Bhatter P, et al. Diabetic Retinopathy Detection Using Collective Intelligence. *Journal of Scientific Innovation in Medicine*. 2020; 3(1): 1. DOI: https://doi.org/10.29024/jsim.47

RESEARCH

Diabetic Retinopathy Detection Using Collective Intelligence

Param Bhatter[1], Emily Frisch[1], Erik Duhaime[2], Anant Jain[2] and Chanel Fischetti[2]

[1] UC Irvine School of Medicine, US

[2] Centaur Labs, US

Corresponding author: Chanel Fischetti (chanel@centaurlabs.com)

Much attention has been focused on describing the utility of artificial intelligence (AI) applied to diabetic retinopathy data. It has been determined that there are ample opportunities for AI algorithms within medicine and that AI is even superior to what we can determine with the professional human eye. However, fewer studies actually have looked at a combined model, or rather, a collective intelligence approach of both human and computer/machine efforts. We attempt to describe and demonstrate the power of collective intelligence in the future of medicine and to offer ways to consider a more complementary approach to both humans and computers.

Keywords: artificial intelligence; medicine; diabetic retinopathy; collective intelligence; ensemble learning

1. Introduction

Diabetic retinopathy is the number one cause of vision loss in the world. An estimated 275 million people in the world have diabetes mellitus, with about 10% having vision threatening diabetic retinopathy [1]. Given the effort needed for massive screening worldwide, numerous research teams are attempting to develop and implement screening programs based on AI. Groups of researchers at companies like Google's DeepMind have worked on using complex AI architectures to build, train, and screen for disease with high sensitivity and specificity [2, 3]. Companies such as IDx have been approved by the FDA in 2018 to screen for diabetic retinopathy in primary care offices [4]. As ML methods improve, and AI technology becomes more accessible worldwide, integrating the technology within healthcare systems for practical use becomes a greater challenge.

These clinical models are often built on a triage based approach [5], referring AI analyzed images with identifiable manifestations of diabetic disease to ophthalmologists for further evaluation. Despite this approach, issues in image quality, diversity of images used to build and validate ML models, patient acceptability, and physician augmentation continue to arise [6–8]. A recent study attempting to evaluate model based grading assistance augmented with well-trained retinal specialists demonstrated the potential to increase overall grading sensitivity without decreasing specificity, though time needed for screening images increased [9].

Given that AI model development is becoming easier and better understood, screening programs in diabetic retinopathy, and medicine in general, are now moving towards understanding how human expertise can be augmented. One approach has been through developments of methods for screening and referring for DR with autonomous, offline phone applications, with no human analysis [10]. While this method has been shown to be effective for screening, we believe human augmentation will be necessary for dealing with diverse data sets to increase diagnostic accuracy and alleviating some of the concerns of "black box" model explainability in the clinical setting. A model's value is not just held in its algorithmic performance, but also in how it classifies disease and explains this to physicians and patients, increasing its overall acceptance [11].

In this paper, we will demonstrate our approach to factoring human opinion in conjunction with AI by emsembing AI models and crowd opinions (which were collected with a crowdsourcing application, DiagnosUs) to investigate our hypothesis: that a collective intelligence approach (human + computer) not only increases model acceptance but is also more accurate than AI or human expertise alone. Ensembling is the method that will allow us to combine these multiple models to perform the same task by way of a singular model with higher diagnostic accuracy [12].

2. Methods and Materials

Data Understanding

Sometimes datasets are initially too large to sort through, so transfer learning models are adopted. Transfer learning is often employed in machine learning [13]. It offers a way to train and utilize a smaller set of data (originating from a larger source) and thereafter an ability to apply it to a larger collection of data in the future.

We used the EyePACS diabetic retinopathy dataset [14]. The dataset contained 35,126 images, which came from different models and types of cameras. This was important because even the way a photo is taken can affect the visual appearance of whether the macula is located on either the left or the right. Some images were shown as one would see the retina anatomically (macula on the left, optic nerve on the right for the right eye). Others were shown as one would see through a microscope condensing lens (i.e. inverted, as one sees in a typical live eye exam). Some images contained artifacts, were out of focus, underexposed, or overexposed. Nonetheless, the dataset was ideal to develop a robust algorithm that would be able to function in the presence of noise and variation.

