Geographic variation and ethnicity in diabetic retinopathy detection via deep learning

Ali SERENER1*, Sertan SERTE1

1Department of Electrical and Electronic Engineering, Faculty of Engineering, Near East University, Nicosia, North Cyprus, Mersin 10, Turkey

Abstract: The prevalence of diabetes is on the rise steadily around the globe. Diabetic Retinopathy (DR) is a result of damage to the blood vessels in the retina due to diabetes and its fast treatment is crucial for preventing possible blindness. The diagnosis of DR is done mostly using a comprehensive eye exam where the eye is dilated for a better inspection. The analysis by an ophthalmologist is prone to human error and thus automatic and highly accurate detection of DR is preferred for an earlier and better diagnosis. It is important, however, that automatic detection is accurate for all data collected from patients of different geographic and ethnic backgrounds. In this paper, the automatic detection of DR with a deep learning algorithm is analyzed when geographic and ethnic information of the patients are also integrated into the architecture. It is shown that robust and generalizable DR detection performance is linearly related to the correlation of geographic and ethnic patient information between the training and the testing datasets. The deep learning model created eliminates geographic variation in the detection and works for patients of all ethnicities.

Key words: Deep learning, diabetic retinopathy, ethnicity, fundus images, geographic variation

1. Introduction

Diabetes mellitus (DM) is a group of diseases that affects 425 million people around the world1. According to the 2017 International Diabetes Federation (IDF) diabetes atlas1, over one million children and adolescents have type 1 diabetes. What is more, one in two adults with diabetes is left undiagnosed. It is estimated by IDF that the number of people with diabetes is going to increase to 629 million by the year 2045.

Diabetic Retinopathy (DR) is a major complication of DM that may lead to permanent vision loss if left untreated. The severity of DR can be categorized into five classes [1]: no retinopathy, mild nonproliferative DR (NPDR), moderate NPDR, severe NPDR, and proliferative DR (PDR). Using this terminology, nonreferable DR (NRDR) can be defined as no or mild NPDR, referable DR (RDR) as moderate NPDR or worse, and vision threatening DR (VTDR) as severe NPDR or worse. Figure 1 shows a captured fundus image of a patient diagnosed with mild DR.

For the automatic classification of eye diseases, deep learning methods have been used by many researchers, such as [2–4]. One of the most well-known deep convolutional neural networks is ResNet [5]. Along

*Correspondence: ali.serener@neu.edu.tr

with ResNet, several other deep convolutional neural networks have also been introduced, such as GoogLeNet [6, 7], VGG [8], AlexNet [9] and DenseNet [10].

The DR research carried out in the past have used public or private datasets. A high number of research papers exist (e.g. [11] and [12]) that aim to obtain excellent DR detection rates using deep convolutional networks. They compare the simulation results to the diagnosis of ophthalmologists in order to show that deep learning can many times do better. However, their models are mostly built on a specific dataset related to a specific geographic region or ethnic group. As a result, the models created allow DR detection only for that geographic region or ethnic group.

A few researchers who carried out such work [11, 13, 14] have further reported that the performance of DR classification should ideally be investigated by also taking into account the race and ethnicity of patients, geographic variation within the testing and training datasets and the factors, such as pupil dilation, affecting the quality of images. Their desire to involve the race and ethnicity of patients had to do with the well-known fact that retinal fundus appearance changes due to pigmentation of the retina and ocular structure [15], which in turn is related to the person’s ethnic group and eye color. It was this fact that also led the researchers such as Giancardo et al. [12] and Ting et al. [13] to use datasets involving patients of various ethnic groups.

Still, these researchers or the others have not so far investigated the relationship between patient’s geography of residence or ethnicity and DR detection. Furthermore, past research has not fully reported if their systems could work with patients of different geographic locations and ethnicities. As such, it is the goal of this research to analyze for the first time the impacts of geographic variation and ethnicity of the patients on DR classification performance using a deep learning ResNet architecture. During this analysis, five publicly available fundus image datasets, namely Kaggle², Messidor [16], E-Optha [17], HRF [18], and IDRID (2018 IEEE ISBI Challenge) [19] were used. Sample fundus images from three datasets are shown in Figure 2.

