x512x3 crop was acquired from the same location for each image in the set. When images were generated, the 512x512x3 output image was resized to 640x480x3 to match the size of retinal fundus images. Finally, a black, circular mask was applied around the outside of the image to better mimic the appearance of images captured by retinal fundus cameras. pix2pixHD was trained using the default settings, as they are already optimized to generate large, high-quality images. However, just as with pix2pix, a black, circular mask was applied around the outside of the image.

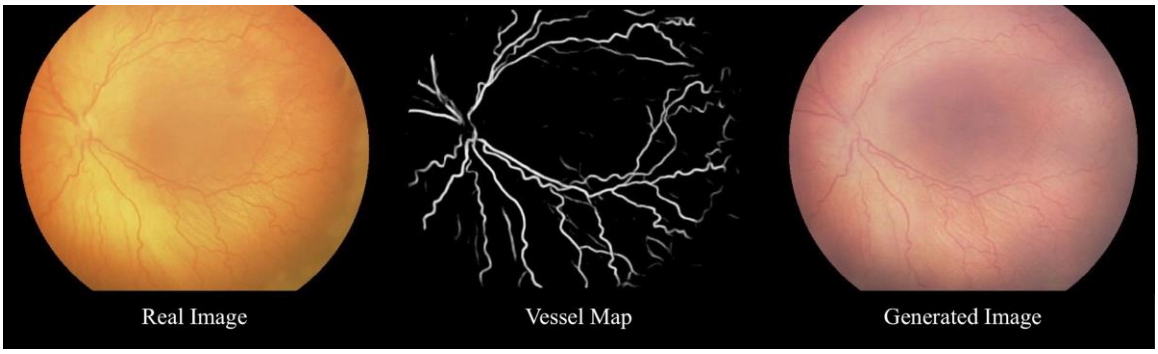

Figure 3: Retinal image generation process. Blood vessels of real images (left) are segmented and converted into retinal vessel maps (center). Pix2pix is used on raw or filtered retinal vessel maps (no discernible difference in image appearance) to generate images with similar vascular patterns (right).

For pix2pix, two separate models were trained: one on raw, grayscale vessel maps produced by the U-Net model (pix2pix-raw), and the other on the same vessel maps thresholded at pixel values greater than 25 (pix2pix-filtered). After the first training session, it was noted that although the U-Net was only trained to segment major retinal blood vessels, the reconstructed images contained similar choroidal blood vessel patterns. Upon further investigation, it was found that choroidal blood vessels were segmented, but at pixel intensities indistinguishable from the background to the human eye (pixel intensity < 26). Therefore, for the second training iteration, pixel values less than or equal to 25 were set to 0 to remove information about choroidal blood vessel patterns. The pix2pix models were trained for 1,000 epochs using the Adam optimizer with a β value of 0.5. During the first 500 epochs, the learning rate was constant at $2x10^{-4}$, and was linearly decayed to 0 over the remaining 500 epochs. pix2pixHD was trained for 200 epochs using the Adam optimizer with a β value of 0.5. During the first 100 epochs, the learning rate was constant at $2x10^{-4}$, and was linearly decayed to 0 over the remaining 100 epochs. Discriminator and generator loss functions on both the training and validation test sets were monitored to ensure learning was occurring at an

equal rate between objective functions, and that overfitting was not occurring. The quality of image reconstructions were evaluated using the structural similarity index (SSIM) and the peak signal to noise ratio (PSNR), metrics often used to describe the quality of image reconstruction.[111,112] pix2pixHD was trained on thresholded images only at a constant learning rate of $3x10^{-4}$ for 100 epochs, and linearly decayed to 0.0 over another 100 epochs.

pix2pix Image Grading

Of the 880 true retinal fundus images in the test dataset, 30 images were randomly selected for grading. Raw vessel maps and thresholded vessel maps were generated for each real image and used to generate reconstructions from their respective models. In total, there were three image sets (90 images) to be graded: 30 ground truth retinal fundus images, 30 reconstructions from pix2pix-raw, and 30 reconstructions from pix2pix-filtered. Using a custom, online grading system, a set of three independent ROP experts graded each image as normal, pre-plus, or plus, and also assessed whether they believed the image was real or generated.[23] All images were presented at a resolution of 640x480x3. In the event of a three-way tie for normal, pre-plus, or plus, the image was classified as pre-plus. The majority diagnoses of real images were used to compare agreement of diagnoses to generated image sets.

pix2pixHD Image Grading

Of the 880 true retinal fundus images in the test dataset, 50 images were randomly selected for grading. Thresholded vessel maps were generated for each real image and synthetic images were generated by pix2pixHD to create a total dataset size of 100 images, 50 real and 50 synthetic. Using the same custom, online grading system, a set of four separate ROP experts assessed whether they believed each image was real or generated.[23] All images were presented at a resolution of

640x480x3. Individual expert predictions were compared to ground truth, as well as the expert majority. In the event of a tie, the expert majority was deemed "fake."

Data Analysis

All analyses were performed in R. Majority diagnoses were determined for all images in a set, in addition to a majority vote on whether images were real or generated. Fisher's Exact Test for Count Data was used to determine if experts were statistically able to identify generated images from real images. In order to determine if expert diagnoses were affected by generated images, the Cochran-Mantel-Haenszel test was used to compare the pix2pix-raw and pix2pix-filtered contingency tables to the real image contingency table. Further, Cohen's kappa (κ) was used to measure agreement of diagnoses between generated images and real images.

RESULTS

The [discriminator, generator] losses on the training dataset for pix2pix-raw and pix2pix-filtered were [0.315, 0.224] and [0.170, 0.078], respectively. The [train, validation] PSNR values for pix2pix-raw and pix2pix-filtered were [16.882, 16.584] and [12.563, 12.014], respectively. The [train, validation] SSIM for pix2pix-raw and pix2pix-filtered were [0.617, 0.559] and [0.505, 0.448], respectively. A generator loss function value that is lower than a discriminator loss function value indicates that the generator can trick the discriminator into classifying its images as real more often than not. This occurred for all three models and suggested that each can generate realistic images. This was further confirmed by the SSIM and PSNR values of pix2pix models. The higher SSIM and PSNR value of pix2pix-raw images, as compared to pix2pix-filtered images, suggested that its generated images were more similar to true retinal fundus images; this was likely due to the presence of choroidal blood vessels (**Figure 3**). pix2pixHD was a small followup experiment to the original pix2pix experiments, which determined that SSIM and PSNR roughly tracked the loss statistics of the generator and discriminator, so these metrics were not explicitly monitored during pix2pixHD training. Rather, loss was evaluated.

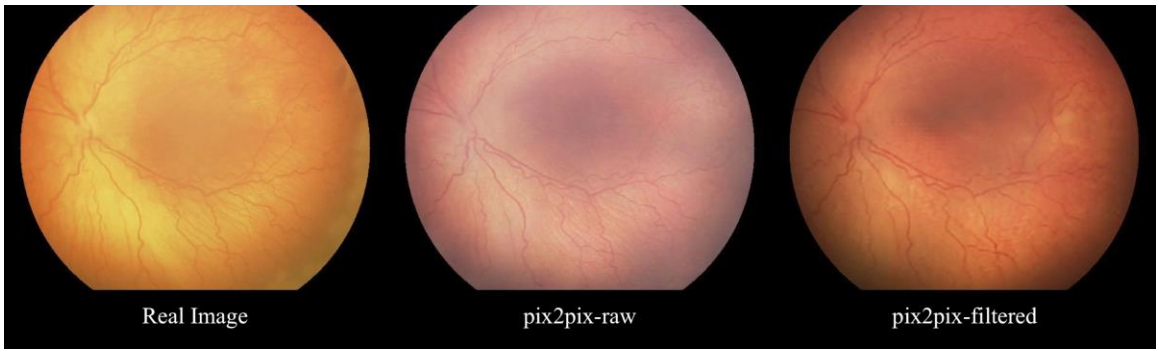

Figure 4: Examples of real and generated retinal fundus images. From left to right: a real retinal fundus image, a generated retinal fundus image from pix2pix-raw that uses raw vessel maps to create images with choroidal blood vessels patterns, and a generated retinal fundus image from

pix2pix-filtered that uses filtered vessel maps to generate images without choroidal blood vessel patterns.

In general, experts were able to discern between real and generated images produced by pix2pix-raw and pix2pix-filtered (Accuracy: 92.2%, **Table 1**). Images without choroidal blood vessel patterns (pix2pix-filtered) were identified as generated in 100% of cases. Some (16.7%) generated images that contained choroidal blood vessel patterns (pix2pix-raw) were classified as real images. This corroborates the difference in test set PSNR values between the two models. Nearly all (93.3%) of real images were identified as such. The $\chi 2$ test statistically confirmed that, overall, experts could identify real versus generated images ($\chi 2 \cong 64.019$; $p \cong 1.254x10\text{-}14$). The Fisher's Exact Test confirmed this finding ($p < 2.2x10\text{-}16$). Generated images with choroidal blood vessel patterns were, statistically, not more likely to be identified as real than those without ($\chi 2$: $p \cong 0.062$; Fisher: $p \cong 0.052$).

Table 1: Expert majority determination of pix2pix images

		Expert Majority Determination	
		Real	**Generated**
Image Type	**Real Images**	28	2
	pix2pix-raw	5	25
	pix2pix-filtered	0	30

Expert majority diagnoses for each image set are presented in **Table 2**. The majority diagnoses for real images were used as the ground truth. For normal images, experts diagnosed with accuracies of 92.3% and 100% on images generated by pix2pix-raw and pix2pix-filtered, respectively. For

pre-plus images, experts had 91.7% accuracy on both pix2pix-raw and pix2pix-filtered images. Plus disease was diagnosed with 80% accuracy on both pix2pix-raw and pix2pix-filtered images. A Cochran-Mantel-Haenszel test confirmed that real, pix2pix-raw, and pix2pix-filtered images were not graded dissimilarly ($p \cong 0.501$). This suggests that generated images, even those without choroidal blood vessel patterns, have the same diagnostic power as real images.

Table 2: Expert majority diagnoses of real images versus expert majority diagnoses of generated images

		Real Images			pix2pix-raw			pix2pix-filtered		
		Normal	**Pre-Plus**	**Plus**	**Normal**	**Pre-Plus**	**Plus**	**Normal**	**Pre-Plus**	**Plus**
Real Image Majority Diagnosis	**Normal**	13	0	0	12	1	0	13	0	0
	Pre-Plus	0	12	0	0	11	1	1	11	0
	Plus	0	0	5	0	1	4	0	1	4

To further investigate and confirm this finding, intergrader agreement of diagnoses and overall agreement of diagnoses between image sets were determined using weighted κ statistics for chance-adjusted agreement in ordinal diagnosis, using a well-known scale: [0, 0.20] = slight agreement, [0.21, 0.40] = fair agreement, [0.41, 0.60] = moderate agreement, [0.61, 0.80] = substantial agreement, and [0.81, 1.00] = near-perfect agreement (**Table 3**). For images generated by pix2pix-raw, individual expert diagnoses had substantial to near-perfect agreement (κ = [0.680, 0.880]) to real images, and the majority diagnoses had near-perfect agreement (κ = 0.880). For images generated by pix2pix-filtered, individual expert diagnoses had moderate to near-perfect agreement (κ = [0.498, 0.980]) to real images, and the expert majority diagnoses had near-perfect agreement (κ = 0.902).

Table 3: Expert agreement of diagnoses between generated images and real images

	Majority Diagnosis	Expert 1	Expert 2	Expert 3
pix2pix-raw	0.880	0.743	0.680	0.880
pix2pix-filtered	0.902	0.663	0.498	0.980

For pix2pixHD, Fisher's Exact Test p-values for the Expert Majority and Experts 1–4, respectively, were: 0.100, 0.505, 0.158, 1.000, and 0.043. This suggests that the majority of experts could not discern between real and synthetic images (**Table 4**). An example of a real image and one generated by pix2pixHD is presented in **Figure 5**.