The following labels were used to scale images and rate the presence of diabetic retinopathy:

0 – No DR
1 – Mild
2 – Moderate
3 – Severe
4 – Proliferative

However, one issue with this original larger dataset was that the labels were not balanced. As you can see in **Figure 1** demonstrating the distribution of labels in the dataset, most of the images were representative of either mild or moderate diabetic retinopathy or none at all. This was the dataset made available through Kaggle. There were 25,810 images labeled as "No diabetic retinopathy" or (0), 2,443 images labeled as "Mild" or (1), 5,292 images labeled as "Modeate" or (2), 873 images labeled as "Severe" or (3), and 703 images labeled as "Proliferative" or (4).

Data Cleaning and Preparation

We partitioned the dataset by taking out 572 images from each labeled class, such that we had two sub-datasets to work with. The first fraction of the dataset contained 2,858 images with 572 images belonging to each class. This fraction was then uploaded to DiagnosUs to accumulate crowd opinions and was later used as the testing set for AI models. **Figure 2** shows the distribution of labels in this subset.

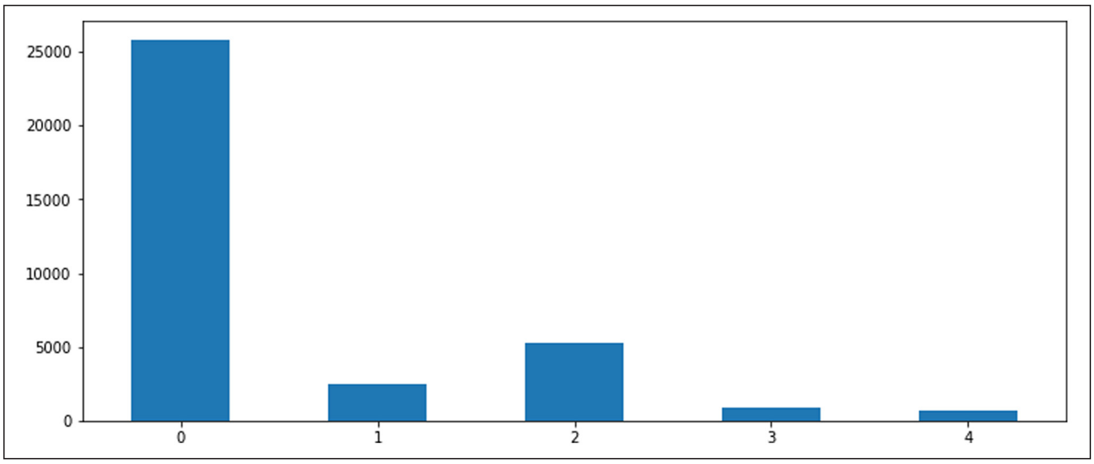


Figure 1: Distribution of labels in EyePACS dataset.

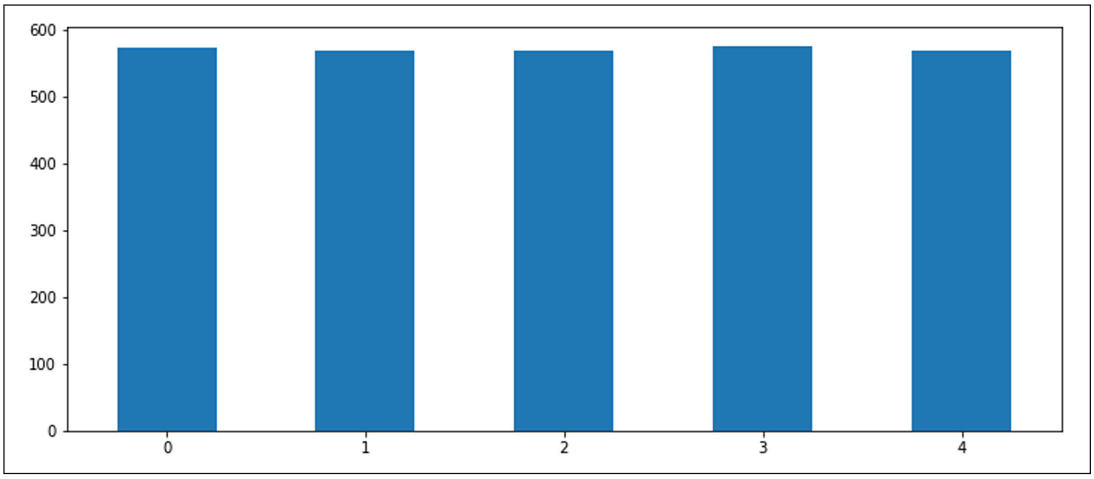


Figure 2: Distribution of labels in the crowd/testing subset.