Specifically, the contributions of this paper to the research community are a) integration of DR patients’ geographic location information into DR detection and analysis of the performance when the training and testing datasets consist of patients from the same country, the same continent as well as different continents b) integration of patients’ ethnicity information into DR detection and investigation of the effects of having training and testing datasets consisting of coethnic and multiethnic patients c) creation of a robust and generalizable deep learning model that eliminates geographic variation in detection and works for patients of all ethnicities.

The organization of this paper is as follows. First, related work on DR disease classification using fundus images is given. Then, ResNet architecture and databases used for performance evaluation are explained. Finally, performance results are tabulated and analyzed. Conclusions based on this analysis are also outlined.

2. Related Work

Abràmoff investigated the benefits and limitations of Iowa automated DR detection system in [11]. He concluded that having a detection system that works on well-defined populations with different ethnic and racial backgrounds is more important than the automated detection algorithm itself. During his investigation, he found out that the detection system employed failed if it was tested on a dataset that was collected from a population with different DR incidence, race or ethnic backgrounds. He further speculated that the gender of the patient or the resolution of the camera used for fundus imaging had no effect on the performance of the system used.

Similarly, there were a few studies done later on involving the ethnicity of the patients. In [13], diabetic retinopathy and related eye diseases were detected by evaluating fundus images obtained in Singapore and further validated on other DR cases from 10 additional multietnic datasets. Giancardo et al. [12] introduced a new publicly available dataset of various ethnic groups and levels of diabetic macular edema (DME), a result of DR, in order to test their new methodology based on a novel set of features. Cree [20] used color normalization in order to reduce inter-patient and intra-patient variability in microaneurysm detection of populations with diverse races.

Note that even though the Iowa DR detection system of Abràmoff [11] used many retinal images from DR screening centers around the world, only publicly available dataset at the time was Messidor dataset [16]. This dataset was later used by a few researchers for further DR research, including the work by Giancardo et al. [12] who detected DME by finding the exudates on the fundus images of [16] using a novel set of features. In [21], both microaneurysms and hemorrhages were able to be detected on the images of [16] using a new set of shape features, called Dynamic Shape Features, for better screening and grading of diabetic retinopathy. Pires [22] used the bag-of-visual-words algorithm in order to detect DR-related regions on three public datasets, including [16]. In [23], a method based on combining both deep learned and domain knowledge was used for red lesion detection of images in various datasets, including [16]. In [24], images from the dataset of [16] were also used to evaluate a method that was proposed for faster convolutional neural network training by dynamically selecting misclassified negative samples.

The number of publicly available datasets has since increased, with Drive [25], DIARETDB1 [26] and Aria datasets being among them. The other publicly available dataset provided by EyePACS [27] and used

---

in Kaggle DR competition has created a boom in the research of DR. With the fundus photographs from this
competition and a private dataset of almost 110,000 photographs, Quellec [28] used ConvNet to produce high-
quality heatmaps in order to detect DR by segmenting microaneurysms, hemorrhages, exudates and cotton-wool
spots. Pratt [29] used convolutional neural networks to classify the severity of DR into five classes, namely no
DR, mild DR, moderate DR, severe DR, and proliferative DR, using Kaggle dataset. In [30], deep convolutional
neural networks were used to classify the retinal images of Kaggle dataset into the five stages of DR. Yu [2]
proposed a novel method that combines unsupervised features from saliency map and supervised features from
convolutional neural networks to detect the quality of the retinal fundus images of Kaggle dataset. In [3], the
dataset of Kaggle was used in order to see the effects of changing certain simulation parameters, such as batch
size, epoch, and training dataset size, on the detection of diabetic retinopathy using ConvNet deep learning
architecture.

The goal of this paper is to improve upon past research and use five publicly available datasets in order
to take into account the geographic location and ethnicity of the patients in DR detection. It further aims to
observe how detection performance changes when the effects of variation in a geographic location or patient
ethnicity are eliminated.

3. Method

A ResNet deep learning algorithm is used in this paper. A set of images is first used to train the ResNet
architecture. Then, this is used to model the fundus images for DR classification. In order to create a robust
and generalizable model and eliminate the geographic location and patient ethnicity dependency, the training
dataset is included images from patients who are in the same geographic region or the same ethnicity as the
patients whose fundus images are in the testing dataset. Finally, this model is used to evaluate the system
performance. During this evaluation, a new set of fundus images is used.