Table 4: Expert majority determination pix2pixHD images

		Expert Majority		Expert 1		Expert 2		Expert 3		Expert 4	
		Real	**Fake**	**Real**	**Fake**	**Real**	**Fake**	**Real**	**Fake**	**Real**	**Fake**
Actual Image Type	**Real**	35	15	38	12	32	18	43	7	34	16
	Fake	26	24	34	16	24	26	44	6	23	27

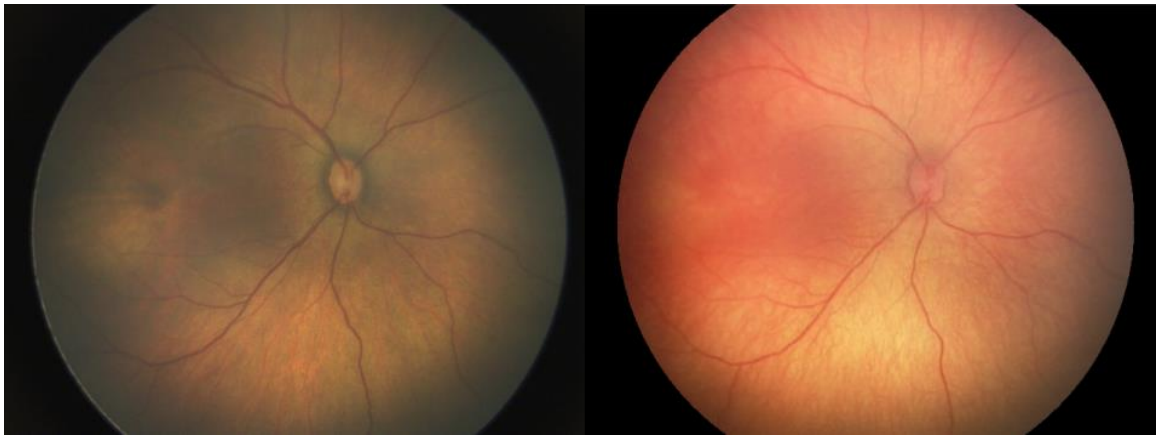

Figure 5: Example pix2pixHD image. A real retinal fundus image (left) and its corresponding generated image produced by pix2pixHD from a retinal vessel map (right).

DISCUSSION

This study aimed to generate and evaluate synthetic retinal fundus images, for the diagnosis of plus disease in retinopathy of prematurity, by segmenting the vasculature of real retinal fundus images into grayscale vessel maps using a U-Net, and generating realistic color retinal fundus images from said vessel maps using pix2pix and pix2pixHD. There are three key findings: (1) images generated by pix2pix, regardless of whether choroidal blood vessel patterns were present, were easily identified by experts as generated, (2) images generated by pix2pixHD, with synthetic choroidal blood vessel patterns, were not easily identified by experts as generated, and (3) generated images retained information relevant for the detection of plus disease in retinopathy of prematurity. The Chi-squared test suggested that pix2pix images were not realistic enough to deceive experts into believing they were real images ($p \approx 1.254 \times 10^{-14}$, **Table 1**). This was confirmed using a Fisher's Exact test ($p < 2.2 \times 10^{-16}$). However, a Cochran-Mantel-Haenszel test showed that images were realistic enough to be diagnosed similarly to real images ($p \approx 0.501$, **Table 2**). This was further confirmed by measuring the agreement of individual and majority expert diagnoses across image sets using Cohen's kappa (**Table 3**). For real images, experts had near-perfect agreement with the diagnoses of the same images reconstructed by pix2pix-raw and pix2pix-filtered ($\kappa = 0.880,\ \kappa = 0.902$). The Fisher's Exact test suggested that pix2pixHD images were realistic enough to deceive experts into believing they were real images ($p \approx 0.100$, **Table 4**). These results suggest that, although pix2pix images may be recognized as generated, these models retain relevant information that is required to reconstruct retinal fundus images for the diagnosis of plus disease in retinopathy of prematurity. Further, pix2pixHD creates highly realistic images that contain the same major retinal vascular information, but with greater detail in non-vascularized areas.

This work has many important implications. First, it confirms that, at least for plus disease diagnosis in ROP, pix2pix- and pix2pixHD-generated images are of high enough quality and fidelity for expert physicians to form accurate diagnoses. This is important, as numeric metrics and subjective layperson evaluations are often used to evaluate the quality and realism of generated images.[104,106,113,114] However, because the goals of such systems are often to generate images for synthetic datasets, training physicians, or diagnosis from image reconstructions, the ability of physicians to form diagnoses from generated images should be evaluated prior to implementation. It should also be noted that clinical findings unrelated to plus disease diagnosis that were present in original retinal fundus images, such as hemorrhages, were not present in reconstructions. This is likely because the U-Net was not trained to segment retinal hemorrhages, so pix2pix and pix2pixHD were unaware of their existence. While not detrimental for the diagnosis of plus disease, it serves as a warning to those looking to generate images of highly complex diseases where rare clinical findings may be highly-relevant for a given diagnosis. In essence, although an image may appear real or diagnosable, it may be lacking pertinent information that was present in the original image.

Second, retinal scans are listed as protected health information (PHI) according to the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule; however, they are de-identified via the Expert Determination method (§ 164.514(b)(2)).[115,116] This method essentially states that the risk of re-identification using retinal fundus images is negligible; however, the European Union's General Data Protection Regulation (GDPR) currently informs that this method is not sufficient for de-identification. It is conceivable that the pix2pix methods we have trained could be used to further de-identify retinal fundus images. From a purely observational standpoint, it was found that generated images were often pigmented differently than original images (**Figure**

3, 4). Second, generated images without choroidal blood vessel patterns were just as diagnosable as real images. Finally, other clinical findings present in retinal fundus images, such as hemorrhages, may be specific to unique diagnoses and increase the identifiability of images. As mentioned, in the i-ROP dataset, there were a limited number of images with these features; therefore, the few retinal fundus images that did have these highly identifiable features were reconstructed without them. Although this method would not fully de-identify images, it could further reduce the risk of re-identification by removing highly identifiable features while still providing physicians with the information required to form accurate diagnoses. One could argue that the retinal vessel maps used to generate retinal fundus images are computationally easier to generate and modify and might be considered further de-identified. However, physicians are not formally trained to form diagnoses from these types of images, and the accuracy and reliability of said diagnoses could suffer. Nonetheless, it is one of a few interesting future directions for this work.

CONCLUSION

We have implemented the pix2pix and pix2pixHD generative adversarial networks for the generation of retinal fundus images in retinopathy of prematurity and can successfully generate highly realistic retinal fundus images from retinal vessel maps. These generated images have the same diagnostic power as real images; images were easily diagnosed for the presence of plus disease by retinopathy of prematurity experts, and their diagnoses were highly correlated. This is important, as previous studies have not formally evaluated the ability of physicians to form diagnoses from generated medical images.

AIM 3B: AUGMENTING THE SEVERITY OF RETINAL VESSEL MAPS

ABSTRACT

Retinopathy of prematurity (ROP) is a blinding disease that affects prematurely born infants. A significant indicator of the need for treatment of ROP is the presence of plus disease, described as venous dilation and arterial tortuosity. The diagnosis of plus disease is, unfortunately, a dichotomous decision based on subjective comparison to an outdated reference standard image. In this work, we aim to create personalized reference standard images of plus disease for individual patients.

To do so, we explore the modification of retinal vessel maps. We use unpaired-image GANs to increase the severity of plus disease present in retinal vessel maps from normal or pre-plus disease to plus disease, or to augment the severity along a novel vascular severity score (VSS) scale ranging from 1.0 to 9.0. The retinal vessel maps were then converted into retinal fundus images using a previously trained paired-image GAN, and evaluated via DeepROP, an automated plus disease screening tool.

We found that converting images from normal or pre-plus disease vasculature to plus disease vasculature worked as intended almost every time. Converting from VSS 1 to VSS 2, VSS 3, …, VSS 9 proved to be more challenging, likely due to dataset size limitations. However, there was a general trend of increasing vascular severity produced by each model. Ultimately, this work may allow for generation of personalized reference standard images, which may better alert non-ROP

experts as to when an infant may require treatment, or at the very least be referred to an ROP expert.

INTRODUCTION

Although retinopathy of prematurity (ROP) is a highly-treatable disease, if diagnosed in a timely fashion, it remains one of the world's leading causes of childhood blindness.[1,2,18] While there are a handful of reasons for this phenomenon, a significant issue revolves around the accurate diagnosis of plus disease. The presence of plus disease, defined as venous dilation and arterial tortuosity, is required for the diagnosis of treatment-requiring (TR-) ROP in five out of six conditions under which ROP should be treated (**Figure 1**).[1,16] The International Classification of Retinopathy of Prematurity (ICROP) refers physicians to use a "standard" photograph to define the minimum amount of vascular dilatation and tortuosity required to make the diagnosis of plus disease (**Figure 2**).[1,3] However, many multi-centered clinical trials suggested that the diagnosis of plus disease should be made if sufficient vascular dilatation and tortuosity were present in at least 2 quadrants of the eye.[2] Furthermore, the original fundus image from the ICROP is outdated — the resolution of the image is low and it has a narrow field-of-view as compared to modern retinal fundus cameras and ophthalmoscopes and depicts a more severe case of plus disease. Put together, these issues have increased the subjectivity of plus disease diagnosis, thereby decreasing ROP-expert agreement on what constitutes plus disease. Ultimately, this has led to overtreatment in some children, and undertreatment in others, both of which can lead to lifelong visual impairment.[1,3] While ROP experts may not agree on diagnostic cut points for plus disease diagnosis, they do tend to rank vascular severity similarly.[22,23] That is, although there may be disagreement between whether or not an eye has plus disease, they can agree that the vasculature of one eye may be more severe than another. In this work, we aim to take advantage of this by generating personalized reference standard plus disease images for a given patient's eyes.

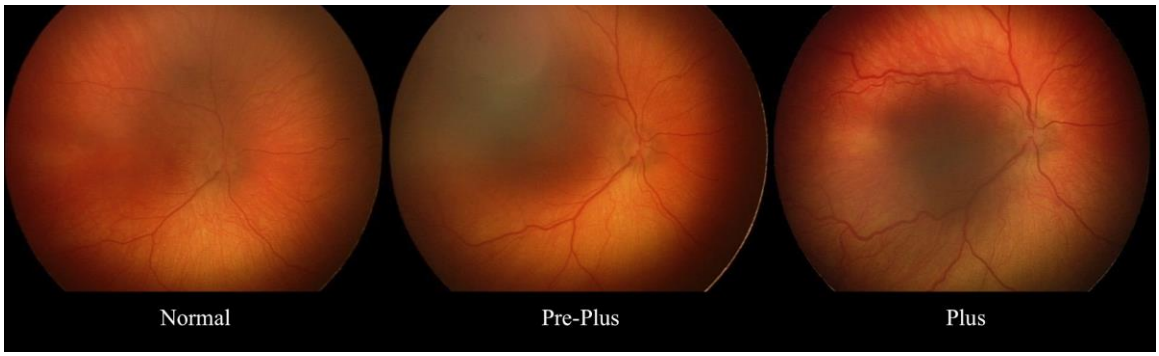

Figure 1: Example retinal fundus images of increasing vascular severity. From left to right, retinal fundus images of an eye that was originally diagnosed normal, developed pre-plus disease, and then plus disease. In plus disease images, retinal blood vessels are dilated and tortuous as compared to normal images. The degree of dilation and tortuosity of pre-plus blood vessels is less than that of plus disease blood vessels, but greater than normal.

Figure 2: Original fundus photograph depicting plus disease from the International Committee for the Retinopathy of Prematurity.

To do so, we use a series of generative adversarial networks (GANs). GANs have an uncanny ability to generate highly-realistic images, and are slowly being introduced into the field of medicine.[113,117] GANs contain both discriminative and generative convolutional neural networks

that are trained to deceive one another.[47,49,50] A discriminative network attempts to estimate an output given a set of inputs, whereas a generative network attempts to model the distribution of the input given an output.[47,49,50] To train these networks, data are supplied in pairs – inputs and corresponding output(s). The two models are pitted against one another, and as training progresses, the ability of each model improves (i.e., as the discriminator better learns to discern between real and generated images, the generator must also learn how to better simulate data). Ideally, this results in a generator that consistently fools a well-trained discriminator into classifying its outputs as real using highly realistic generated images.