The rest of the 32,268 images became the other subset, which we used as the training set for our AI models. Because the dataset contained a disproportionately large amount of images without diabetic retinopathy (labeled as 0), compared to other labels, it made sense to randomly throw out ~20,000 of these images to make the distribution less unbalanced. The training fraction finally contained 12,268 images distributed across 5 classes. **Figure 3** shows the distribution of labels in the training subset (about 12,268 images).

3. Crowd Intelligence Results

Exploratory Analysis

Once the subsets were constructed, we employed an application, DiagnosUs, to collect human annotations on these images. DiagnosUs, is freely downloadable on iOS devices through Apple's Application Store in which users are incentivized to annotate medical images. In this study, the "crowd" is defined as a variety of professionals, semi-professionals and untrained lay people. Users can volitionally "compete" for prize money by viewing images and "reading" or labeling them with pre-assigned binary tasks or other numerical tasks, such as diabetic retinopathy. There is significance in the diversity of the crowd.

With a diverse crowd, there are certain strengths of some users that are more apparent compared to the strengths of other users, regardless of education or background [15–17]. This is believed to be because different individuals' errors can cancel out sometimes even the best individual [15–18]. Hong and Page conclude that the diversity of a crowd can also contribute a diversity of intuition and different problem solving abilities that make them more reliable and accurate than a comparative group of specialists [19]. Additionally, the diversity of the crowd, including both the lay public and medical specialists allows for a more collaborative approach to both reproduce and legitimize annotation results, as is demonstrated best in the recent Halabi et al. study [20]. Additionally, the diversity of the crowd also minimizes any inherent bias that could be potentially introduced [16].

It is with this understanding, that DiagnosUs reads were accumulated for a 40 day period from 10/10/2019 to 11/19/2019. Ultimately we obtained 52,412 reads on the 2,858 images subset from the EyePACS dataset, which were evaluated by ~200 users under the topic of Diabetic Retinopathy. Collectively, we assessed and analyzed the accuracy of each user compared to the others and compared to the gold standard of that medical image – an annotation or label from a board-certified physician.

As demonstrated in **Figures 4** and **5**, most of the images from this initial collection had about 14 reads and were each annotated by about 14 unique users. This ensured high confidence in the results and helped in identifying cases which were truly controversial.

Aggregation and Evaluation

All the accumulated reads were aggregated modally by each annotator for each image to get a majority opinion of the annotator for an image, and thereafter all majority opinions by all annotators for an image were modally aggregated to get the majority crowd opinion for that image.

Once we had crowd labels for all 2,858 images, we used Quadratic Weighted Kappa score, a measure of inter-rater reliability [21], to evaluate majority crowd labels against the correct labels. Ultimately, we obtained a kappa score of 0.36 and an AUC score of 0.68.

4. Artificial Intelligence Results

Preprocessing

We used auto-cropping method with a Gaussian filter to increase the robustness (data sufficiently immune to noise) and to enhance the distinctive features in the images. This increased the performance when we trained our AI models.

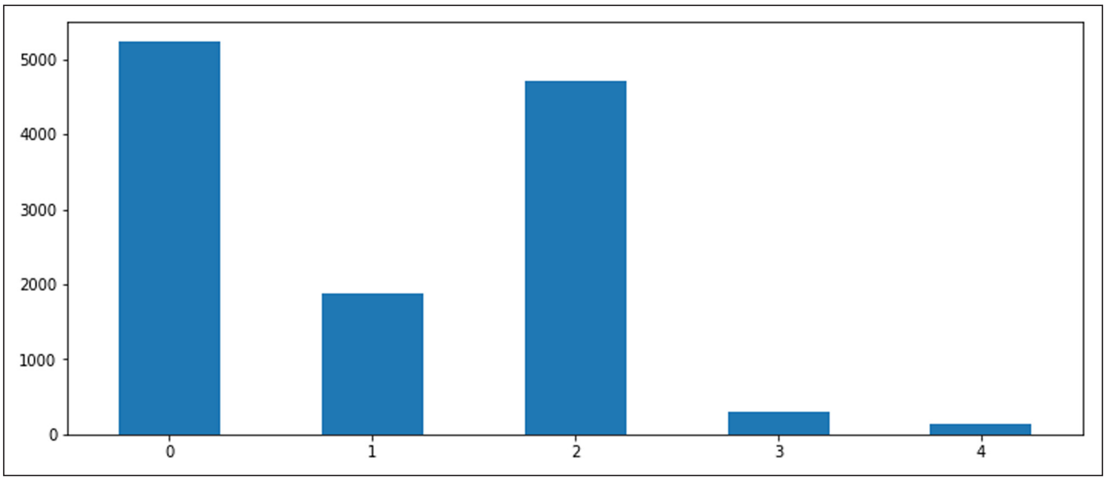


Figure 3: Distribution of labels in the training subset.