The proposed approach is based on fine-tuned 18 layer ResNet (ResNet-18) network architecture (see
Figure 3). This network consists of 18 convolutional layers. First four convolutional layers contain $3 \times 3 \times 64$
filter. The following four convolutional layers after that contain $3 \times 3 \times 128$ filters. Finally, the last four layers
contain $3 \times 3 \times 512$ filters. The last layer of the ResNet-18 model is the fully connected layer (fc), where fc shows
the predictions of classes. The ResNet model is based on residual learning, as also shown in Figure 3. Classic
convolutional neural network models directly calculates $H(x)$ mapping for identity $x$. In residual learning,
however, a different mapping is employed i.e. $F(x) = H(x) - x$. Residual learning allows better network
performance even though the deeper networks (higher number of convolutional layers) are known to be difficult
to optimize.

Fine-tuned model allows modeling retinal images more accurately. Since the number of some of the retinal
image sets are small, the network parameters might not be estimated accurately. On the other hand, adapting
already trained network parameters to new data leads to a well-trained network. The ResNet-18 model used is
trained on ImageNet dataset. This dataset contains many images and allows good parameter estimation. The
model training is performed using GeForce GTX 1080 Ti GPU and Caffe deep learning framework.

All images are resized to $256 \times 256$ image size and $224 \times 224$ patches are extracted while ResNet-18 model
is trained. Image augmentation (25, 40, and 120 degrees rotation) is done whenever necessary in order to create
large and balanced datasets. However, no image quality improving preprocessing methods have been used.
4. Datasets

System performance is evaluated using five datasets. These datasets are Kaggle, Messidor, E-Optha, HRF, and IDRID. Their details are described in the following sections.

4.1. Kaggle Dataset

This dataset contains a total of 83702 fundus images. These images are categorized as healthy (no DR), mild NPDR, moderate NPDR, severe NPDR, and proliferative DR. A total of 65343, 13153, 2087 and 1914 images, respectively, are present in each category. Table 1 lists the number of images in each class of DR for this dataset. Note that this dataset contains images taken using various camera models from patients with different ethnic backgrounds in different clinics located in North America.

4.2. Messidor Dataset

There are a total of 1200 fundus images in this dataset. Unlike Kaggle dataset, there is no proliferative DR class present in this dataset and thus the images are categorized into four classes. The total number of images in each category are 546, 153, 247 and 254, respectively (Table 1). The images in this dataset were taken at Hôpital Lariboisière located at Paris/France, Brest University Hospital located at Brest/France, and Saint-Etienne University Hospital located at St Etienne/France. The images, which have resolutions of 1440 × 960, 2240 × 1488 or 2304 × 1536 pixels, were acquired by using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinograph with a 45-degree field of view (FOV).

4.3. E-Optha Dataset

This dataset has 463 images in three categories, namely healthy (no DR), microaneurysm (an indication of DR) and exudates. There are a total of 268, 148 and 47 images in each category, respectively (Table 1). The images
Table 1: Number of Images in Each Category of Five Datasets.

<table>
<thead>
<tr>
<th>Category</th>
<th>Kaggle</th>
<th>Messidor</th>
<th>E-Optha</th>
<th>HRF</th>
<th>IDRID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>65343</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
</tr>
<tr>
<td>Mild DR</td>
<td>6205</td>
<td>Mild DR</td>
<td>Microaneurysms</td>
<td>DR</td>
<td>Mild DR</td>
</tr>
<tr>
<td>Mod. DR</td>
<td>13153</td>
<td>Mod. DR</td>
<td>Exudates</td>
<td>Glaucoma</td>
<td>Mod. DR</td>
</tr>
<tr>
<td>Severe DR</td>
<td>2087</td>
<td>Severe DR</td>
<td></td>
<td></td>
<td>Severe DR</td>
</tr>
<tr>
<td>Prolif. DR</td>
<td>1914</td>
<td></td>
<td></td>
<td></td>
<td>Prolif. DR</td>
</tr>
<tr>
<td>Total</td>
<td>83702</td>
<td>1200</td>
<td>463</td>
<td>45</td>
<td>512</td>
</tr>
</tbody>
</table>

were collected from OPTHDIAT teleophthalmology network in France.