There are two main types of GANs: (1) those that require paired image data and (2) those that do not.[106] Paired-image GANs require that both input and output images are structurally similar, but represent different modalities. Typically, this involves a real image and a labeled image. **Figure 3** best illustrates this concept, where labels are converted to scenes, or vice versa.[49] Not only do paired-image GANs produce higher-quality images than unpaired-image GANs, but they typically require fewer images for training. This is because paired data allows a GAN to home in on what the true differences are between two imaging modalities, and learn how it can best model the mapping between said modalities. However, collecting paired images can be difficult, if not impossible. In contrast, unpaired-image GANS do not have this limitation, but require more images to train and results are often of lesser quality.[49,50] Prime examples of this are displayed in **Figure 4** where, for instance, finding paired image data of zebras and horses posed in the same manner would be virtually impossible. Although the images are believable, close examination of the "horses" reveals that they are somewhat blurry and their faces are not sharp. Similarly, in the horse to zebra scene, the grass has been modified to be slightly browner — likely because zebras

are typically photographed in savannas and deserts, whereas horses are photographed in prairies and grasslands — even though this was not an intended outcome.

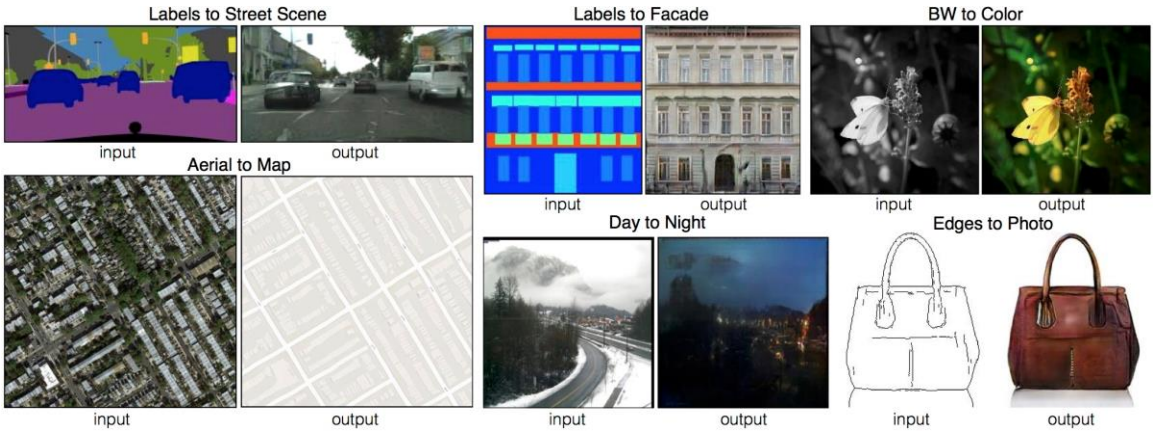

Figure 3: Example implementations of the pix2pix generative adversarial network. This model uses paired-image data to convert feature maps to real images (Labels to Street Scene, Labels to Facade, Edges to Photo), and real images to feature maps (Aerial to Map). The results are realistic and of relatively-high resolution. Figure adapted from Image-to-Image Translation with Conditional Adversarial Networks.[49]

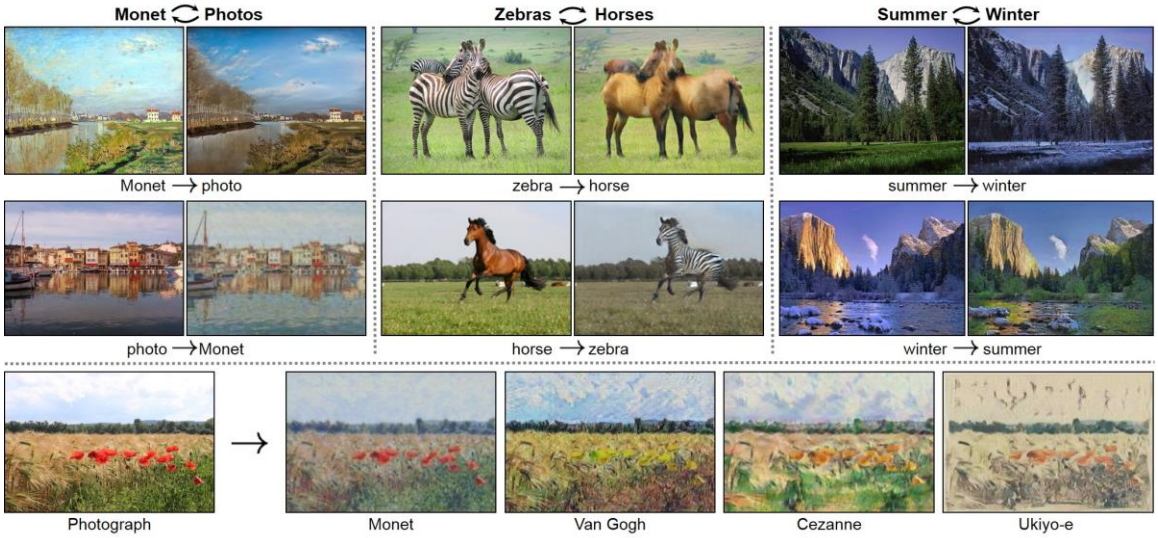

Figure 4: Example implementations of the CycleGAN generative adversarial network. This model converts the style of images from one to another using unpaired images. Figure adapted from Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.

This is because the task of transferring the style of color images is not trivial, due to image size and the vast array of colors (millions) that must be learned.[47,49,50,106] On the other hand, black and white or grayscale images are better suited for this task, as there are only 2 or 256 possible values per pixel, respectively. In this work, we use an unpaired-image GAN to modify grayscale retinal vessel maps, as they are theoretically easier for a model to learn and simulate. Specifically, we modify the vessel maps of normal or pre-plus disease eyes to appear as plus disease. The eventual goal is to then convert said vessel maps into retinal fundus images using a paired-image GAN. Ultimately, this allows for the synthesis of personalized reference standard images to which physicians may compare infants' eyes.

METHODS

Institutional Review Board

This study was approved by the Institutional Review Board at the coordinating center (OHSU) and at each of 8 study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Weill Cornell Medical Center, Asociación para Evitar la Ceguera en México [APEC]). This study was conducted in accordance with the Declaration of Helsinki. Written informed consent for the study was obtained from parents of all infants enrolled.

Retinal Fundus Image Dataset

As part of the multicenter ROP cohort study, i-ROP, over 30,000 nasal, temporal, inferior, superior, and posterior-pole retinal fundus images were collected from 970 preterm infants during routine ROP screening examinations. Between three and eight independent experts labeled each image set as normal, pre-plus, or plus disease, and an expert consensus diagnosis was formed, which established the ground truth diagnosis. Experts were all ophthalmologists with extensive experience in both ophthalmoscopic and image-based diagnosis of ROP. In addition to an ROP diagnosis, every posterior-pole image for a given subject was analyzed by DeepROP, an automated plus disease classifier. DeepROP provided a soft-max probability for each image as having normal, pre-plus, or plus disease vasculature, with all three values summing to one. From these values, a vascular severity score (VSS) ranging [1.0, 9.0] was created:

$$Vascular\ Severity\ Score = P(normal) + 5 * P(preplus) + 9 * P(plus).$$

Images of stage 4 (partial retinal detachment) and 5 (total retinal detachment) ROP were not included in the dataset, nor were images where image quality was deemed "not acceptable for diagnosis."

Model Setup and Training

Ten separate GANs were trained to modify the retinal vasculature in the following ways: normal to plus, pre-plus to plus, VSS 1 to VSS 2, VSS 1 to VSS 3, and so on to VSS 1 to VSS 9. Each model had unique datasets that were downsampled by plus disease diagnosis or VSS. The normal to plus and pre-plus to plus datasets were randomly split, 70/10/20, into train, validation, and test datasets, respectively. Because the number of training images for training VSS models was low, especially for the rarer VSS scores 7–9, the tuned parameters used for the normal to plus and pre-plus to plus models were used to train the VSS models, and generator learning was evaluated by comparing discriminator versus generator losses, and by manual inspection of generated images in the training dataset. A test dataset (20%) was used for final model evaluation. Additionally, because a subject may be represented in the dataset more than once (multiple imaging sessions and multiple image views), it was ensured that subjects were unique to either the train, validation, or test datasets. This by-subject split was common across all trained models, so that they could later be compared.

Models were built and trained in Python using PyTorch on an Nvidia V100 GPU (Santa Clara, CA).[109] For each image in the training, validation, and test datasets, vessel maps were generated using a previously-trained U-Net.[6,45] The open-source CycleGAN code was forked from a Github repository, hosted by its authors, and applied to the i-ROP training datasets.[50,118] Models were trained for 100 epochs at an initial learning rate of 0.0002, and linearly decayed to a learning rate

of 0 over another 100 epochs. During training, image pairs (e.g., normal and plus disease vessel maps) were input at a resolution of 640x480, scaled to 572x572, and randomly cropped to 512x512 in the same location for each image in the set. Discriminator and generator loss functions on both the training and validation test sets were monitored to ensure learning was occurring at an equal rate between objective functions, and that overfitting was not occurring. Manual inspection of the training set images occurred to verify that generated vessel maps were medically plausible.

Image Severity Verification

As mentioned, because GANs are an unsupervised learning method, there is theoretically not a "target" outcome. While visual inspection of generated images could confirm increased vessel dilation and tortuosity, whether the images were converted to plus disease remained a question, as they could have been converted to pre-plus disease. Therefore, we chose to evaluate the diagnoses of generated images from the validation and test datasets using DeepROP, which provided both a diagnosis and a VSS. After a GAN was successfully trained (discriminator and generator loss curves tracked one another) and visual inspection of generated training dataset images appeared appropriate, the vasculature of validation set images was modified. If the diagnoses of modified validation set images were accurate and no adjustments to the model were needed, the final model was evaluated on the test dataset.

RESULTS

Augmenting Vessel Severity to Plus Disease

During the first training pass of the normal to plus disease CycleGAN, it was found that the discriminator loss quickly approached zero while the generator loss did not, suggesting that the generator was not easily able to trick the discriminator. This was further confirmed by generating plus disease images from normal images using the validation dataset — these images were evaluated via DeepROP, which showed that many images were either unchanged, not medically-plausible or, at best, were converted to pre-plus (**Figure 5**).