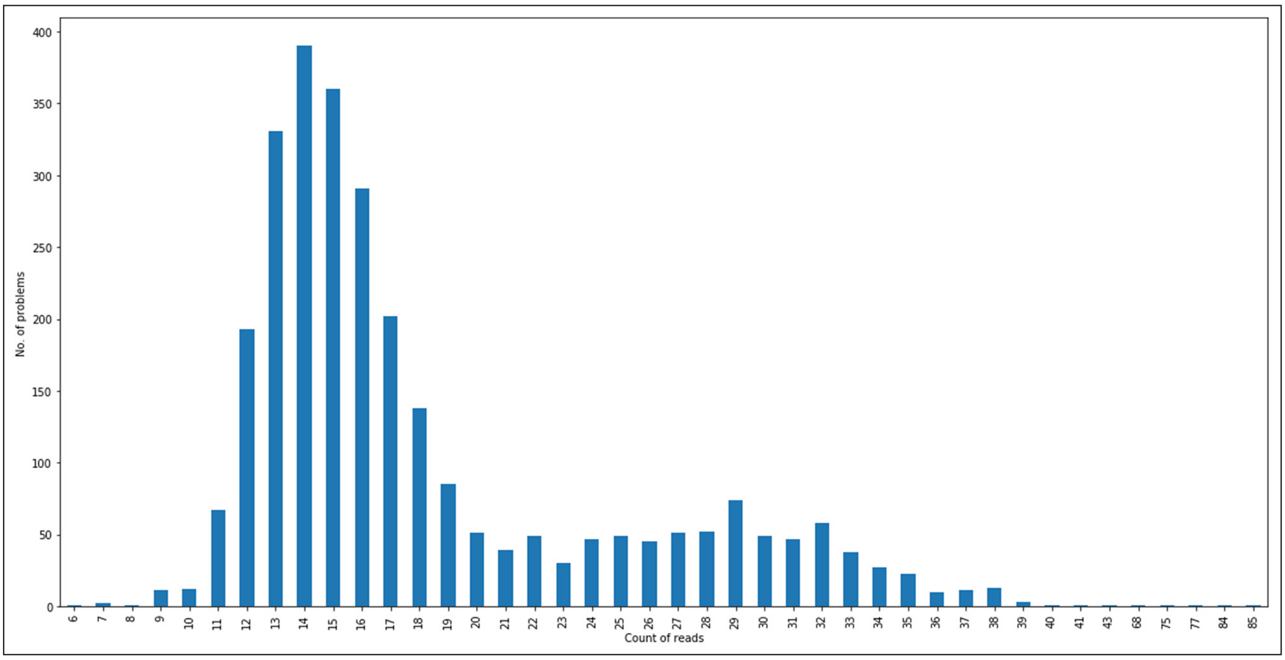


Figure 4: Distribution of counts of reads for images.