4.4. HRF Dataset
HRF dataset contains 45 images divided equally among three categories, namely healthy (no DR), DR and glaucomatous (Table 1). They were taken using a Canon CF-60UVi camera with a resolution of $3504 \times 2336$ pixels. The database was provided by the Department of Ophthalmology of Friedrich-Alexander University located at Erlangen/Germany and the Department of Biomedical Engineering of the Brno University of Technology, located at Brno/Czech Republic.

4.5. IDRID Dataset
This dataset contains a total of 516 fundus images. These images are categorized into the same five classes as the Kaggle dataset. A total of 168, 25, 168, 93 and 62 images, respectively, exist in each category. Table 1 lists the number of images in each class of IDRID dataset. The images in this dataset were taken at a clinic located in India. Each image has a resolution of $4288 \times 2848$ pixels and was captured using a Kowa VX-10 alpha digital fundus camera with a 50-degree FOV.

5. Performance Evaluation

5.1. Geographic Variation in DR Detection
In this section, we would like to show the results of experiments ran to investigate the relationship in terms of algorithm performance between the patient’s geography of residence and the DR detection. Specifically, two cases have been looked at. First, we wanted to see how the detection performance is affected if the training and testing datasets have patients who are from the same country. Second, we ran two sets of experiments to see the relationship when training and testing datasets contain data from patients who reside in the same continent or who are from different continents.

5.1.1. Case I: DR Detection within a Country
The first case includes four experiments. The goal of these experiments is to find out the change in DR detection performance if the DR images in the training and testing datasets contain DR images of patients who live in the same country.

Initially, two experiments are carried out that contain as their training and testing datasets images from patients who live in France and the United States, respectively. The results are given in Table 2. Note the trade-off between sensitivity and specificity results [31].
Then, another experiment is run that contains training dataset images from Kaggle. These DR images are from patients who live in the United States (mainly the state of California [27]). One of the reasons we preferred to use Kaggle images here was that it contains a large number of images of all DR severity levels [32]. The testing dataset of this experiment, obtained from E-Optha dataset, includes fundus images of patients who live in France. As shown in Table 2, the accuracy, the sensitivity and the specificity of this experiment are 56.25%, 55.59%, and 57.43%, respectively. The accuracy results here are less than that of the second experiment since now the training and testing datasets contain images from patients who live in different countries.

Next, a fourth experiment is carried out where images from a different dataset (Messidor), containing images of patients who are also from France, are added to the initial training dataset of the third experiment. Messidor dataset contains images with four different DR severity levels. It is observed that with the addition of this dataset, the accuracy compared to the third experiment increases by about 25%, the sensitivity increases by about 33% and the specificity increases by about 11% (see Table 3).

According to the 2013 U.S. Census Bureau data, the number of French Americans in the United States was about 10.7 million, about 0.8 million of which were living in the state of California. As the accuracy of the third experiment is a little larger than 50%, it is highly likely that Kaggle dataset does not include many images of patients with French ancestries. However, with the addition of more DR images of patients from France to the training dataset, the accuracy increases in the fourth experiment. This increase in performance is an indication of a higher correlation between the retinal pigmentations and ocular structures of French patients whose images are in the training and testing datasets.

---

5.1.2. Case II (a): DR Detection Within a Continent

Next, we ran two more experiments to see how having training and testing datasets of patients who live in the same continent is linked to the DR detection performance.

The training dataset of the first experiment again includes DR images of patients from the United States (obtained from Kaggle dataset), whereas, the testing dataset now includes those of patients from Germany and the Czech Republic (through HRF dataset). An accuracy of 80%, sensitivity of 80% and specificity of 80% are obtained from this experiment, as shown in Table 2.

The second experiment of this case adds DR images of patients from France (obtained from Messidor dataset) to the training dataset. With this addition, the accuracy, compared to the first experiment, now increases by about 13% and the sensitivity increases by 25%. The specificity stays the same (Table 3).

The total number of German and Czech Americans in the United States was about 49.2 million and 3.25 million of those lived in the state of California, according to the 2013 U.S. Census Bureau data. Thus, there is a higher chance of Kaggle dataset to include images of patients from Germany and the Czech Republic than France and hence the accuracy of the second experiment of this case is higher than the first experiment.

When the data of the patients from France were added to the training dataset, the accuracy has increased by about 13%. This shows that there is a higher correlation among the retinal pigmentations and ocular structures of the patients from France, Germany and the Czech Republic and that had a positive impact on the deep learning performance.