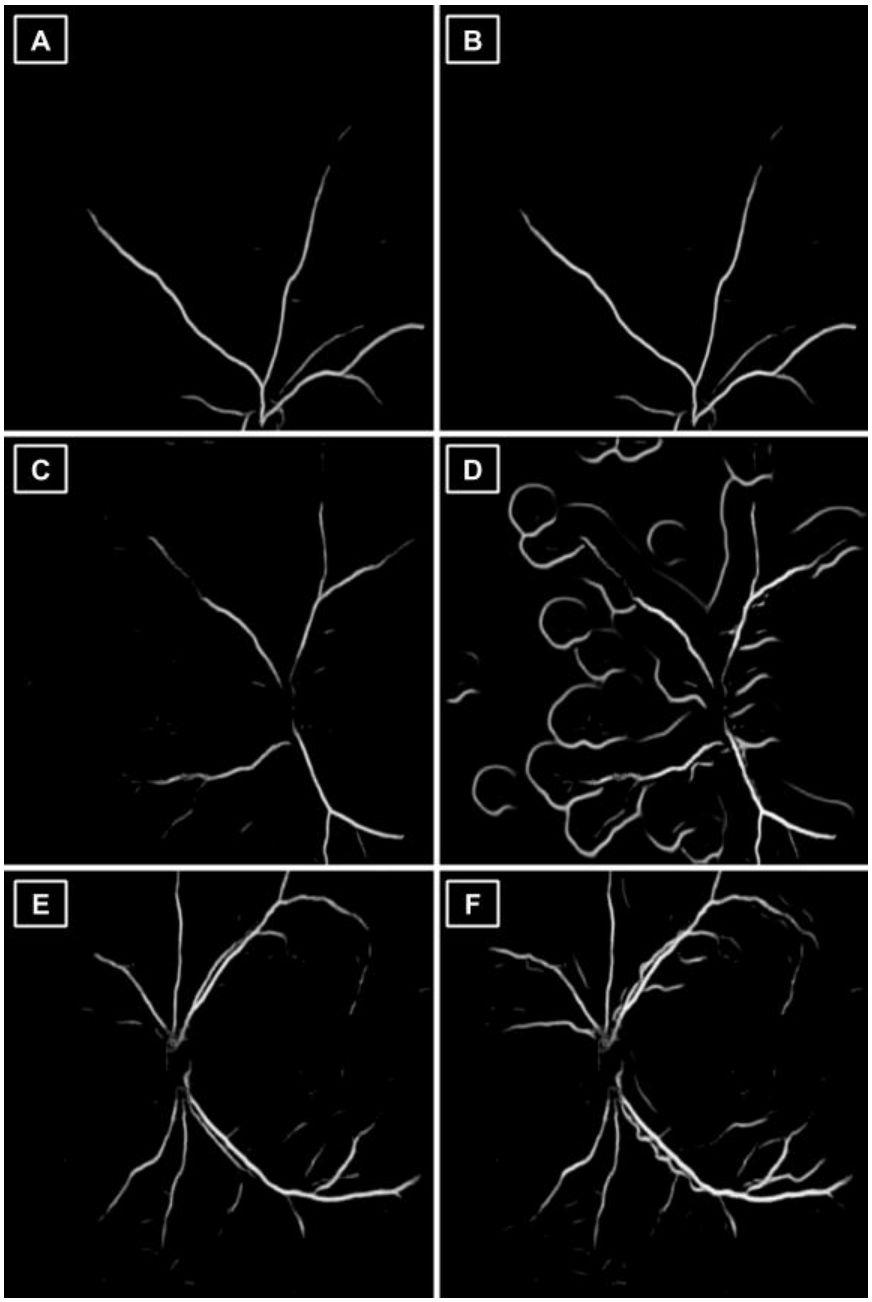

Figure 5: Examples of failed CycleGAN transformations. A, C, and E represent original vessel maps with normal retinal vasculature. CycleGAN either (B) failed to change the vessel map at all, (D) changed it in a way that was not medically plausible, or (F) only converted the image to pre-plus disease.

Thus, this model was fine-tuned by reducing the number of filters in the first layer of the discriminator from 64 to 32. This had the effect of reducing the capacity of the discriminator, thereby slowing its learning rate and allowing for the generator to learn how to better mimic the appearance of plus disease. After 200 epochs, the discriminator loss was 0.254 and the generator loss was 0.253, suggesting that the generator was likely able to trick the discriminator and that generated images were of the desired diagnosis. Using this model, generated images were input to pix2pixHD, and generated plus disease images from the validation dataset were evaluated via DeepROP. All 32 generated images were diagnosed as "plus" by DeepROP (**Figure 4**). To further confirm that this tuned model was performing as expected, the test dataset was also evaluated. Again, 100% of the 64 test dataset images were transformed from normal to plus (**Figure 5**). Manual inspection of both validation and test dataset images revealed medically plausible results (e.g., there were not any abnormal vasculature patterns that could lead to a plus disease diagnosis, such as loops or abnormal branching patterns).

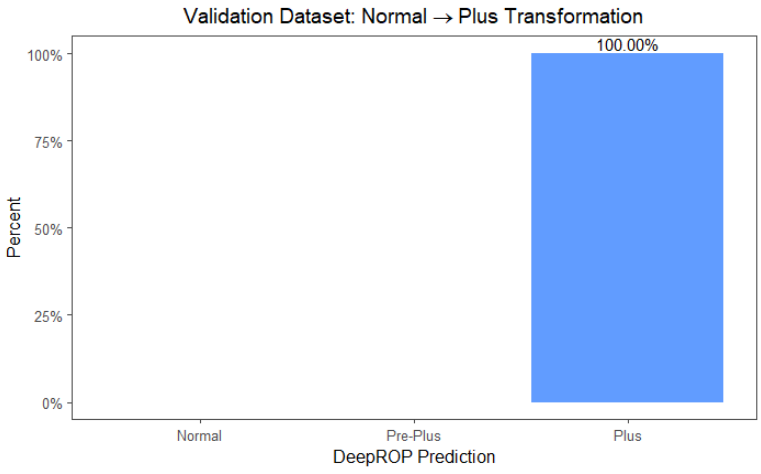

Figure 4: Validation dataset results for the normal to plus disease CycleGAN. All 32 validation dataset images were successfully converted from normal to plus disease vasculature.

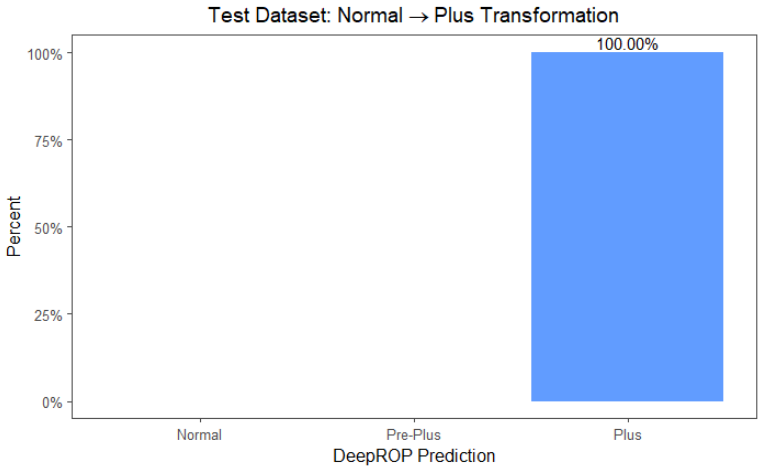

Figure 5: Test dataset results for the normal to plus disease CycleGAN. All 64 test dataset images were successfully converted from normal to plus disease vasculature.

Similar to the normal to plus disease CycleGAN, it was found that the discriminator loss of the pre-plus to plus disease CycleGAN quickly approached zero while the generator loss did not, suggesting that the generator was not able to easily trick the discriminator. This was again confirmed by generating plus disease images from pre-plus images in the validation dataset. Generated images were not evaluated by DeepROP, as they were not medically plausible.

Therefore, this model was also tuned by reducing the number of filters in the first layer of the discriminator from 64 to 32. After 200 epochs, the discriminator loss was 0.248 and the generator loss was 0.250, suggesting that the generator was able to trick the discriminator. Using this model, generated pre-plus disease images from the validation dataset were evaluated via DeepROP. All 32 generated images were diagnosed as "pre-plus" by DeepROP (**Figure 6**). This desired model behavior was confirmed using the test dataset. 94% of the 64 test dataset images were transformed from pre-plus to plus — four images were not diagnosed as plus (**Figure 7**). Manual inspection of both validation and test dataset images revealed medically plausible results.

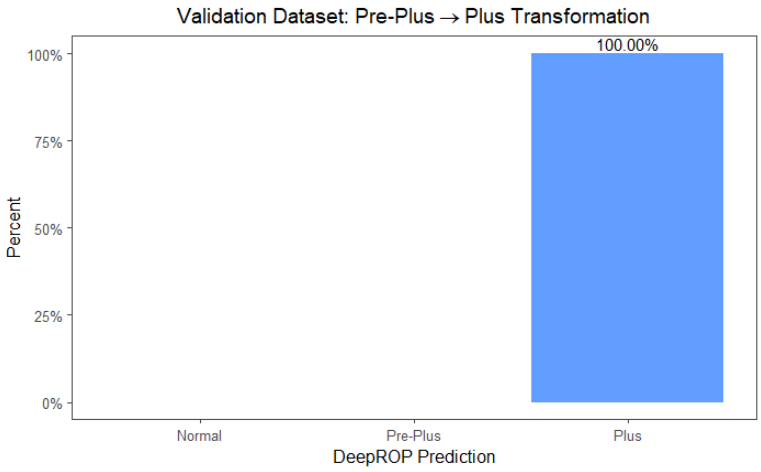

Figure 6: Validation dataset results for the pre-plus to plus disease CycleGAN. All 32 validation dataset images were successfully converted from pre-plus to plus disease.

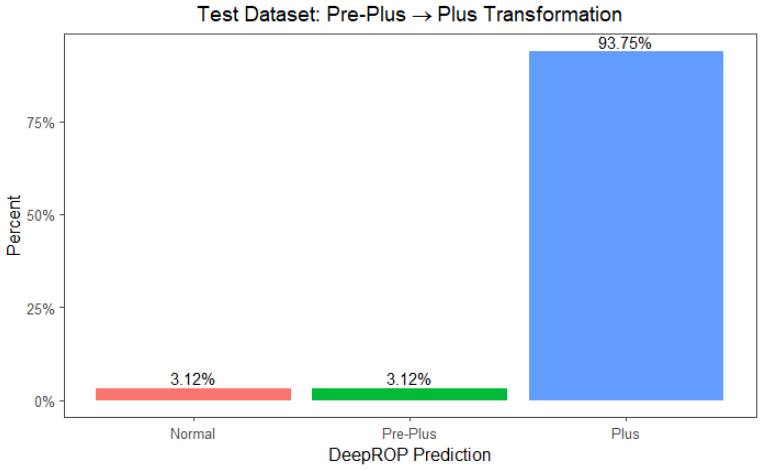

Figure 7: Test dataset results for the pre-plus to plus disease CycleGAN. Most test dataset images were successfully converted from pre-plus to plus disease vasculature. A total of four out of the 64 images were not diagnosed as plus disease by DeepROP.

A subset of normal eyes and pre-plus eyes that actually developed plus disease over time were compared with GAN predictions of plus disease in those specific eyes. In general, the predicted retinal vascular trees were not exact matches to the actual retinal vascular trees, however there were similarities. An example is presented in **Figure 8**. Here, the model did not predict that vitreous haze would occlude portions of the retina. Although, it did predict that the retinal vasculature would dilate, and that tortuosity would increase in the areas depicted by white arrows. However, the overall severity of the predicted plus disease appeared to be slightly increased over the level of plus disease in the real image. In practice, this could be problematic, as the overall severity of an eye for the diagnosis of plus disease has yet to be fully agreed upon by ROP experts. Therefore, in addition to increasing vessel severity from normal or pre-plus to plus to plus, we also investigated increasing vessel severity in a more granular fashion, using the novel VSS.

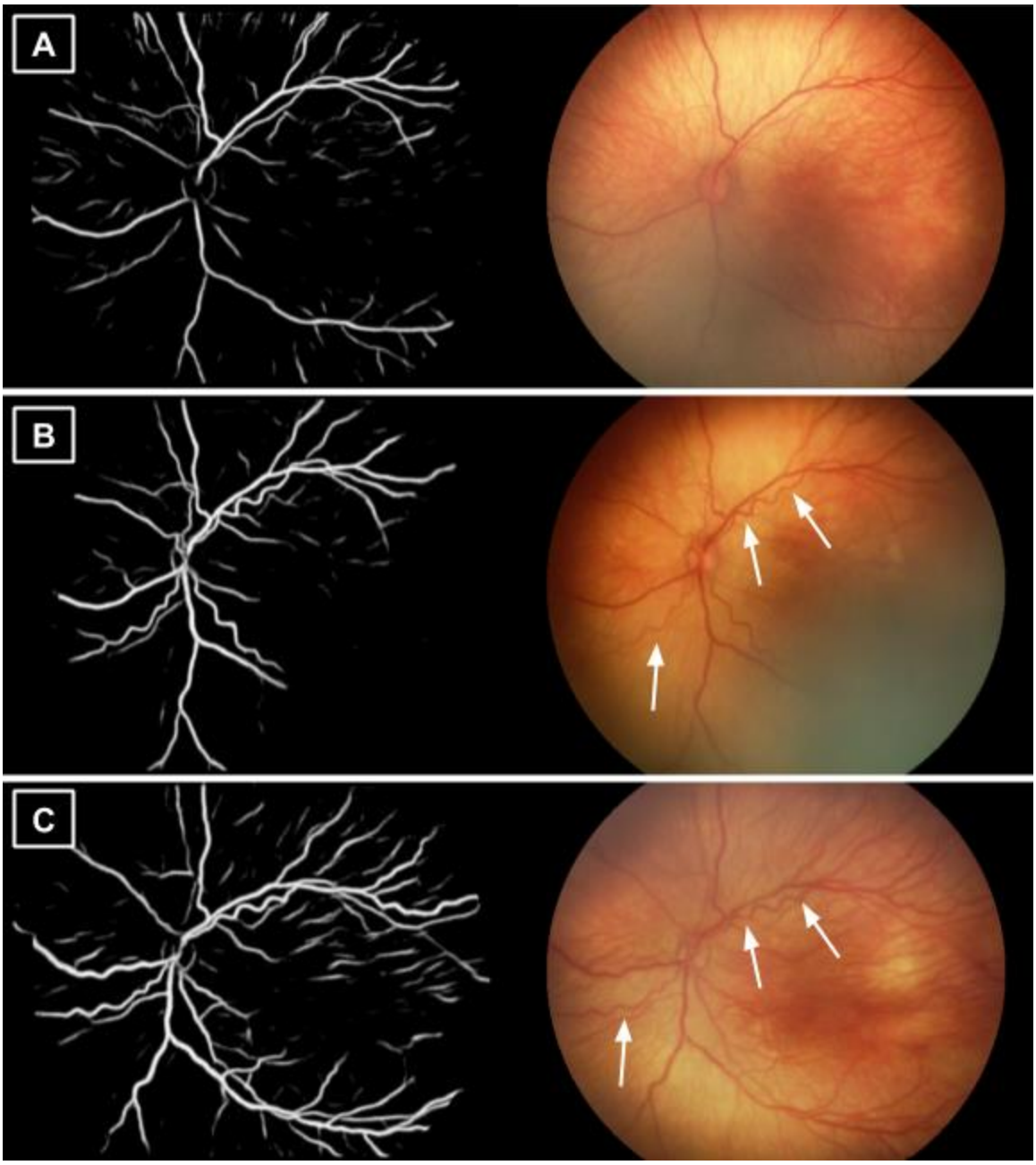

Figure 8: Example of normal vasculature transformed to plus disease vasculature. (A) An eye with normal vasculature that eventually developed (B) plus disease. (C) A prediction of how plus disease would appear in the eye, given the image in A.