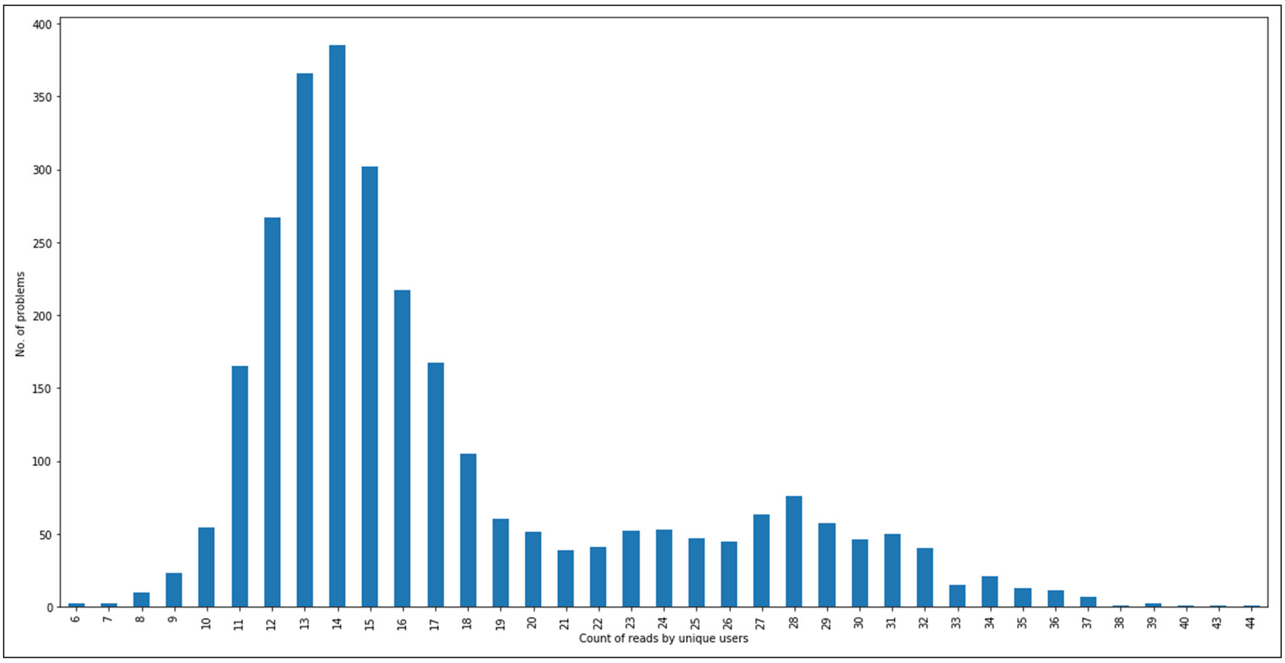


Figure 5: Distribution of counts of reads for cases by unique annotators.

Transfer Learning with EfficientNet

EfficientNet was released this June (2019) by Google AI [22] and is the new state-of-the-art on ImageNet (**Figure 6**). It introduces a systematic way to scale Convolutional Neural Networks (CNNs) in a nearly optimal way.

Since we wanted to optimize the Quadratic Weighted Kappa score we formulated this problem as a regression task. In this way, we were more flexible in our optimization and could yield higher scores than solely optimizing for accuracy. We optimized pre-trained EfficientNetB3, EfficientNetB4 and EfficientNetB5 with a few added layers. Our final AI model was a simple average/ensemble of these three optimized models.

Loss, Augmentations, Pooling

We used Mean Squared Error as the loss function.

Since we did not have a huge amount of data (12,268 images, only 133 of which were labeled 4's), we augmented the images to make the model more robust. We rotated the images on any angle. Also, we flipped the images both horizontally and vertically. Lastly, we divide the image matrices by 128 for normalization.

Additionally, we replaced all the Batch Normalization layers with Group Normalization layers as Batch Normalization becomes unstable with small batch sizes (<16).

We employed RAdam optimizer, since it often yields better convergence than Vanilla Adam optimizer.