5.1.3. Case II (b): DR Detection Between Continents

The aim of this third case is to show the relationship between DR detection and patient’s geography of residence if now the data in the training and testing datasets are collected from two different continents.

To do this, first, an experiment is run with a training dataset that is a mixture of data from North America and Europe (using Kaggle and Messidor datasets). The testing dataset has DR images obtained from Asia (through IDRID dataset). The accuracy, sensitivity, and specificity of this experiment are 58.54%, 56.54%, and 72%, respectively, as shown in Table 2.

The second experiment of this case includes a testing dataset which has DR images obtained from patients from Europe instead (through HRF dataset). The training dataset is kept the same as the first experiment. It is observed from the performance results that the accuracy now increases by about 54%, the sensitivity increases by about 77% and the specificity by about 11% (Table 3).

It is known that the total number of Indian Americans in the United States was about 3.5 million in the year 2013 and about 0.7 million of those lived in the state of California. Therefore, there is a lower chance that Kaggle dataset contains images collected from patients of Asian Indian ancestries than from patients of German and Czech descents (Messidor dataset does not contain any data from patients of these three ancestries). Hence, when the deep learning algorithm was trained on the same datasets but tested on HRF dataset, there was an increase in accuracy performance of about 54%. That, again, is a clear indication of the fact that the higher the correlation between the retinal pigmentations and eye structures of the patients, the higher the deep learning performance.

5.2. Ethnicity in DR Detection

Having investigated the relationship between the DR patients’ places of residence and the deep learning performance, we next would like to analyze the impact of the ethnicities of the patients’ on the performance.
Table 4: Performance Results of Ethnicity Investigation.

<table>
<thead>
<tr>
<th>Case I: Coethnicity</th>
<th>Train Dataset</th>
<th>Test Dataset</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaggle (Diverse)</td>
<td>E-Optha (French)</td>
<td>56.25%</td>
<td>55.59%</td>
<td>57.43%</td>
<td></td>
</tr>
<tr>
<td>Kaggle (Diverse) + Messidor (French)</td>
<td>E-Optha (French)</td>
<td>70.19%</td>
<td>73.88%</td>
<td>63.51%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case II: Multiethnicity (a)</th>
<th>Train Dataset</th>
<th>Test Dataset</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaggle (Diverse)</td>
<td>E-Optha (French)</td>
<td>56.25%</td>
<td>55.59%</td>
<td>57.43%</td>
<td></td>
</tr>
<tr>
<td>Kaggle (Diverse) + Messidor (French)</td>
<td>HRF (German/Czech)</td>
<td>80%</td>
<td>80%</td>
<td>80%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case II (b)</th>
<th>Train Dataset</th>
<th>Test Dataset</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaggle (Diverse) + Messidor (French)</td>
<td>IDRID (India)</td>
<td>58.54%</td>
<td>56.54%</td>
<td>72%</td>
<td></td>
</tr>
<tr>
<td>Kaggle (Diverse) + Messidor (French)</td>
<td>E-Optha (French)</td>
<td>70.19%</td>
<td>73.88%</td>
<td>63.51%</td>
<td></td>
</tr>
<tr>
<td>Kaggle (Diverse) + Messidor (French)</td>
<td>HRF (German/Czech)</td>
<td>90%</td>
<td>100%</td>
<td>80%</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Percent Change Between Performance Results of Ethnicity Investigation Experiments.

<table>
<thead>
<tr>
<th>Case</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case I</td>
<td>24.78%</td>
<td>32.9%</td>
<td>10.59%</td>
</tr>
<tr>
<td>Case II (a)</td>
<td>42.22%</td>
<td>43.91%</td>
<td>39.3%</td>
</tr>
<tr>
<td>Case II (b)</td>
<td>19.9%</td>
<td>30.67%</td>
<td>-11.79%</td>
</tr>
<tr>
<td></td>
<td>28.22%</td>
<td>35.35%</td>
<td>25.96%</td>
</tr>
</tbody>
</table>

For that, three sets of experiments are designed. The first set is aimed at analyzing the coethnicity effect, that is, how having patients of the same ethnicity in the training and testing datasets affects the deep learning performance. The other two sets of experiments are designed to analyze the multiethnicity effect, that is, how the deep learning performance changes when the training and testing datasets contain images from patients of multiple ethnicities.