Augmenting Vessel Severity Along the Vascular Severity Score

In this set of experiments, the goal was to transform images in the i-ROP dataset with a VSS of 1 (the most commonly presented VSS in the i-ROP dataset) to VSS 2–9 (**Figure 8**). Because the groups are more granular (i.e., there are nine VSS designations versus three plus disease diagnoses) and adequately-sized train, validation, and test datasets were not possible, we opted to forgo the extra validation dataset and use knowledge gained from previous experiments (i.e., set the number of filters used in the first layer of the discriminator to 32, train for 100 epochs with a static learning rate and linearly decay to 0 over another 100 epochs). The cycle consistency loss was monitored as a measure of over-/under-fitting, and generated training dataset images were visually inspected for plausibility. The models were then evaluated on their respective test datasets.

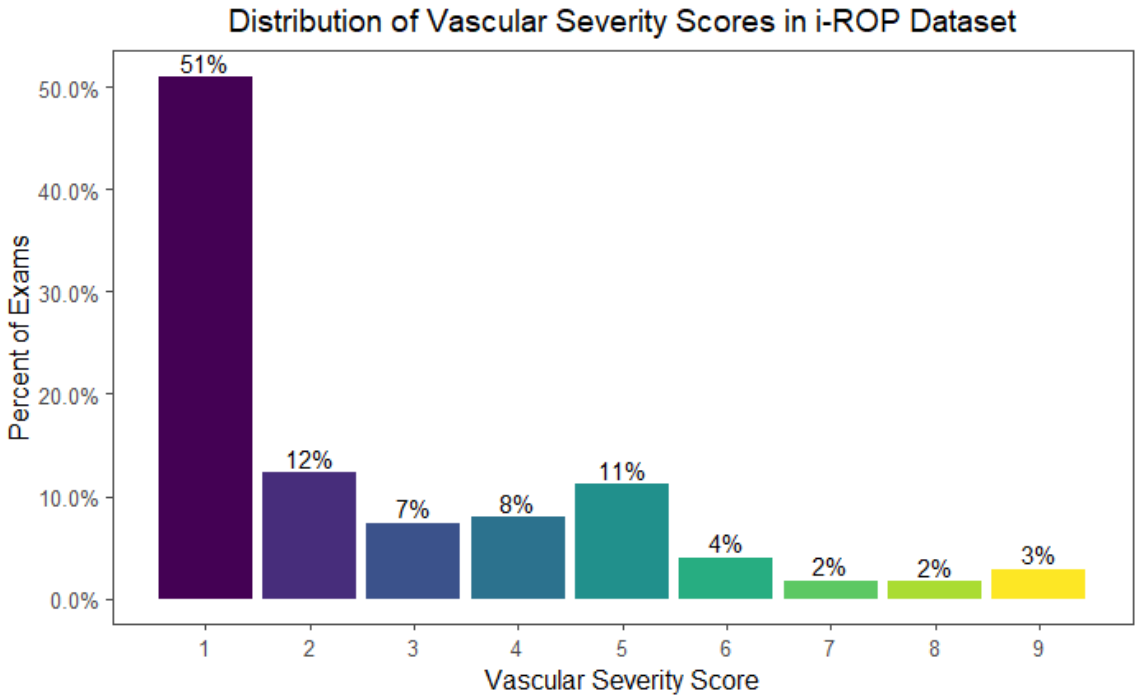

Figure 8: Distribution of the vascular severity score in the i-ROP Dataset. As is common with normal versus pre-plus versus plus, the majority of images are VSS 1, meaning they are normal. Around 7% of images are VSS 7–9, in line with the percentage of images that have plus disease.

After 200 epochs, discriminator and generator losses were compared to one another — models did not appear to be severely over- or under-fitting (**Table 1**). However, the discriminators in models VSS 1→7, VSS 1→8, and VSS 1→9 all appeared to be outperforming their respective generators. The number of filters in the first layers of these discriminators was reduced from 32 to 16 in an attempt to slow the learning of the discriminator as compared to the generator, but results did not improve. This suggests that this disparity between discriminator and generator losses is due to a data limitation, which aligns with the VSS distribution in the i-ROP dataset (**Figure 8**).

Table 1: Discriminator and generator losses for each of the VSS models.

Model	Discriminator Loss	Generator Loss
VSS 1→2	0.253	0.266
VSS 1→3	0.244	0.262
VSS 1→4	0.264	0.241
VSS 1→5	0.216	0.231
VSS 1→6	0.224	0.298
VSS 1→7	0.191	0.368
VSS 1→8	0.140	0.375
VSS 1→9	0.206	0.365

Using the best-trained models for each VSS transformation, test dataset images were transformed from VSS 1 to the desired VSS for each model (**Figure 9**). Although the models did not perform perfectly, especially for VSS 7 and VSS 8, there appears to be a gradual increase in VSS for each image ranging from desired VSS 2 to desired VSS 9 (**Figure 9, Table 2**). This was confirmed via a one-way analysis of variance ($p < 2e^{-16}$); a post-hoc pairwise t-test with Bonferroni correction for multiple comparisons is presented in **Table 3**. Statistically dissimilar VSSs were not produced by the following models: VSS 1→2 and VSS 1→3, VSS 1→4 and VSS 1→5, VSS 1→6 and VSS 1→7, VSS 1→6 and VSS 1→7, and VSS 1→7 and VSS 1→8.

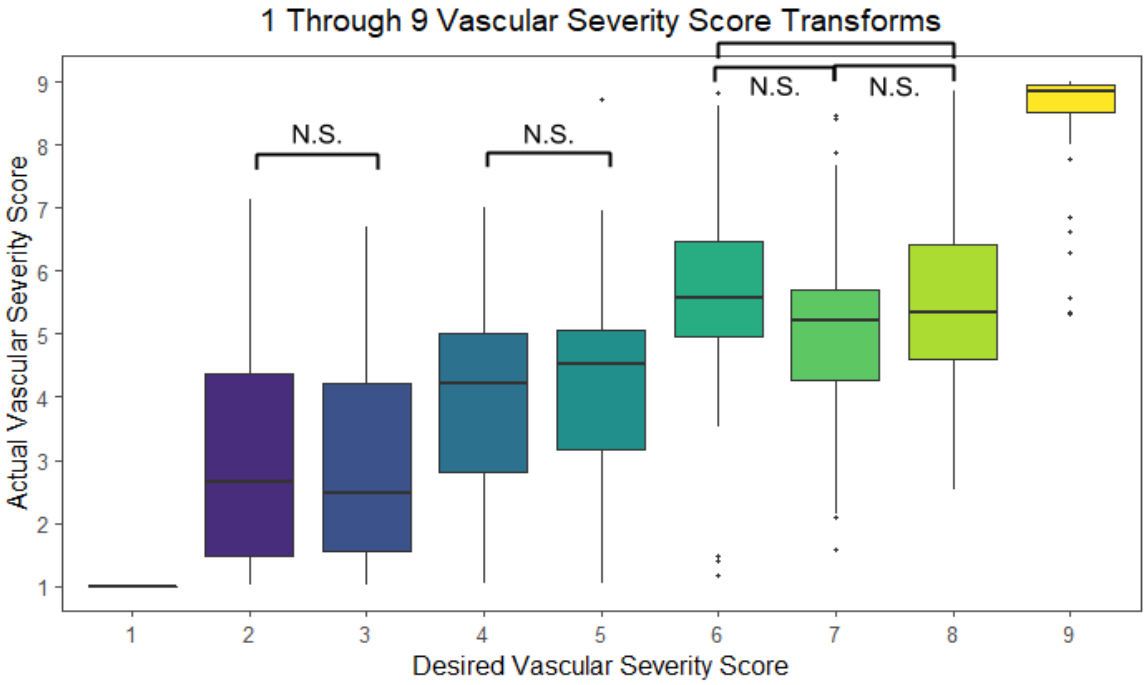

Figure 9: Actual DeepROP vascular severity scores for test set images versus the desired transformation. N.S. indicates models that did not produce images with significantly different vascular severity scores.

Table 2: Mean ± standard deviation VSS of each CycleGAN model.

Model	Mean ± SD
VSS 1→2	3.0 ± 1.6
VSS 1→3	2.9 ± 1.5
VSS 1→4	3.8 ± 1.5
VSS 1→5	4.1 ± 1.4
VSS 1→6	5.6 ± 1.4
VSS 1→7	5.1 ± 1.7
VSS 1→8	5.5 ± 1.5
VSS 1→9	8.4 ± 0.9

DISCUSSION

This study aimed to simulate plus disease in retinal fundus images collected from preterm infants who did not, at the time of imaging, have plus disease. The overall goal was to generate personalized reference standard images for plus disease for individual subjects. To accomplish this, retinal fundus images were segmented into retinal vessel maps, where the vascular patterns were augmented, and then back-converted into retinal fundus images using a previously trained GAN. There are two key findings: (1) the severity of retinal blood vessels in retinal vessel maps can be augmented to appear as plus disease and (2) the severity of plus disease can be augmented in a granular fashion along the scale of a novel 1–9 vascular severity score.

This work has many important implications. First, it allows for generation of personalized reference standard images for individual patients. This is important, as the current reference standard image is outdated — it is blurry and has a narrow field-of-view (**Figure 2**)[1,3] Additionally, the degree of plus disease presented in the image is severe compared to the consensus treatment-requiring plus disease presented in **Figure 1**. The plus disease images used in this study were all diagnosed by ROP experts with decades of experience from around the United States, who came to a consensus diagnosis for each image. This arguably better represents the true degree of plus disease that warrants treatment. However, it is expected that the ICROP will be updated relatively soon, and that the issues associated with the image of plus will be addressed. Therefore, we also felt it was necessary to produce images that exhibited increasingly worse vascular patterns. In this way, the vasculature can be increased to a desired level of plus disease. For example, it may be that the updated ICROP will suggest that VSS 9 is beyond the severity level required for treatment, and that VSS 8 is sufficient. The model can then be amended to produce images of VSS 8, rather than just “plus disease.”

Another interesting potential application of these methods centers around synthetic datasets. In many fields of research, it is often the case that data is hard to obtain, either due to practical limitations, privacy concerns, or disease rarity. For numeric data, new cases can often be simulated using machine learning algorithms. However, until now, image data has proven difficult to simulate realistically [47,106]. In this study, we have shown that we can increase the vascular severity of retinal fundus images in a manner that is indistinguishable from real images by ophthalmic experts. This is important as it may allow for the training of machine learning methods on datasets of limited size by generating synthetic images from real images. Then, real images could be used for testing to evaluate real-world performance.

