Binary Mode Multinomial Deep Learning Model for more efficient Automated Diabetic Retinopathy Detection

A. Trivedi[*†] J. Desbiens[*‡] Ron Gross[*‡] S. Gupta[§] J. Lavista Ferres[†]
R. Dodhia[†]

Abstract

The ability to rapidly and accurately classify diabetic retinopathy from color fundus photographs is vital to maximize the ability to assess diabetic eye disease early. Our paper compares the performance of binary classification (Refer/No Refer) to multinomial (Diabetic Retinopathy Severity) classification using deep learning models. The binary mode multinomial experiment achieved very high performance for Refer/No Refer DR on clinical datasets with accuracy up to 97.69%. We show how annotating images using image processing improves Multinomial classification in binary mode on a set of fundus images and yields equal if not better performance than a simple binary classification on the same dataset.

1 Introduction

An estimated 314 million people worldwide are visually impaired with 80% of the vision loss being preventable or treatable (World Health Organization). In 2013, the annual financial burden to the U.S. from vision problems was $48.4 billion. Diabetes is the leading cause of disability and blindness in the U.S. affecting 8% of the American population and growing . It is highest in those 65 and older, with less than a high school education and with the lowest income. The number of Americans with diabetic retinopathy (DR) is expected to triple between 2005 and 2050. With delayed or no treatment, DR leads to impaired vision and blindness. Persons with DR are 25 times as likely as the general population to develop blindness, and DR is the leading cause of blindness among working-aged Americans. Telemedicine is an efficient, reliable, and cost-effective diagnostic process. Instead of patients traveling long distances to consult experts, telemedicine leverages technology to enable remote and timely consultations. The benefits of retinal screening with digital photography programs and telemedicine are well studied and documented, allowing quick and accurate patient screening resulting in early detection to treat this largely preventable (>90%) disease. As the population with diabetes grows, the contribution of DR to the overall burden of visual impairment will become increasingly substantial. It can improve patient care, is more cost-effective, and allows screening for common causes of vision loss [1,2,3]. Recently, deep learning (DL) has become an exciting opportunity in clinical ophthalmology for the automated detection of pathology and assessment of retinal diseases [4,5,6]. This can be accomplished in a very short time frame, within seconds facilitating availability of a rapid result to guide follow-up and treatment decisions. An advantage of DL compared to feature-based machine learning is that it uses the raw data with a defined outcome variable, not utilizing the identification of known clinical features to construct the model. The primary objective of this study is to compare the performance of DL to automatically predict the Refer/ No Refer outcome. Specifically, to perform automated detection for the presence of referable diabetic retinopathy in fundus color photographs of patients with diabetes, obtained in the non-eye care setting, comparing multinomial classification in binary mode and simple binary classification. In this paper, we show that using multinomial classification specialized as a binary classifier yields better performance despite the well-known fact that $Pr(AUB) >= Pr(A) + Pr(B)$.

[*]These authors contributed equally as first author
[†]Microsoft
[‡]Intelligent Retinal Imaging Systems
[§]Retina Specialty Institute

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

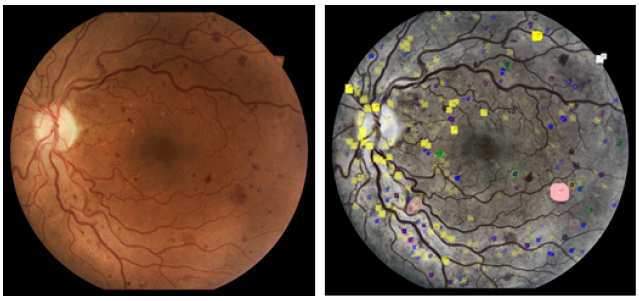


Figure 1: Fundus image preprocessing using Retinex Algorithm

2 Data

This analysis is performed utilizing the Intelligent Retinal Imaging Systems (IRIS) [Pensacola, Florida] FDA-cleared, HIPPA compliant, commercial platform database. The IRIS software is a software-as-a-service (Saas) application that is hosted on the internet which give clinicians the ability to scan a patient's retina with a fundus camera, store the images in a cloud-based datastore, view the images and their associated information, and offer an opinion on the scans. The data set consisted of color digital fundus photography obtained in a non-Eye-Care Professional's environment operated by staff with minimal training. The IRIS software solution is designed to be camera agnostic. For this study, a total of 6,988 graded $45°x40°$ CenterVue DRS camera (CenterVue SpA, Freemont,CA) were utilized. All images are monoscopic, single images centered on the fovea. These images represent a random sample from over 430,000 orders within the IRIS database. Categorization of the images was performed using the Diabetic Retinopathy Severity Scale (DRSS) based on the International Classification of Diabetic Retinopathy (ICDR) criteria. **Refer** image was defined as **No apparent DR, Mild non-Proliferative DR** while **No Refer** image was defined as **Moderate DR, Severe non-Proliferative DR, Proliferative DR**. Ground truth for each image was determined by agreement of two experienced grading board-certified ophthalmologists. If there was disagreement among the two, a third ophthalmologist acted as adjudicator. All images were evaluated in the standard color image as well as an IRIS proprietary image enhancement. The main disagreements lie along the Mild/Moderate boundary. Agreements between the clinicians were assessed by Cohen's kappa. The equation defining it is:

$$\kappa = \frac{p - q}{1 - q}$$

where p is the observed probability of agreement and q is the probability of chance agreement. The larger the kappa, the better is the agreement. The kappa value we obtained was 0.83383 which shows a pretty good agreement. Before training our model, the initial image dataset was randomly divided by a ratio of roughly 80/20 into two datasets, one for training and one for validation.

3 Method

Deep Learning, more precisely the so-called Convolution Networks which belong to a category of Neural Networks that have proven very effective in areas such as image recognition and classification, have achieved superior performance in medical imaging [7]. Transfer learning [8], a machine learning technique where a model trained on one task is re-purposed on a second related task, has become a $de - facto$ method for deep learning applications.

Implementation Details Inception is an architectural addition to CNN to allow for some scale invariance in object recognition. To detect small objects like a dot hemorrhage for instance, Inception can convolve through the image with various iteration sizes, all at the same computational step in the network. We use transfer learning and fine-tuning on the Inception-V3 [9] architecture to address both the binary and multi-class classification tasks. This architecture is very deep, as needed to better learn from images, while having a relatively small number of parameters, which helps in prevent overfitting.

Data Preprocessing Preprocessing is required to ensure that the dataset is consistent and displays only relevant features and to counter the great variability in fundus color images as well. A modified version of the Retinex algorithm [10], a kind of contrast limited adaptive histogram equalization to enhance the features of the images, is performed (see Figure 1). No additional preprocessing was performed on the images except for resizing. ‘

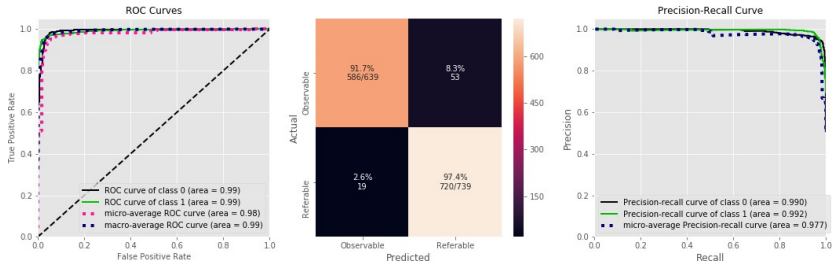
Binary versus Multinomial Classification The dataset was split as follows: 80% for training, 10% for testing, and 10% for validation (i.e., hold-out set). Two DL models were created: one for DR Refer/No Refer binomial classification, the other one for DR multinomial classification (Normal|Mild|Moderate|Severe|Proliferative). The metrics to evaluate the model are computed on the validation sets. The area under the receiver operator characteristic curve (ROC-AUC) is used to assess the performance of the DL models for classifications.

Binary Model The following parameters were used for training the CNN: 50 epochs for fine-tuning, and Adam optimizer with values of learning rate ranging in $[10^{-5}, 10^{-2}]$. Once the model is trained, all training and validation probability vectors are collected and, based on accuracy with cross-validation measure, a best classifier is chosen among a set of eight Scikit-Learn [11] classifiers. The best accuracy achieved with a cross-validation ratio of 20 was 95.86% for the MLPClassifier (see Figure 2a).

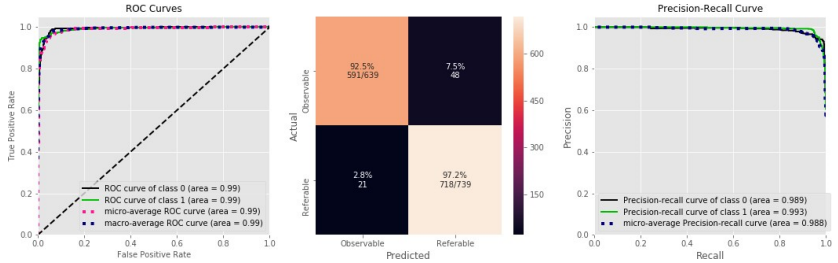
Multinomial Model The same DL parameters were used as in the binary case. Once the model was trained, all 5 training and validation probability vectors were collected and the categories were re-grouped into two upper-categories: Normal + Mild = No Refer, and Moderate + Severe + Proliferative = Refer. The best classifier GaussianNB achieved an accuracy of 96.61% with a cross-validation ratio of 20 (see Figure 2b).

4 Conclusion

CNNs are very effective at image classification, but CNNs trained on the entire retinal scan do not reliably detect the subtle changes present in early-stage retinopathy. Moderate and severe diabetic retinopathy contain macroscopic lesions at a scale that current CNN architectures are optimized to classify. However, the lesions that distinguish normal retinas from mild disease reside in less than 1% of the total pixel volume, a level of subtlety that is difficult for both human interpreters and CNNs to detect. This is a problem that is not necessarily remedied with additional data . For example, Gulshan et al. [4] reported a 93% to 96% recall for their binary classification tasks. However, this was not improved when training with 60,000 samples versus 120,000 samples. To build on existing techniques, research groups have instead augmented existing methods rather than deploy CNNs alone [12]. Several groups have sought to locate these subtle lesions specifically. In the past, microaneurysm detection has been studied using a smaller dataset of 50 training and test images annotated by x, y coordinates and radii to identify these tiny lesions. The performance criteria in this study evaluated the proportion of x and y coordinates falling within a set distance of the ground truth coordinates. Medical images are fraught with subtle features that can be crucial for diagnosis. Fortunately, the most often deployed architectures have been optimized to recognize macroscopic features such as those present in the ImageNet dataset. We therefore require a new paradigm for diagnosing diseases via CNN models. This could be a two stage lesion detection pipeline that involves feature localization followed by classification and further preprocessing steps to segment out pathologies difficult to discern by manual inspection, and finally rebalancing network weights to account for class imbalances seen in medical datasets. Overall, our model goal involves improving detection of mild disease and transitioning to more challenging and beneficial multi-grade disease detection. In summary, this study demonstrates that a sliding window approach using neural networks trained on clinician-selected regions of interest is able to detect subtle pathologic lesions with significantly fewer examples than traditional CNNs. This proposed method is a promising step toward development of screening algorithms for other, less-common retinal diseases.



(a) AUC & Precision/Recall curves for Binary classification



(b) AUC & Precision/Recall curves for Multinomial classification

Figure 2: AUC & Precision/Recall curves of our models

References

1. Gupta A, Cavallerano J, Sun JK, Silva PS. Evidence for telemedicine for diabetic retinal disease. Semin Ophthalmol. 2017;32(1):22-28. doi:10.1080/08820538.2016.1228403

2. Silva PS, Aiello LP. Telemedicine and eye examinations for diabetic retinopathy: a time to maximize real-world outcomes. JAMA Ophthalmol. 2015;133(5):525-526. doi:10.1001/jamaophthalmol. 2015.0333

3. Mansberger SL, Sheppler C, Barker G, et al. Long-term comparative effectiveness of telemedicine in providing diabetic retinopathy screening examinations: a randomized clinical trial. JAMA Ophthalmol. 2015;133(5):518-525. doi:10.1001/jamaophthalmol.2015.1

4. 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;315:2402–2410.

5. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology. 2017;124: 962–969.

6. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318:2211–2223.

7. Litjens, G. et al. . (2017). A Survey on Deep Learning in Medical Image Analysis. arXiv:1702.05747v2 [cs.CV] 4 Jun 2017.

8. J. Yosinski, J. Clune, Y. Bengio. How transferable are features in deep neural networks?. arXiv . 2014;1411.1792v1.

9. C. Szegedy, V. Vanhoucke, S. Ioffe, others . Rethinking the Inception architecture for computer vision. arXiv. 2015;1512.00567v3.

10. Land, E.H., McCann, J.J.: Lightness and retinex theory. J. Opt. Soc. Am. 61, 1–11 (1971)

11. Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-2830.

12. Abramoff MD, Lou Y, Erginay A,et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest Ophthalmol Vis Sci. 2016; 57: 5200– 5206.