5.2.1. Case I: Coethnicity

This case includes two experiments. The goal of these experiments is to analyze the relationship between DR detection performance and the existence of coethnicity between the patients whose fundus images are in the training dataset and the patients whose fundus images are in the testing dataset.

In the first experiment, DR images from Kaggle dataset is used in the training dataset. Kaggle dataset is a collection of DR images collected from people with diverse ethnic backgrounds [14]. The testing dataset is DR images from E-Optha dataset. E-Optha dataset consists of images from patients with French ethnicity. The accuracy of this experiment is 56.25%, the sensitivity is 55.59% and the specificity is 57.43% (see Table 4). As noted in Section 5.1.1, the low performance occurring in this experiment is believed to be due to Kaggle dataset containing a low number of DR images of patients with French ancestries.

In the second experiment, DR images from Messidor dataset is now added to the training dataset of the first experiment. Messidor dataset contains DR images from patients with French ancestries. With this addition, the accuracy increases by about 25%, sensitivity by about 33% and specificity by about 11%, as shown in Table 5. These results certainly show the linear relationship between the ethnic correlation among the patients and the DR detection performance.
5.2.2. Case II (a): Multiethnicity - Implicit Ethnic Correlation

In this case, two experiments are carried out to show the relationship between patient ethnicity and DR detection performance. The training datasets of these experiments have fundus images of patients with diverse ethnic backgrounds. The testing datasets, on the other hand, consist of DR images from two different datasets that contain fundus images of patients with two different ethnic backgrounds. Here, any ethnicity correlation between the training and testing datasets are implicit as the training datasets contain images from ethnically diverse patients.

Both experiments use images from Kaggle dataset as their training datasets. The testing datasets are images from E-Optha and HRF datasets. Again, these datasets are obtained from patients with French and German/Czech ethnicities, respectively.

It is observed from the results tabulated in Table 4 that the accuracy, sensitivity and specificity results of the first experiment are subpar. However, after the second experiment, there is an improvement of about 40% in accuracy performance results. It is believed that this is due to the fact that in the state of California there is a higher population of people with German and Czech ethnicities than French. Hence, the Kaggle dataset has a higher chance of having DR images from patients with German and Czech ethnic backgrounds than French. That further means that there is a higher correlation in terms of retinal pigmentation and ocular structure between the DR images of the patients in Kaggle and HRF datasets than Kaggle and E-Optha datasets.

5.2.3. Case II (b): Multiethnicity - Explicit Ethnic Correlation

The final case includes a set of three distinct experiments. The goal here is to observe the effect on the DR performance of manually adding the DR images of the patients in the Messidor dataset to the training datasets of the two experiments described in Section 5.2.2. Additionally, a third experiment is designed here, this time with the same training dataset as these two experiments but with a testing dataset that contains the DR images of IDRID dataset. With the inclusion of Messidor dataset, we have explicitly added to the training dataset the DR images of patients with French ethnic backgrounds.

Table 4 shows the accuracy, sensitivity and specificity results of these three experiments. The results are displayed in the order of increasing accuracy. Table 5 displays the percent change between the performances of the first two as well as the performances of the second and the third experiments.

When the results are analyzed, it is observed that the accuracy results are the lowest when the testing dataset contains DR images from IDRID dataset and the highest when it includes images from HRF dataset. As shown in Table 5, the accuracy improvement of the second experiment over the first is about 20% and the accuracy improvement of the third over the second is about 28%.

Messidor, IDRID and HRF datasets, included in the first and third experiments, have DR images from patients of different ethnic backgrounds. Thus, the only overlap in terms of ethnicity between the training and testing datasets of the first and the third experiment is the patients with Asian Indian, German and Czech ethnic backgrounds whose fundus images are included in the Kaggle dataset.

The third experiment has higher accuracy performance than the first experiment by reason of the analysis given in Section 5.1.3. Even though the second experiment includes patients with French ethnic backgrounds in both the training and the testing datasets, still the performance results of it are not any better than the third experiment. The reason for this we believe is that in the state of California, where the Kaggle dataset images are from, the population of people with German and Czech ethnic backgrounds is much higher than the population with French descent.
Table 6: DR Severity Stages Investigation.