Although the normal to plus disease, pre-plus to plus disease, and VSS 1–9 models performed well according to DeepROP, there are still limitations to these models. First, it cannot be guaranteed that all images will be transformed to the desired VSS. This was likely due to dataset size limitations[47] However, it's possible that more fine-tuning would allow for the VSS models to better learn the desired severities. That said, even with a larger amount of data, there will still be images that are not properly converted, as was witnessed with the pre-plus to plus disease test set. Second, these models do not predict exactly how plus disease will present in a given patient. This was demonstrated in **Figure 8**. Although there are similarities between true and predicted vascular patterns that should alert physicians to the presence of plus disease, there are also differences. Some of the vasculature may not be visible, or more or less tortuous. However, the overall severity should be similar. Therefore, an examiner would need to keep in mind that predicted images are reference standards that better represent an individual, but are not exact predictions.

CONCLUSION

In this work, we have trained multiple generative adversarial networks to increase the severity of the retinal vasculature to better mimic the appearance of plus disease in retinopathy of prematurity. Future work will aim to better simulate VSSs. Ideally, we will be able to train GANs that can accurately model various VSSs that are increasing in a statistically different manner. The ultimate goals will be to (A) implement this method for better reference standards for physicians, especially non-ROP experts who examine infants for ROP in rural areas and (B) to generate synthetic datasets so that image-based risk models may be trained from them. Overall, these models produce highly realistic images that are diagnosed as plus disease by DeepROP, a plus disease screening tool currently in the FDA approval process. It is our hope that these simulated images can provide personalized reference standard images to assist ROP examiners with plus disease diagnoses.

DISCUSSION

In this work, we have developed multiple algorithms to assist with accurate and reliable diagnosis and prediction of treatment-requiring (TR-) retinopathy of prematurity (ROP). The first algorithm, a convolutional neural network (CNN), was able to detect acceptable quality retinal fundus images — those from which accurate and reliable ROP diagnoses can be formed — from those which were not. The second algorithm, ElasticNet regularized logistic regression, used a combination of clinical factors and a CNN-derived vascular severity score to predict all infants who were at risk of eventually developing TR-ROP, while accurately ruling out more than half of those who would not. The final algorithms were generative adversarial networks (GANs). The first set of GANs, cycleGANs, operate on retinal vessel maps to augment the severity of the retinal vasculature to appear as plus disease or to incrementally augment it along a novel 1–9 vascular severity scale. Another GAN, pix2pixHD, converts the augmented retinal vessel maps into realistic retinal fundus images. These images can then be used as personalized reference standard images to assist non-experts with monitoring ROP in children who are not predicted to develop TR-ROP. There are three key findings that arose from this study: (1) the quality of retinal fundus images can be quickly and accurately detected by a CNN, (2) a parsimonious logistic regression model that uses a vascular severity score can detect eyes that are likely to develop TR-ROP, and (3) GANs can accurately augment the appearance of plus disease in retinal vessel maps and generate realistic retinal fundus images from retinal vessel maps. These models have potential to be used alone or in conjunction and have attempted to address the issues in the three previously outlined specific aims.

Aim 1: Quality Control for Retinal Fundus Images. We will ensure that high quality retinal fundus images are used for both telemedicine and automated diagnosis of ROP by training a convolutional neural network to detect images that are acceptable for the diagnosis of ROP from those which are not.

During the training phase for the image quality model, the mean (SD) area under the receiver operating characteristics curve (AUROC) of five-fold cross-validation was 0.958 (0.005), which suggested that the individual models were performing well and that they were not overfit to the training data. Evaluation on a held-out test dataset confirmed this (AUROC=0.965). These AUROCs suggested that this model would have high discriminatory power along many decision thresholds for image quality. To test this, experts ranked a subset of 30 test set images from worst to best quality. The CNN's rank was established by ordering the probabilities of images being of acceptable quality; it was highly correlated with the expert consensus rank (correlation coefficient: 0.90). Given these results, there are two conclusions that can be made: (1) the model was able to distinguish between images of acceptable quality and images of low or questionable quality with a high degree of confidence, and (2) because the model could rank image quality similarly to experts, the decision threshold can be altered based upon the application to which this algorithm is applied.

As mentioned, a major limitation to image-based ROP diagnosis, whether via telemedicine or automated methods, is the lack of sufficient quality control.[7,9,21] Standard image quality metrics, such as the peak signal to noise ratio (PSNR), do not capture some of the nuances of ROP.[72,80] For example, vitreous haze can be a symptom of severe ROP, which can cause occlusions [1]. However, not all occlusions are detrimental to an accurate ROP diagnosis. Unlike the PSNR, this CNN presumably identifies not only vitreous haze, but the degree and extent of occlusion, and its

location relative to other retinal features, such as the optic disc or major retinal blood vessels. Other methods have attempted to subdivide retinal fundus images into smaller blocks that can be examined alone, and an overall score produced.[72] While these methods are better at judging the degree and extent of occlusions than PSNR, they still fail to take into account where in the image it occurs (i.e., small occlusions in the periphery are far less damaging to an accurate ROP diagnosis than large occlusions near the center of an image).

This model has a few potential applications. The image quality labels used to train this model ("Acceptable", "Not Acceptable") were determined via consensus agreement by six ROP experts from across the United States, each of which had extensive experience in image-based diagnosis of ROP. Therefore, because this model had such high agreement with experts' determination of image quality, it could be used, as-is, in telemedicine pipelines. It could also be used by automated methods for the diagnosis of ROP; however, the decision threshold at which images are classified as "Acceptable" or "Not Acceptable" may need to be adjusted, as it is possible that automated methods may need slightly higher quality images, or may even be able to use lower quality images. While this method addresses issues specific to ROP, it should be noted that it certainly has potential for application in other ophthalmic diseases that are evaluated via retinal fundus images. It is also likely that a CNN can be trained to determine image quality for a variety of other imaging modalities and non-ophthalmic diseases.

Aim 2: Prediction of Treatment-Requiring ROP Patients. We will develop a risk model for the prediction of TR-ROP. Data suggests that the severity of the retinal vasculature may be an early and significant predictor of TR-ROP. The goal will be to have 100% sensitivity, with specificity above 50%, in order to reduce the number of subjects needing to be screened by more than half.

In this aim, we successfully demonstrated that just two features could accurately predict, more than one month in advance, eyes that would develop TR-ROP, while correctly ruling out more than half of those that would not. Using the training dataset, we first performed a correlation analysis in order to discover which clinical features were tied to the development of TR-ROP. After dropping features with low correlation values or low representation in the dataset, we used five-fold cross-validation to train and tune multiple ElasticNet logistic regression models on all possible combinations of the remaining features, namely: birthweight (BW), gestational age (GA), and a deep learning-based vascular severity score (VSS). All infants were screened between 32 and 34 weeks PMA. We found that a combination of GA and VSS produced the model with the highest AUPR. We tuned this model's decision threshold, via five-fold cross-validation, using the F_2 score. On a held-out test dataset and on an independent test dataset, the model had 100% sensitivity and specificity equal to 55% and 68%, respectively. There are two key takeaways from this work: (1) VSS, evaluated at 32–33 weeks PMA, is correlated with TR-ROP, and (2) a risk model that uses VSS and GA as predictors is not only highly sensitive, but also specific.

These are important findings, as TR-ROP risk models with 100% sensitivity and high specificity are rare.[10,18,19] Those that have demonstrated 100% sensitivity and high specificity, when evaluated on more diverse datasets, such as the i-ROP dataset, often begin to fail.[11] For example, the Children's Hospital of Philadelphia (CHOP) developed what is arguably the best-performing ROP risk model, to date. However, it was only trained on infants that were admitted to a single hospital

in Philadelphia, PA.[10] When the CHOP ROP model was evaluated on a larger and more diverse set of patients from all across North America, the specificity of the model had to be reduced to 6% in order to achieve 100% sensitivity.[11]

Our model, however, was trained and evaluated on a large, diverse dataset, and further evaluated on an entirely independent test dataset. In both cases, sensitivity was 100% and specificity was greater than 50%, suggesting that this model generalizes well to larger North American populations. If performance were to remain consistent, this means that more than half of the infants screened for ROP could either be discharged or screened far less frequently. This would effectively reduce the screening burden, which equates to more time and attention spent evaluating and treating those who are predicted to be at a higher risk of developing TR-ROP, while simultaneously reducing the physiological stress placed on those who are predicted to be at low or no risk of developing TR-ROP. Ultimately, these results, coupled with the fact that this model is extremely parsimonious — it consists of only two, easy-to-obtain features that are input to a highly-interpretable logistic regression model — suggest that it could quickly and easily be applied, clinically.

Aim 3: Development of Personalized Reference Standard Images. We will synthesize personalized reference standard images to assist non-experts with identification of TR-ROP. To accomplish this, we will use a series of generative adversarial networks. The first networks will augment the retinal vasculature present in retinal vessel maps (generated from retinal fundus images) to appear as plus disease (a significant indicator of TR-ROP) or to increase the vascular severity incrementally. The last network will be used to transform the augmented retinal vessel maps into realistic retinal fundus images.

The overall aim of this work was to create personalized reference standard images of plus disease; it consisted of two parts. The first part was to convert retinal vessel maps into highly realistic retinal fundus images. The second was to augment the vascular severity of normal and pre-plus disease retinal vessel maps — segmentations of the retinal vasculature produced by a previously trained U-net — to appear as plus disease. Similarly, we also wished to incrementally increase the vascular severity of vessel maps along the novel 1–9 VSS. When put together, we can segment a given patient's retinal fundus image into a retinal vessel map, modify said vessel map to appear as plus disease or any severity along the 1–9 VSS scale, then convert the image back into a retinal fundus image so that it may assist a physician during the screening and diagnosis of ROP.

In the first set of experiments, we found that we could transform retinal fundus images into retinal vessel maps, and then convert them back into realistic retinal fundus images using a GAN. To further ensure that images were realistic and gradable, a set of physicians diagnosed all real and synthetic retinal fundus. Inter- and intra-expert diagnoses were nearly identical. This suggested that data was not lost during the conversions of retinal fundus images into retinal vessel maps, and then back. However, physicians were able to identify which images were real and which were synthesized. This was likely due to the capacity of the GAN used, which could only generate

images of size 256x256x3. When these images were upsampled to the common retinal fundus image size of 640x480x3, they appeared slightly blurry and pixelated. A follow-up experiment was performed some months later using a GAN with a much larger capacity — it could produce realistic images much larger than 640x480x3. We then found that a new set of ROP experts, when asked to identify real images from synthesized, could not do so. That is, the synthesized images were so realistic looking, that not even experts with decades of experience could identify them.

In the second set of experiments, we found that CycleGAN converted 100% of images in the validation and test datasets from normal to plus disease, as determined by i-ROP DL. Furthermore, the converted images were medically plausible. This is an important note, since, prior to tuning, the GAN was simply attempting to create circular patterns in retinal vessel maps that would undoubtedly be diagnosed as plus disease by i-ROP DL, but are not realistic. 100% and 93.75% of validation and test dataset images, respectively, for the conversion of pre-plus to plus disease, were successful and medically plausible. CyleGAN operates by first determining whether an image is already of the desired class, and then converting it if it is not. The “converted” pre-plus images that were not diagnosed as pre-plus by i-ROP DL were not actually modified at all. This is, presumably, because CycleGAN evaluated the images and determined that they already represented plus disease. Whether CycleGAN or i-ROP DL is correct needs to be determined by trained ROP experts. Unfortunately, for this set of experiments, experts did not have time to evaluate the accuracy of thousands of synthetic retinal fundus images. Because the majority of infants, upon their first ROP screening examination, have a VSS of 1, we also attempted to convert images of VSS 1 to the eight remaining VSSs. Overall, there was an increasing trend of vascular severity. However, what was likely due to a lack of data given the high similarity between VSSs, images converted to VSS 2 and VSS 3 did not have statistically dissimilar VSSs. The same

occurred for images converted to VSS 4 and VSS 5, as well as VSS 6 and VSS 7, VSS 6 and VSS 8, and VSS 7 and VSS 8. However, all images were within a couple VSSs of the desired VSS.