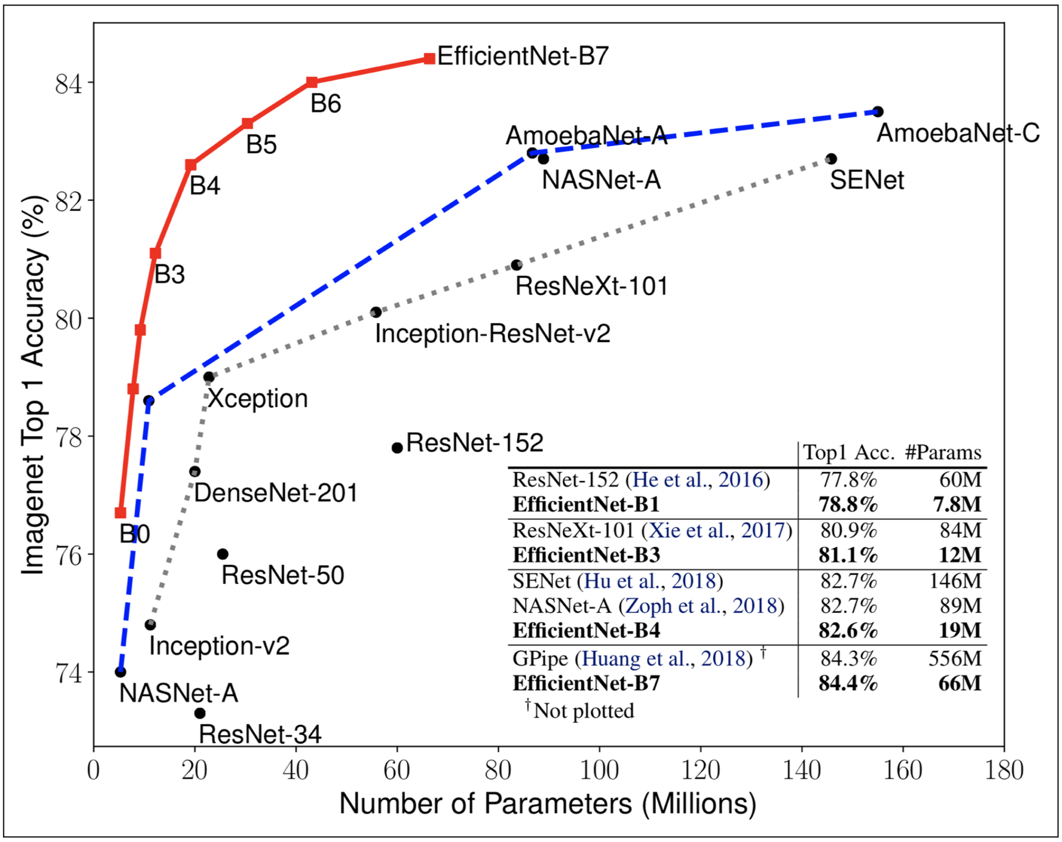
Evaluation

In the end, our evaluation yielded the validation and test weighted kappa quadratic results below (**Table 1**). As visualized, each test set had a different result, but cumulatively, the ensemble together was superior than any set alone.

5. Collective Intelligence Results

Ensembling Crowd Intelligence and AI

Contrary to many of the ways ML, and in particular transfer learning, is represented in most literature- our approach is novel in its employment of collective intelligence (or a crowd approach to image labeling). Often times, models are built on the opinions of only a select one or two professionals, or even via natural language processing (NLP). But with crowd intelligence, one can obtain a diversity of opinions and minimize potential for bias. As demonstrated by **Table 2**, applying ensemble learning to even this collective intelligence approach provides for a synergistic opportunity to maximize both human and machine intelligence [23]. As reflected by Halabi et al. [20], "model heterogeneity is an important aspect of ensemble learning," which includes both the model ensemble diversity, but also the user diversity within each model. It is with this logic that a more optimistic and harmonious future of AI with humans is likely to contribute far more to medicine than either alone.



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**78.8%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.1%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.6%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.4%**	**66M**

†Not plotted

Figure 6: An overview of model architectures and their performance on ImageNet. We can see that EfficientNet achieves state-of-the-art results and uses fewer parameters than most modern CNN architecture [22].

Table 1: Weighted Kappa Artificial Intelligence Evaluation Results.

#	Method	Quadratic Weighted Kappa
1.1	B3 base	0.345
1.2	B4 base	0.399
1.3	B5 base	0.373
1.4	Ensemble (B3+B4+B5)	0.401

Evaluation

Table 2: Collective Intelligence Evaluation Results.

#	Method	Quadratic Weighted Kappa
2.1	Crowd Intelligence	0.36
2.2	Artificial Intelligence	0.40
2.3	Ensemble (Crowd+AI)	0.47

6. Conclusion

AI is transforming the way medicine is practiced. Based on our results, we have demonstrated that a collective intelligence and ensemble approach to labeling medical imaging data is superior to crowd ML or AI alone. But this is not the first time centaur or collective intelligence has been praised within medicine. Poses et al. demonstrated that a group of less experienced residents and fellows were statistically superior in predicting critically ill patient outcomes compared to their attending counterparts [24]. Similar collective intelligence support has been seen in studies with nevi [25], mammograms [26], and even in emergency medicine decision making [15]. More recently Halabi et al. reiterated the importance of ensemble or collective intelligence approaches, as datasets can be replicated and legitimized to continue to improve dataset results [20]. By undertaking a collective intelligence approach, we can optimize the best aspects of both machines and humans, as evidenced by the example provided from our initial analysis of these datasets.

One of the strengths of this collective intelligence approach to AI in medicine is that human error is reduced and the opportunity for the "second opinion" is inherently provided by the "wisdom of the crowd" [26]. We have demonstrated that to achieve a labeling accuracy comparable, if not superior, to a medical professional is through a reliable crowd and well classified dataset. Balanced data is key to achieving validated datasets. This is true for both "easy" and "difficult" cases, and this is where a harmony of generalizability and fine tuning become extremely helpful within collective intelligence.

AI learns by example. Ultimately, the validating data is only as good as the training data. In order to ensure successful results from a collective intelligence approach, the training data must be balanced and high quality. Thus, to replicate a study, we recommend that practitioners take this into account, in the same way any patient would for a single physician's opinion.

As the development of AI progresses, physicians are encouraged to stay up to date on the evolving technologies that impact our patient outcomes, experiences, and perceptions. Effective and consciousness implementation of AI requires the investment and flexibility of the medical community. Medical providers have a unique role within this new era of medicine– and the role will evolve alongside the technology. Through dissemination of understandable content to the medical community, it is our hope that providers can more seamlessly integrate AI into medical care.

7. Limitations

The limitations to this study obviously are most apparent in the inherent bias that can be introduced from any crowd, a problem experienced by other companies and AI algorithms [27] already. Additional bias may be introduced through the incentives we offer to users. As mentioned, users are incentivized to use the application, DiagnosUs, for either personal learning interests or by the money that is set and awarded to the most accurate user.

The other limitations within this study revolve around the data used. As mentioned earlier, the Kaggle dataset employed was initially imbalanced. Minimizing overfitting and bias is possible with appropriately preparing balanced data [28] and using a diverse crowd [16]. Test and validation datasets inherently also help with this process.

This particularly affected our results because the dataset we used required extrication resulting in a significantly smaller training dataset. Because the initial larger Kaggle dataset was substantially biased with labels of 0,1, and 2, we had to account and proportion an appropriately balanced dataset. As a result, our analysis is limited because of the variation and amount of available and equally distributed data.

8. Future Work

The most useful implication from this experiment is its applicability with better balanced datasets. With larger and more balanced data, the results of both the AI and the collective intelligence (built with the ensemble approach of human and machine intelligence) will be far superior.

Lastly, there are many future implications for this study. This approach is not limited to diabetic retinopathy. In fact, there are plenty of opportunities within medicine today that would allow for data analysis that could further improve diagnostic efficiency and accuracy. The implications of what this could mean, both for developed and developing countries has yet to be explored with the advent of this new technology, but is arguably promising and exciting.

Competing Interests

AJ, ED, and CF are employees of Centaur Labs, affiliated with the application, DiagnosUs, which was used as part of the data acquisition for this study. AJ, ED, and CF have shares in Centaur Labs, which supplied the analysis and data for this project. PB and EF are interns with Centaur Labs. The authors disclose no further competing interests.

References

1. **Lee R, Wong TY, Sabanayagam C.** Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and vision.* 2015; 2(1): 17. DOI: https://doi.org/10.1186/s40662-015-0026-2
2. **Gulshan V,** et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama.* 2016; 316(22): 2402–2410. DOI: https://doi.org/10.1001/jama.2016.17216
3. **Gulshan V,** et al. Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India. *JAMA ophthalmology*; 2019. DOI: https://doi.org/10.1001/jamaophthalmol.2019.2004
4. **Abràmoff MD,** et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *Npj Digital Medicine.* 2018; 1(1): 39. DOI: https://doi.org/10.1038/s41746-018-0040-6
5. **Bellemo V,** et al. Artificial Intelligence Screening for Diabetic Retinopathy: the Real-World Emerging Application. *Curr Diab Rep.* 2019; 19(9): 72. DOI: https://doi.org/10.1007/s11892-019-1189-3
6. **Grzybowski A,** et al. Artificial intelligence for diabetic retinopathy screening: A review. *Eye.* 2019: 1–10.
7. **Sayres R,** et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology.* 2019; 126(4): 552–564. DOI: https://doi.org/10.1016/j.ophtha.2018.11.016
8. **Wong TY, Bressler NM.** Artificial Intelligence With Deep Learning Technology Looks Into Diabetic Retinopathy Screening. *JAMA.* 2016; 316(22): 2366–2367. DOI: https://doi.org/10.1001/jama.2016.17563
9. **Abramoff MD, Niemeijer M, Russell SR.** Automated detection of diabetic retinopathy: barriers to translation into clinical practice. *Expert Rev Med Devices.* 2010; 7(2): 287–96. DOI: https://doi.org/10.1586/erd.09.76
10. **Natarajan S,** et al. Diagnostic Accuracy of Community-Based Diabetic Retinopathy Screening With an Offline Artificial Intelligence System on a Smartphone. *JAMA ophthalmology.* 2019; 137(10): 1182–1188. DOI: https://doi.org/10.1001/jamaophthalmol.2019.2923
11. **Ting DSW,** et al. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology.* 2019. 103(2): 167–175. DOI: https://doi.org/10.1136/bjophthalmol-2018-313173
12. **Ju C, Bibaut A, van der Laan M.** The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics.* 2018; 45(15): 2800–2818. DOI: https://doi.org/10.1080/02664763.2018.1441383
13. **Pan SJ, Yang Q.** A survey on transfer learning. *IEEE Transactions on knowledge and data engineering.* 2009; 22(10): 1345–1359. DOI: https://doi.org/10.1109/TKDE.2009.191
14. **Kaggle.** *Diabetic Retinopathy Detection*; 2015. Available from: https://www.kaggle.com/c/diabetic-retinopathy-detection/data.
15. **Kämmer JE,** et al. The potential of collective intelligence in emergency medicine: pooling medical students' independent decisions improves diagnostic performance. *Medical decision making.* 2017; 37(6): 715–724. DOI: https://doi.org/10.1177/0272989X17696998
16. **Page SE.** The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies-New Edition; 2008. Princeton University Press. DOI: https://doi.org/10.1515/9781400830282
17. **Page SE.** Making the difference: Applying a logic of diversity. *Academy of Management Perspectives.* 2007; 21(4): 6–20. DOI: https://doi.org/10.5465/amp.2007.27895335
18. **Krause S,** et al. Swarm intelligence in humans: diversity can trump ability. *Animal Behaviour.* 2011; 81(5): 941–948. DOI: https://doi.org/10.1016/j.anbehav.2010.12.018
19. **Hong L, Page SE.** Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences.* 2004; 101(46): 16385–16389. DOI: https://doi.org/10.1073/pnas.0403723101
20. **Halabi SS,** et al. The RSNA pediatric bone age machine learning challenge. *Radiology.* 2018; 290(2): 498–503. DOI: https://doi.org/10.1148/radiol.2018180736
21. **McHugh ML.** Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica.* 2012; 22(3): 276–282. DOI: https://doi.org/10.11613/BM.2012.031
22. **Tan M, Le QV.** EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv preprint arXiv:1905*,11946: 2019.
23. **Dietterich TG.** Ensemble methods in machine learning. In *International workshop on multiple classifier systems*; 2000. Springer. DOI: https://doi.org/10.1007/3-540-45014-9_1
24. **Poses RM,** et al. Are Two (Inexperienced) Heads Better Than One (Experienced) Head?: Averaging House Officers' Prognostic Judgments for Critically Ill Patients. *Archives of internal medicine*, 1990; 150(9): 1874–1878. DOI: https://doi.org/10.1001/archinte.1990.00390200068013
25. **King AJ,** et al. Skin self-examinations and visual identification of atypical nevi: Comparing individual and crowdsourcing approaches. *Cancer epidemiology.* 2013; 37(6): 979–984. DOI: https://doi.org/10.1016/j.canep.2013.09.004
26. **Wolf M,** et al. Collective Intelligence Meets Medical Decision-Making; 2015.
27. **Crawford K.** Artificial intelligence's white guy problem. *The New York Times*; 2016: 25.
28. **Japkowicz N, Stephen, S.** The class imbalance problem: A systematic study. *Intelligent data analysis.* 2002; 6(5): 429–449. DOI: https://doi.org/10.3233/IDA-2002-6504

How to cite this article: Bhatter P, Frisch E, Duhaime E, Jain A, Fischetti C. Diabetic Retinopathy Detection Using Collective Intelligence. *Journal of Scientific Innovation in Medicine*. 2020; 3(1): 1. DOI: https://doi.org/10.29024/jsim.47

Submitted: 03 December 2019 **Accepted:** 22 January 2020 **Published:** 28 April 2020

Levy Library Press

Journal of Scientific Innovation in Medicine is a peer-reviewed open access journal published by Levy Library Press.

OPEN ACCESS