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>USA</td>
<td>France</td>
<td>USA</td>
<td>France</td>
</tr>
<tr>
<td>No DR</td>
<td>91%</td>
<td>65%</td>
<td>80%</td>
<td>69%</td>
</tr>
<tr>
<td>Mild DR</td>
<td>85%</td>
<td>87%</td>
<td>77%</td>
<td>80%</td>
</tr>
<tr>
<td>Moderate DR</td>
<td>70%</td>
<td>89%</td>
<td>73%</td>
<td>75%</td>
</tr>
<tr>
<td>Severe DR</td>
<td>48%</td>
<td>39%</td>
<td>70%</td>
<td>74%</td>
</tr>
</tbody>
</table>

Figure 4: **Top row:** Detected images, **Middle row:** Undetected DR images, **Bottom row:** Images falsely detected as DR

6. Geographic Variation and Ethnicity in DR Severity Stages

We next investigate the effects of geographic variation and ethnicity in DR severity stages. For this, we run two experiments using Kaggle (USA) and Messidor (France) datasets and we specifically look at no DR, mild DR, moderate DR, and severe DR stages, as shown in Table 6. For these classifications, a few images are listed in Figure 4 where the top row contains detected images i.e. images that are correctly identified to be from healthy eyes or from eyes with DR, the middle row contains undetected DR images i.e. images with DR that went undetected as healthy, and the bottom row contains images that are falsely detected as DR i.e. images that are healthy but identified to have DR. Figure 5 shows the confusion matrices for these two experiments showing the true (actual) and predicted DR classifications for each severity stage of Kaggle and Messidor datasets.

The results show that in general, the deep learning model resulted in a better performance for ethnically diverse people of United States when the area under the receiver operating characteristic curve (AUC), accuracy, and sensitivity performances of no DR and severe DR are concerned. Similarly, the performance was better for ethnically uniform French for AUC, accuracy, and specificity performance results of mild and moderate DR. Notice again the performance trade-off between sensitivity and specificity [31].

Note that, as pointed out in [29] and [33], Kaggle dataset contains more than 10% of ungradable images. This might have affected the classification results of mild and moderate DR (also addressed in [29] and [34]). If ungradable images were exempted, it could be that the AUC, accuracy and specificity results would have been all better for the models using this dataset for all DR severity stages. That, in turn, would have meant DR severity stages are better classified for an ethnically diverse population than an ethnically uniform population.
It is reported in [29] that their preprocessing included performance results of five class model using Kaggle dataset were 75% accuracy, 30% sensitivity, and 95% specificity. This is comparable to our four class Kaggle dataset model given in Table 6 having an average of 75% accuracy, 22% sensitivity, and 88.5% specificity. Similarly, in [35] a no DR versus DR classification using Messidor dataset gave an AUC of 90%, sensitivity of 94%, and specificity of 50%, again with preprocessing. Some of these results are obviously better than our no DR results given in Table 6 and they show the positive effect of having image quality improving preprocessing methods on the performance.

7. Discussions
With the integration of patients’ geographic information in the deep learning architecture for DR detection, it is observed that there is a direct relationship between DR detection and the similarity between the geographic regions where the images in the training and testing datasets are collected from.

Similarly, when the patients’ ethnic background information is integrated into the architecture for DR detection, the analyses show that there is again a direct relationship between DR detection and ethnic similarity of the patients whose images are present in the training and testing datasets.

8. Conclusions
DR is one of the top causes of blindness globally. Artificial intelligence algorithms, including deep learning, have been used in the past in order to efficiently detect the presence of DR in the eye using fundus images.

In this paper, experiments have been run to investigate for the first time the effect of variation in the patients’ place of residence and the ResNet deep learning DR detection performance. Then, another set of experiments were used to analyze again for the first time the relationship between variation in the patients’ ethnic background and such detection performance. The geographic variation investigation included an analysis in terms of the patients’ country and the continent of residence. The ethnicity investigation, on the other hand, covered experiments with coethnic and multiethnic relationships among the patients whose DR images were in the training and testing datasets. The results showed that it is important for testing and training datasets

Figure 5: Confusion Matrices for Geographic Variation and Ethnicity in DR Severity Stages Investigation.
to be obtained from patients who are similar in terms of geographic residence and ethnicity if a robust and generalizable deep learning model is to be looked for.

Future work will involve analyzing the effects on deep learning performance of data collected from patients of different races.

References