Taken together, we can segment the retinal vasculature of subjects' retinal fundus images into retinal vessel maps. We can then increase the severity of the segmented vasculature in a stepwise fashion or simply convert it to what is considered "plus disease" vasculature. From there, the vessel map can be converted into a highly realistic retinal fundus image. Overall, this method can help create personalized retinal fundus images of plus disease that have better quality and a wider field-of-view than the current, yet outdated, reference standard image of plus disease.

SUMMARY AND CONCLUSIONS

As previously mentioned, the increasing incidence of ROP — attributed to the ability to preserve the lives of younger, smaller infants — and the shortage of ROP experts has increased the prevalence of ROP-related visual loss. This has, unfortunately, resulted in ROP becoming a leading cause of childhood blindness in both developed and developing countries. It was our goal to solve some of the issues associated with the accurate and timely diagnosis of TR-ROP. First, a robust image quality algorithm was developed. This algorithm quickly alerts clinicians and researchers as to whether images are of high enough quality for the diagnosis of ROP. It can (and currently does) have applications in both telemedicine pipelines and automated methods for the diagnosis of ROP. Second, a risk model that can identify all subjects who will develop TR-ROP and correctly rule out more than half of the subjects in a screening pool who will never develop TR-ROP was developed. This model can significantly reduce the screening burden and the physiological stress placed on low-risk infants, allowing experts to prioritize those who are most at-risk of developing TR-ROP. Finally, to further reduce the screening burden, we proposed that non-ROP experts in developing countries assist with screening low-risk infants. To do so, we developed a model that can produce personalized reference standard images of plus disease, the most prominent indicator of the need for treatment of ROP. Using these images, non-experts can feel more confident in examining low-risk children, and can do so far less frequently than telemedicine or automated methods require, thereby reducing the ROP expert screening burden and the frequency of physiological stress placed on premature infants. Ultimately, we have addressed some of the major issues associated with ROP care, and hope for their implementation.

REFERENCES

1. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. *Arch Ophthalmol*. 2005;123(7):991-999.

2. Good WV, Hardy RJ, Dobson V, et al. The incidence and course of retinopathy of prematurity: findings from the early treatment for retinopathy of prematurity study. *Pediatrics*. 2005;116(1):15-23.

3. International Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. The Committee for the Classification of Retinopathy of Prematurity. *Arch Ophthal*. 1984;102(8). doi:10.1001/archopht.1984.01040030908011

4. Braverman RS, Enzenauer RW. Socioeconomics of retinopathy of prematurity care in the United States. *Am Orthopt J*. 2013; 63:92-96.

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

6. 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.

7. Richter GM, Williams SL, Starren J, Flynn JT, Chiang MF. Telemedicine for retinopathy of prematurity diagnosis: evaluation and challenges. *Surv Ophthalmol*. 2009;54(6):671-685.

8. Quinn GE, Ying GS, Daniel E, et al. Validity of a telemedicine system for the evaluation of acute-phase retinopathy of prematurity. *JAMA Ophthalmol*. 2014;132(10). doi:10.1001/jamaophthalmol.2014.1604

9. Chiang MF, Wang L, Busuioc M, et al. Telemedical retinopathy of prematurity diagnosis: accuracy, reliability, and image quality. *Arch Ophthal*. 2007;125(11). doi:10.1001/archopht.125.11.1531

10. Binenbaum G, Ying GS, Quinn GE, et al. The CHOP postnatal weight gain, birth weight, and gestational age retinopathy of prematurity risk model. *Arch Ophthal*. 2012;130(12). doi:10.1001/archophthalmol.2012.2524

11. Binenbaum G, Ying GS, Tomlinson LA. Validation of the Children's Hospital of Philadelphia Retinopathy of Prematurity (CHOP ROP) Model. *JAMA Ophthalmol*. 2017;135(8). doi:10.1001/jamaophthalmol.2017.2295

12. Hurt KJ, Guile MW, Bienstock JL, Fox HE, Wallach EE. *The Johns Hopkins Manual of Gynecology and Obstetrics*. Lippincott Williams & Wilkins; 2012.

13. Early Treatment For Retinopathy Of Prematurity Cooperative Group. Revised indications for the treatment of retinopathy of prematurity: results of the early treatment for retinopathy of prematurity randomized trial. *Arch Ophthal*. 2003;121(12). doi:10.1001/archopht.121.12.1684

14. Lawn JE, Davidge R, Paul VK, et al. Born too soon: care for the preterm baby. *Reprod Health*. 2013;10 Suppl 1:S5.

15. Blencowe H, Cousens S, Chou D, et al. Born Too Soon: The global epidemiology of 15 million preterm births. *Reproductive Health*. 2013;10(Suppl 1):S2. doi:10.1186/1742-4755-10-s1-s2

16. Fierson WM, American Academy of Pediatrics Section on Ophthalmology, American Academy of Ophthalmology, American Association for Pediatric Ophthalmology and Strabismus, American Association of Certified Orthoptists. Screening Examination of Premature Infants for Retinopathy of Prematurity. *Pediatrics*. 2018;142(6). doi:10.1542/peds.2018-3061

17. Fierson WM. Screening examination of premature infants for retinopathy of prematurity. *Pediatrics*. 2013;131(1). doi:10.1542/peds.2012-2996

18. Blencowe H, Lawn JE, Vazquez T, Fielder A, Gilbert C. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. *Pediatr Res*. 2013;74 Suppl 1:35-49.

19. Kim SJ, Port AD, Swan R, Campbell JP, Chan RVP, Chiang MF. Retinopathy of prematurity: a review of risk factors and their clinical significance. *Surv Ophthalmol*. 2018;63(5):618-637.

20. Campbell JP, Kim SJ, Brown JM, et al. Evaluation of a novel retinopathy of prematurity severity scale applied by clinicians and deep learning. *Ophthalmology*. Published online October 26, 2020. doi:10.1016/j.ophtha.2020.10.025

21. Briggs R, Bailey JE, Eddy C, Sun I. A methodologic issue for ophthalmic telemedicine: image quality and its effect on diagnostic accuracy and confidence. *J Am Optom Assoc*. 1998;69(9). Accessed October 14, 2020. https://pubmed.ncbi.nlm.nih.gov/9785735/

22. Campbell JP, Kalpathy-Cramer J, Erdogmus D, et al. Plus Disease in Retinopathy of Prematurity: A Continuous Spectrum of Vascular Abnormality as a Basis of Diagnostic Variability. *Ophthalmology*. 2016;123(11):2338-2344.

23. Kalpathy-Cramer J, Campbell JP, Erdogmus D, et al. Plus Disease in Retinopathy of Prematurity: Improving Diagnosis by Ranking Disease Severity and Using Quantitative Image Analysis. *Ophthalmology*. 2016;123(11):2345-2351.

24. 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.

25. Coyner AS, Swan R, Campbell JP, et al. Automated Fundus Image Quality Assessment in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. *Ophthalmol Retina*. 2019;3(5):444-450.

26. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. Published online November 14, 2017. Accessed October 13, 2020. http://arxiv.org/abs/1711.05225

27. Jones LD, Golan D, Hanna SA, Ramachandran M. Artificial intelligence, machine learning and the evolution of healthcare. *Bone & Joint Research*. 2018;7(3):223-225. doi:10.1302/2046-3758.73.bjr-2017-0147.r1

28. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342-1350.

29. Moor J. The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *AI Magazine*. 2006;27(4):87-87.

30. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With*

Applications in R. Springer; 2014.

31. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media; 2013.

32. Salian I. NVIDIA Blog: What's the Difference Between Supervised & Unsupervised Learning? Published August 2, 2018. Accessed October 13, 2020. https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/

33. Molnar C. Interpretable Machine Learning. Published 2019. Accessed October 13, 2020. https://christophm.github.io/interpretable-ml-book/

34. Bogunovic H, Waldstein SM, Schlegl T, et al. Prediction of Anti-VEGF Treatment Requirements in Neovascular AMD Using a Machine Learning Approach. *Investigative Ophthalmology & Visual Science*. 2017;58(7):3240. doi:10.1167/iovs.16-21053

35. Mei S, Montanari A, Nguyen P-M. A mean field view of the landscape of two-layer neural networks. *Proc Natl Acad Sci U S A*. 2018;115(33):E7665-E7671.

36. Do H. Animal and physiological psychology. *Annu Rev Psychol*. 1950;1. doi:10.1146/annurev.ps.01.020150.001133

37. Rosenblatt F. PRINCIPLES OF NEURODYNAMICS. PERCEPTRONS AND THE THEORY OF BRAIN MECHANISMS. Published online 1961. doi:10.21236/ad0256582

38. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986;323(6088):533-536.

39. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.

40. Hahnloser RH, Sarpeshkar R, Mahowald MA, Douglas RJ, Seung HS. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*. 2000;405(6789):947-951.

41. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*. 2017;60(6):84-90. doi:10.1145/3065386

42. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science*. 2006;313(5786):504-507.

43. Kaehler A, Bradski G. *Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library*. "O'Reilly Media, Inc."; 2016.

44. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. *arXiv*. Accessed October 13, 2020. http://arxiv.org/abs/1512.04150

45. Ronneberger O Fischer P Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*. Accessed October 13, 2020. http://arxiv.org/abs/1505.04597

46. Ghosh S, Das N, Das I, Maulik U. Understanding Deep Learning Techniques for Image Segmentation. Published online July 13, 2019. Accessed October 13, 2020. http://arxiv.org/abs/1907.06119

47. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks. Published online

June 10, 2014. Accessed October 13, 2020. http://arxiv.org/abs/1406.2661

48. Google. Generative Adversarial Networks. Published 2020. Accessed October 13, 2020. https://developers.google.com/machine-learning/gan

49. Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv.org*. Published online November 21, 2016. Accessed October 13, 2020. http://arxiv.org/abs/1611.07004

50. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Published online March 30, 2017. Accessed October 13, 2020. http://arxiv.org/abs/1703.10593

51. de Sisternes L, Simon N, Tibshirani R, Leng T, Rubin DL. Quantitative SD-OCT imaging biomarkers as indicators of age-related macular degeneration progression. *Invest Ophthalmol Vis Sci*. 2014;55(11):7093-7103.

52. Schmidt-Erfurth U, Waldstein SM, Klimscha S, et al. Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence. *Invest Ophthalmol Vis Sci*. 2018;59(8):3199-3208.

53. Vogl W-D, Waldstein SM, Gerendas BS, Schlegl T, Langs G, Schmidt-Erfurth U. Analyzing and Predicting Visual Acuity Outcomes of Anti-VEGF Therapy by a Longitudinal Mixed Effects Model of Imaging and Clinical Data. *Invest Ophthalmol Vis Sci*. 2017;58(10):4173-4181.

54. Bogunovic H, Montuoro A, Baratsits M, et al. Machine Learning of the Progression of Intermediate Age-Related Macular Degeneration Based on OCT Imaging. *Invest Ophthalmol Vis Sci*. 2017;58(6):BIO141-BIO150.

55. Niu S, de Sisternes L, Chen Q, Rubin DL, Leng T. Fully Automated Prediction of Geographic Atrophy Growth Using Quantitative Spectral-Domain Optical Coherence Tomography Biomarkers. *Ophthalmology*. 2016;123(8):1737-1750.

56. Caixinha M, Amaro J, Santos M, Perdigao F, Gomes M, Santos J. In-Vivo Automatic Nuclear Cataract Detection and Classification in an Animal Model by Ultrasounds. *IEEE Transactions on Biomedical Engineering*. 2016;63(11):2326-2335. doi:10.1109/tbme.2016.2527787

57. Kim SJ, Cho KJ, Oh S. Development of machine learning models for diagnosis of glaucoma. *PLoS One*. 2017;12(5):e0177726.

58. Zhang Q, Zhu S-C. Visual Interpretability for Deep Learning: a Survey. Published online February 2, 2018. Accessed October 13, 2020. http://arxiv.org/abs/1802.00614

59. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. Published online November 12, 2013. Accessed October 13, 2020. http://arxiv.org/abs/1311.2901

60. 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). doi:10.1167/tvst.4.6.5

61. 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). doi:10.1001/jamaophthalmol.2016.0611

62. Castellanos FX, Giedd JN, Marsh WL, et al. Quantitative brain magnetic resonance imaging in attention-deficit hyperactivity disorder. *Arch Gen Psychiatry*. 1996;53(7). doi:10.1001/archpsyc.1996.01830070053009

63. Chiang MF. Image analysis for retinopathy of prematurity: where are we headed? *J AAPOS*. 2012;16(5). doi:10.1016/j.jaapos.2012.08.001

64. Lundberg T, Westman G, Hellstrom S, Sandstrom H. Digital imaging and telemedicine as a tool for studying inflammatory conditions in the middle ear--evaluation of image quality and agreement between examiners. *Int J Pediatr Otorhinolaryngol*. 2008;72(1). doi:10.1016/j.ijporl.2007.09.015

65. Chiang MF, Starren J, Du YE, et al. Remote image based retinopathy of prematurity diagnosis: a receiver operating characteristic analysis of accuracy. *Br J Ophthalmol*. 2006;90(10). doi:10.1136/bjo.2006.091900

66. Smith RA, Saslow D, Sawyer KA, et al. American Cancer Society guidelines for breast cancer screening: update 2003. *CA Cancer J Clin*. 2003;53(3). doi:10.3322/canjclin.53.3.141

67. Bartlett E, DeLorenzo C, Parsey R, Huang C. Noise contamination from PET blood sampling pump: Effects on structural MRI image quality in simultaneous PET/MR studies. *Med Phys*. 2018;45(2). doi:10.1002/mp.12715

68. Bartling H, Wanger P, Martin L. Automated quality evaluation of digital fundus photographs. *Acta Ophthalmol*. 2009;87(6). doi:10.1111/j.1755-3768.2008.01321.x

69. Katuwal GJ, Kerekes J, Ramchandran R, Sisson C, Rao N. Automatic fundus image field detection and quality assessment. *IEEE Western New York Image Processing Workshop*. Accessed October 14, 2020. https://ieeexplore.ieee.org/abstract/document/6890980

70. Dietrich TJ, Ulbrich EJ, Zanetti M, Fucentese SF, Pfirrmann CW. PROPELLER technique to improve image quality of MRI of the shoulder. *AJR Am J Roentgenol*. 2011;197(6). doi:10.2214/AJR.10.6065

71. Jiang Y, Huo D, Wilson DL. Methods for quantitative image quality evaluation of MRI parallel reconstructions: detection and perceptual difference model. *Magn Reson Imaging*. 2007;25(5). doi:10.1016/j.mri.2006.10.019

72. Li H, Hu W, Xu ZN. Automatic no-reference image quality assessment. *Springerplus*. 2016;5(1). doi:10.1186/s40064-016-2768-2

73. Maberley D, Morris A, Hay D, Chang A, Hall L, Mandava N. A comparison of digital retinal image quality among photographers with different levels of training using a non-mydriatic fundus camera. *Ophthalmic Epidemiol*. 2004;11(3). doi:10.1080/09286580490514496

74. Niemeijer M, Abràmoff, van Ginneken B. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. *Med Image Anal*. 2006;10(6). doi:10.1016/j.media.2006.09.006

75. Patel T, Peppard H, Williams MB. Effects on image quality of a 2D antiscatter grid in x-ray digital breast tomosynthesis: Initial experience using the dual modality (x-ray and molecular) breast tomosynthesis scanner. *Med Phys*. 2016;43(4). doi:10.1118/1.4943632

76. Smet MH, Breysem L, Mussen E, Bosmans H, Marshall NW, Cockmartin L. Visual grading analysis

of digital neonatal chest phantom X-ray images: Impact of detector type, dose and image processing on image quality. *Eur Radiol*. 2018;28(7). doi:10.1007/s00330-017-5301-2

77. Strauss RW, Krieglstein TR, Priglinger SG, et al. Image quality characteristics of a novel colour scanning digital ophthalmoscope (SDO) compared with fundus photography. *Ophthalmic Physiol Opt*. 2007;27(6). doi:10.1111/j.1475-1313.2007.00512.x

78. Takeda H, Minato K, Takahasi T. High quality image oriented telemedicine with multimedia technology. *Int J Med Inform*. 1999;55(1). doi:10.1016/s1386-5056(99)00017-9

79. Teich S, Al-Rawi W, Heima M, et al. Image quality evaluation of eight complementary metal-oxide semiconductor intraoral digital X-ray sensors. *Int Dent J*. 2016;66(5). doi:10.1111/idj.12241

80. Veiga D, Pereira C, Ferreira M, Gonçalves L, Monteiro J. Quality evaluation of digital fundus images through combined measures. *Journal of medical imaging (Bellingham, Wash)*. 2014;1(1). doi:10.1117/1.JMI.1.1.014001

81. Wang S, Jin K, Lu H, Cheng C, Ye J, Qian D. Human Visual System-Based Fundus Image Quality Assessment of Portable Fundus Camera Photographs. *IEEE Trans Med Imaging*. 2016;35(4). doi:10.1109/TMI.2015.2506902

82. Quinn GE. Retinopathy of prematurity blindness worldwide: phenotypes in the third epidemic. *Eye Brain*. 2016;8:31-36.

83. Coyner AS, Swan R, Brown JM, et al. Deep Learning for Image Quality Assessment of Fundus Images in Retinopathy of Prematurity. *AMIA Annu Symp Proc*. 2018;2018. Accessed October 14, 2020. https://pubmed.ncbi.nlm.nih.gov/30815164/

84. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. Published online December 2, 2015. Accessed October 14, 2020. http://arxiv.org/abs/1512.00567

85. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. Published online September 1, 2014. Accessed October 14, 2020. http://arxiv.org/abs/1409.0575

86. Chiang MF, Gelman R, Martinez-Perez ME, et al. Image analysis for retinopathy of prematurity diagnosis. *J AAPOS*. 2009;13(5). doi:10.1016/j.jaapos.2009.08.011

87. Giardini ME, Livingstone IA, Jordan S, et al. A smartphone-based ophthalmoscope. *Conf Proc IEEE Eng Med Biol Soc*. 2014;2014. doi:10.1109/EMBC.2014.6944049

88. Wu AR, Fouzdar-Jain S, Suh DW. Comparison Study of Funduscopic Examination Using a Smartphone-Based Digital Ophthalmoscope and the Direct Ophthalmoscope. *J Pediatr Ophthalmol Strabismus*. 2018;55(3). doi:10.3928/01913913-20180220-01

89. Saha SK, Fernando B, Cuadros J, Xiao D, Kanagasingam Y. Automated Quality Assessment of Colour Fundus Images for Diabetic Retinopathy Screening in Telemedicine. *J Digit Imaging*. 2018;31(6). doi:10.1007/s10278-018-0084-9

90. Swan R, Kim SJ, Peter Campbell J, et al. Natural course and predictive value of pre-plus disease in retinopathy of prematurity: results of a multicenter prospective cohort study. *Invest Ophthalmol Vis Sci*. 2017;58(8):5545-5545.

91. Chiang MF, Keenan JD, Du YE, et al. Assessment of image-based technology: impact of referral cutoff on accuracy and reliability of remote retinopathy of prematurity diagnosis. *AMIA Annu Symp Proc*. 2005;2005. Accessed December 1, 2020. https://pubmed.ncbi.nlm.nih.gov/16779015/

92. Chiang MF, Keenan JD, Starren J, et al. Accuracy and reliability of remote retinopathy of prematurity diagnosis. *Arch Ophthal*. 2006;124(3). doi:10.1001/archopht.124.3.322

93. Hutchinson AK, Melia M, Yang MB, VanderVeen DK, Wilson LB, Lambert SR. Clinical Models and Algorithms for the Prediction of Retinopathy of Prematurity: A Report by the American Academy of Ophthalmology. *Ophthalmology*. 2016;123(4). doi:10.1016/j.ophtha.2015.11.003

94. Taylor S, Brown JM, Gupta K, et al. Monitoring Disease Progression With a Quantitative Severity Scale for Retinopathy of Prematurity Using Deep Learning. *JAMA Ophthalmol*. 2019;137(9):1022-1028.

95. Bellsmith KN, Brown J, Kim SJ, et al. Aggressive Posterior Retinopathy of Prematurity: Clinical and Quantitative Imaging Features in a Large North American Cohort. *Ophthalmology*. 2020;127(8). doi:10.1016/j.ophtha.2020.01.052

96. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830.

97. Committee on Fetus, Newborn. Age Terminology During the Perinatal Period. *Pediatrics*. 2004;114(5):1362-1364.

98. Toslak D, Ayata A, Liu C, Erol MK, Yao X. Wide-field smartphone fundus video camera based on miniaturized indirect ophthalmoscopy. *Retina*. 2018;38(2). doi:10.1097/IAE.0000000000001888

99. Raju B, Raju NS, Akkara JD, Pathengay A. Do it yourself smartphone fundus camera - DIYretCAM. *Indian J Ophthalmol* . 2016;64(9). doi:10.4103/0301-4738.194325

100. H NK, Nakatsuka A, El-Annan J. Smartphone Fundus Photography. *J Vis Exp*. 2017;(125). doi:10.3791/55958

101. Gilbert C, Fielder A, Gordillo L, et al. Characteristics of infants with severe retinopathy of prematurity in countries with low, moderate, and high levels of development: implications for screening programs. *Pediatrics*. 2005;115(5). doi:10.1542/peds.2004-1180

102. Gilbert C. Retinopathy of prematurity: a global perspective of the epidemics, population of babies at risk and implications for control. *Early Hum Dev*. 2008;84(2). doi:10.1016/j.earlhumdev.2007.11.009

103. Wang H, Peng H, Chang Y, Liang D. A survey of GPU-based acceleration techniques in MRI reconstructions. *Quantitative Imaging in Medicine and Surgery*. 2018;8(2):196-208. doi:10.21037/qims.2018.03.07

104. Andreini P, Bonechi S, Bianchini M, Mecocci A, Scarselli F, Sodi S. A Two Stage GAN for High Resolution Retinal Image Generation and Segmentation. *arXiv*. Accessed October 14, 2020. http://arxiv.org/abs/1907.12296

105. Beers A, Brown JM, Chang K, Campbell JP, Ostmo S, Chiang MF, Kalpathy-Cramer J. High-resolution medical image synthesis using progressively grown generative adversarial networks. *arXiv*. Accessed October 14, 2020. http://arxiv.org/abs/1805.03144

106. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*. 2019;58:101552. doi:10.1016/j.media.2019.101552

107. Costa P, Galdran A, Meyer MI, Mendonça AM, Campilho A. Adversarial Synthesis of Retinal Images from Vessel Trees. *Lecture Notes in Computer Science*. Published online 2017:516-523. doi:10.1007/978-3-319-59876-5_57

108. NVIDIA. NVIDIA/pix2pixHD. Published 2017. Accessed October 14, 2020. https://github.com/NVIDIA/pix2pixHD

109. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv*. Accessed October 14, 2020. http://arxiv.org/abs/1912.01703

110. Isola P, Zhu J, Zhou T, Efros AA. Image-to-Image Translation with Conditional Adversarial Networks. Accessed October 14, 2020. https://phillipi.github.io/pix2pix/

111. Welstead S. Fractal and Wavelet Image Compression Techniques. Published online 1999. doi:10.1117/3.353798

112. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity - IEEE Journals & Magazine. *IEEE Xplore*. Accessed October 14, 2020. https://ieeexplore.ieee.org/document/1284395

113. Yu Z, Xiang Q, Meng J, Kou C, Ren Q, Lu Y. Retinal image synthesis from multiple-landmarks input with generative adversarial networks. *Biomed Eng Online*. 2019;18(1):62.

114. Costa P, Galdran A, Meyer MI, et al. End-to-End Adversarial Retinal Image Synthesis. *IEEE Trans Med Imaging*. 2018;37(3):781-791.

115. Office for Civil Rights (OCR). Summary of the HIPAA Privacy Rule. Published May 7, 2008. Accessed October 14, 2020. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

116. Office for Civil Rights (OCR). Methods for De-identification of PHI. Published September 7, 2012. Accessed October 14, 2020. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

117. Wang T-C, Liu M-Y, Zhu J-Y, Tao A, Kautz J, Catanzaro B. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. Published online November 30, 2017. Accessed October 14, 2020. http://arxiv.org/abs/1711.11585

118. Jun-Yan Zhu Taesung Park. junyanz/pytorch-CycleGAN-and-pix2pix. Published online 2017. Accessed October 14, 2020. https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix