Improving Medical Image Decision Making by Leveraging Metacognitive Processes and Representational Similarity

Eeshan Hasan, (eeshan.hasan@vanderbilt.edu)
Jennifer S. Trueblood (jennifer.s.trueblood@vanderbilt.edu)
Department of Psychology, Vanderbilt University, Nashville, TN 37240, USA

Quentin Eichbaum (quentin.eichbaum@vumc.org)
Adam Seegmiller (adam.seegmiller@vumc.org)
Charles Stratton (charles.stratton@vumc.org
Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN 37232, USA

Abstract

Improving the accuracy of medical image interpretation is critical to improving the diagnosis of many diseases. Using both novices (undergraduates) and experts (medical professionals), we investigate methods for improving the accuracy of a single decision maker by aggregating repeated decisions from an individual in different ways. Our participants made classification decisions (cancerous versus non-cancerous) and confidence judgments on a series of cell images, viewing and classifying each image twice. We first applied the maximum confidence slating algorithm (Koriat, 2012), which leverages metacognitive ability by using the most confident response for an image as the 'final response'. We also examined algorithms that aggregated decisions based on image similarity, leveraging neural network models to determine similarity. We found maximum confidence slating improves classification accuracy for both novices and experts. However, aggregating responses on similar images improves classification accuracy for novices and not experts, suggesting differences in the decision mechanisms of novices and experts.

Keywords: Medical Image Decision Making, Computational Modeling, Neural Networks, Representation, Concepts and Categories, Wisdom of the Crowds, Metacognition

Introduction

Accurate interpretation and classification of medical images is an important step in the diagnosis and treatment of numerous diseases. Despite specialized training and advances in technology, diagnostic errors still occur. One approach to reducing errors is through multiple readings, where the judgments of several medical experts are combined. For example, the misclassification rate decreased from 24.7% to 18.1% in breast histopathology with multiple readings (Elmore et al., 2016). However, multiple readings are not consistently performed in the United States because it is time-consuming and the additional readings are not reimbursed (Waite et al., 2017). In other parts of the world, there is a dearth of pathologists (Nelson et al., 2018), making second opinions difficult if not impossible.

In this paper, we consider whether it is possible for the same individual to act as a second pair of eyes in a series of repeated decisions about medical images. We leverage recent research on the "wisdom of the inner crowd" to reduce errors at the individual level. According to the wisdom of crowds principle, improvements in accuracy are obtained by combining the judgments of different individuals (Surowiecki, 2005). Research on the "crowd within" applies this same idea, but to a single individual who performs repeated judgments.

In our task, participants categorize images of white blood cells as cancerous (i.e., 'blast' cells) or non-cancerous (i.e., 'non-blast' cells). Participants make two separate decisions for each image. We examine both experts (i.e., medical professionals) as well as novices (i.e., undergraduate students). We use novice participants in addition to medical experts for two important reasons. First, data from novice participants provides a baseline for comparing expert participants. Second, there is recent interest in using novices to assist with medical image diagnosis. Particularly relevant for this paper is the possibility of crowdsourcing large numbers of untrained individuals to perform simple diagnostic tasks (Ørting et al., 2020).

We explore various algorithms for aggregating these decisions with the aim of improving individual accuracy. One successful "wisdom of the crowd" algorithm for binary decision-making is the maximum confidence slating algorithm (Koriat, 2012). In this algorithm, one considers the more confident response in a pair of responses made by an individual as their final response. The success of this algorithm hinges on the metacognitive ability of individuals to produce confidence judgments that accurately capture their performance on the task. (Yeung & Summerfield, 2012).

In addition to the maximum confidence slating algorithm, we also explore a set of aggregation algorithms that leverage tools from machine learning and artificial intelligence. Specifically, we propose a set of algorithms that determine final decisions by aggregating an individual's responses on similar images. We use latent representations obtained by convolutional neural networks to calculate the similarity between images. In this paper, we look at two representations, one with general visual features (He, Zhang, Ren, & Sun, 2015) and another one with features that are well tuned to the task at hand (Holmes, O'Daniels, & Trueblood, 2020).

Besides using these algorithms to reduce errors, we will also use these techniques to probe the differences in decision processes of novices and experts. For example, the metacognitive abilities of experts might be better than novices. However, aggregating decisions over similar images might help 'de-noise' novice decisions; but have little impact on expert decisions. Experts might give the same (correct or incorrect) decision for similar images due to systematic biases (or incorrect decision rules) rather than a noisy decision process.

Methods

Participants

We conducted two experiments on undergraduates (novices) at Vanderbilt University and one experiment on medical professionals (experts) at the American Society for Clinical Pathology (ASCP) annual conference held in Baltimore, Maryland in October 2018. All experiments were approved by the Institution Review Board at Vanderbilt University.

A total of 87 undergraduates participated in our experiments, 45 in the Experiment 1a and 42 in Experiment 1b. The sample size was based off of similar studies examining pathology image-based decision-making (Trueblood et al., 2018). 23 pathologists and laboratory professionals participated in Experiment 2. Participants received a $10 Starbucks gift card for participating. The sample size for this experiment was based off of convenience.

The participants primarily identified as female, (Exp. 1a: 76%; Exp. 1b: 70%; Exp 2: 73%). The mean age was 18.9 years (SD=1.2; IQR $18-24$) for Experiment 1a, 19.5 years (SD=2.5; IQR-$18-20$) for Experiment 1b, and 42.4 years (SD=13.5; IQR $30-56$) for Experiment 2.

Materials

The set of stimuli were identical to Trueblood et al. (2018), consisting of 300 digital images of Wright-stained white bloods cells taken from anonymized patient peripheral blood smears at Vanderbilt University Medical Center (VUMC). Examples of these images can be seen in Figure 1. The images were taken by the CellaVision DM96 (CellaVision AB, Lund, Sweden), an automated digital cell morphology instrument. The 300 images consisted of 150 "blast" cell images and 150 "non-blast" cell images. Within these two categories, half of the images were "easy" and half were "hard". Since the 'ground truth' for the image classes was not known, the image classifications (i.e., blast / non-blast) and difficulty ratings (i.e., easy / hard) were based on identification and rating data from three hematopathology faculty from the Department of Pathology at VUMC. The images that were used in the experiment were the ones that all three sub-specialists agreed upon the classification. More details on the rating procedure and image curation can be found in Trueblood et al. (2018).

Procedure

In the experiments, participants gave two categorization responses on the white blood cell images along with their confidence after a brief training phase.

Novices - Experiments 1a & 1b In the novice experiments, participants first completed a familiarization block, training block, and practice trials before starting the main task. The four cell types (blast / non-blast x easy / hard) were counterbalanced in each of these initial blocks. In the 36 familiarization trials, participants viewed cell images with their corresponding labels (either blast or non-blast) one at a time for as long as they wanted. In the 60 training trials, participants viewed two cell images and their task was to select the image that matched a label (either blast or non-blast) at the top of the screen. They received feedback in these training trials. Finally, in the 20 practice trials, they were instructed to indicate whether the cell was a blast cell or a non-blast cell and received feedback.

The main task consisted of two parts, each with 300 trials corresponding to the 300 unique images contained in the stimuli set. Across the two parts of the main task, participants saw a total of 600 images, so that each image was shown twice. On each trial, participants were shown a single image and had to decide if it was a blast or non-blast cell. In addition to making a choice on each trial, participants were also instructed to report how confident they were that they selected the correct response on a scale ranging from 50% (guessing) to 100% (certain correct). In Experiment 1a, they were asked "Is this cell cancerous?" for the first part of the main trials but were asked "Is this cell non-cancerous?" for the second part. In Experiment 1b, the second block was the same as the first (i.e., they were asked "Is this cell cancerous?") for the entire main task. Experiment 1a had 20 practice trials in between both parts of the main trials to help with the transition in instructions. In addition, the images in the second half of the main task of Experiment 1a were rotated 180 degrees. Images were not rotated in Experiment 1b.

Expert - Experiment 2 Experiment 2 with experts was a shorter version of Experiment 1a with novices. This was due to time constraints at the ASCP conference. So that the task was not too easy, expert decisions were only collected for the hard cell types. Since they already had experience with the cells, their training phase was also shortened, consisting of 20 trials of hard cells counterbalanced among blast and non-blast. They received feedback in these trials. The main task consisted of two parts, each containing 60 images. Similar to Experiment 1a, both the main blocks had the same images. After each decision, expert participants were also asked to indicate their confidence in their decision. They did not receive feedback in these trials. Similar to Experiment 1a, there was a change in instruction between the two parts of the main task and the images were rotated 180 degrees in the second half of the main trials.

Behavioral Results

Participants were excluded if their accuracy on the practice trials was less than or equal to 50%. We also excluded participants who gave more than 50 confidence ratings outside the valid range $(50-100)$. We also removed participants that gave the same response for more than 95% of the trials in either parts of the main task. After these exclusions, we retained 34 out of 45 participants in Experiment 1a and 31 out of 42 participants in Experiment 1b. 1 out of the 23 experts was excluded because they did not provide any confidence ratings on the first part of the main task.

For Experiments 1a and 1b, the mean accuracy was 66.1% (SD=8.8; IQR $60.1\%-71.5\%$) and 66.5% (SD=10.7; IQR

59.5% − 74.8%), respectively. The mean accuracy of the experts was 71.6% (SD=14.3; IQR 60.1 − 83.9). Since the stimuli were different for novices and experts, we also report the novice accuracy on the subset of stimuli seen by experts in Experiment 2. On this reduced set of images, the mean accuracy was 61.7% (SD=10.7; IQR 53.3% − 69.2%) for Experiment 1a and 59.0% (SD=9.8; IQR 51.3% − 63.3%) for Experiment 1b. Hence, as expected, experts perform better than novices.

As participants gained experience with the task, it is possible that they changed the way that they used the confidence scale over the course of the experiment. Additionally, in Experiments 1a and 2, the instructions changed between the two parts of the main task. To determine if confidence ratings for the two parts of the main task came from the same distribution, we conducted a Kolmogorov–Smirnov (KS) test. We observed that for 18 out of 34 participants in Experiment 1a, 18 out of 31 participants in Experiment 1b, and 6 out of 22 participants Experiment 2, the distribution of confidence ratings in the two parts of the main task were significantly different ($p < 0.05$). Hence, we normalised the confidence ratings for the two parts of the main task separately. In other words, we calculated the z-score of the confidence ratings for each person separately for each part of the main task.

We also examined how confidence was related to accuracy, image type, and difficulty across all trials in the main task. We conducted a 2 (accuracy: correct, incorrect) x 2 (classification: blast, non-blast) x 2 (difficulty: easy, hard) repeated measures ANOVA for the novices and a 2 (accuracy: correct, incorrect) x 2 (classification: blast, nonblast) repeated measures ANOVA for experts. We observed a significant main effect for accuracy (Exp. 1a: $F(1,33) = 97.9$ $p < 0.0001$; Exp. 1b: $F(1,30) = 71.8$, $p < 0.0001$; Exp. 2: $F(1,21) = 36.6$, $p < 0.0001$) and a main effect of classification (Exp. 1a: $F(1,33) = 34.7$, $p < 0.0001$; Exp. 1b: $F(1,30) = 19.5$, $p = 0.0001$) for novices but no effect for experts (Exp. 2: $F(1,21) = 0.0$, $p = 0.9062$). We also found a main effect of difficulty in Exp. 1a ($F(1,33) = 7.2$, $p = 0.0114$), but not for the Exp. 1b ($F(1,30) = 2.3$, $p = 0.1407$). We also found significant interactions between classification and difficulty in both the novice experiments. In sum, participants gave higher confidence ratings when they were accurate, showing that confidence reflects accuracy, which is critical for the maximum confidence slating algorithm discussed below.

Modeling Methods

As mentioned above, we will explore the possibility of improving the performance of a single individual by aggregating their responses. The algorithms are described in detail in the following sections:

Maximum Confidence Slating Algorithm (MCS)

The maximum confidence slating algorithm uses the two classification decisions for each image along with the two confidence ratings for the image. First, we normalise the confidence ratings as described in the behavioral results section. For each image, we use the more confident classification as the final response on that image.

k Nearest Neighbor (kNN) on Latent Representations

The remaining algorithms attempt to improve individual performance by aggregating the decisions made on similar images. In these algorithms, we first have to calculate the similarity between two images and then use a k Nearest Neighbor (kNN) imputation. Figure 2 provides examples of where this approach might be useful as well as fail. To calculate the similarity between images, we use the Euclidean distance on two representational spaces.

Unsupervised Representation It has been suggested that useful high level visual features for a task can be extracted from neural networks that have been trained on other tasks. To this end, we use a GoogLeNet that was trained on the dataset from ImageNet Large-Scale Visual Recognition Challenge (2014) with objects that are commonly encountered in everyday life (He et al., 2015). We removed the last classification layer and passed every image through the network (Figure 1 top row). The model was not trained on the blast task. As shown in Figure 2, the classes are slightly separated but also overlap in this representation.

Supervised Representation For the supervised representation, we followed the procedure in Holmes et al. (2020). A GoogLeNet trained on ImageNet was additionally trained on the blast task using transfer learning (Figure 1 bottom row). A larger set of 606 images which contained the 300 images used in our experiment was used to train the network. The accuracy of the network was 94% on the validation set and 98% on the training set. This shows that the network did not overfit the images used in the experiment and effectively generalised to novel images. As shown in Figure 2, the classes are distinctly separated in the representation.

kNN imputation For every image, we use the k nearest neighbors to calculate the final response. That is, we examined the k responses on its nearest neighbors. The final decision on the image was taken to be the modal (the most common) decision on all of these k decisions. This included the two decisions on the image in question. Unlike the MCS algorithm, this does not use participants' confidence judgments.

We consider two values of k: $k = 3$ and $k = 7$. When $k = 3$, for a given (target) image, we look for 3 decisions on the most similar images. The first two decisions will be the two separate decisions made on the target image. For the third decision, we randomly pick one of the two decisions on the most similar image to the target image. In the case where the two decisions on the target image are the same - say blast, then the third decision will not be able to overturn the decision on the target image. However, suppose that a person made two different decisions on the target image, then the third decision will be able to break the tie. Therefore, using $k = 3$ amounts

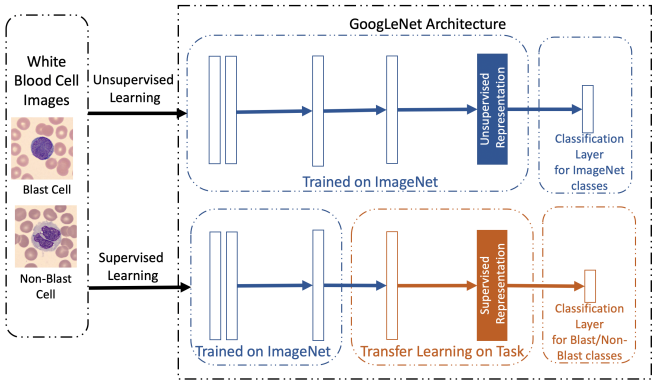

Figure 1: Schematic of the supervised and unsupervised representations. The unsupervised representation was obtained by using GoogLeNet trained on ImageNet. The supervised representation was obtained by using transfer learning on a GoogLeNet trained on ImageNet.

to using one of the judgments on the nearest image to break an inconsistent response on the target image. In no case will it be able to overturn a consistent judgment on a given image. For this algorithm to be successful, with $k = 3$, we need the decision on the nearest image to be better at breaking ties than chance.

In this paper, we also consider $k = 7$, which amounts to using the 7 nearest decisions. In this case, suppose that both of the decisions made on the target image are blast. However, on the 5 remaining decisions ($2 * 2 = 4$ responses from the 2 most similar images and 1 response randomly chosen from the next most similar image), the participant responded non-blast, then the modal response on the set would be non-blast. This is an example where the other decisions can actually overturn the decision made on the target image.

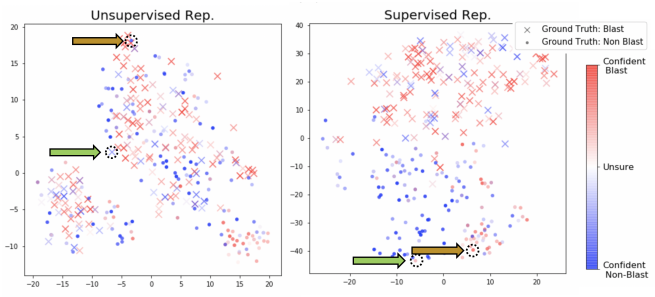

Figure 2: Example of a representative participant's judgments in the unsupervised (left) and supervised (right) latent spaces. The green (gold) arrows illustrate situations where the neighbors point to the correct (incorrect) answer. For example, in the right panel, both the arrows point to non-blast cells that were judged to be blast. In this panel, the green (gold) arrow shows an example where the neighbors were correctly (incorrectly) judged to be non-blast (blast).

Modeling Results

We applied the 5 algorithms to the data. The average performance of each algorithm is in Table 1. A repeated-measures ANOVA showed that there was a significant difference in the performance of the algorithms: Exp. 1a: $F(5,165) = 35.5, p < 0.0001$; Exp. 1b: $F(5,150) = 32.4, p < 0.0001$; Exp. 2: $F(5,105) = 17.9, p < 0.0001$. In the following sections, we present post-hoc t-tests comparing the performance of the algorithms. To control for multiple comparisons, we use the Bonferroni correction to the p-value, setting $p = 0.05/15 = 0.003$. The post-hoc tests are summarized in Table 2.

Table 1: The average performance of each algorithm. The best performing algorithm for each experiment is in bold.

Algorithm	Exp. 1a	Exp. 1b	Exp. 2
Average Response	66.1%	66.5%	71.6%
Max. Conf. Slating	67.4%	67.4%	**73.8%**
Unsupervised k=3	67.0%	67.1%	73.0%
Unsupervised k=7	64.1%	64.4%	62.1%
Supervised k=3	69.2%	68.4%	72.9%
Supervised k=7	**71.0%**	**70.6%**	71.3%

MCS versus Average Performance

The MCS algorithm uses a participant's most confident response as their final response. We first compared the accuracy from MCS to the average accuracy, which is the mean accuracy across both responses. The mean accuracy would be 1(0) for an image where the two responses are correct (incorrect) and consistent. For a cell with inconsistent responses, it would be 0.5. The mean accuracy of an individual is the mean accuracy over all the cell images. As shown in Table 1, the mean MCS accuracy increases by 1.3% to 67.4%, 0.9% to 67.4%, and 2.2% to 73.8% for Exp. 1a, 1b and 2 respectively. As shown in Table 2, in post-hoc tests comparing MCS to average performance, the difference was significant for Exp. 1a ($t(33) = -3.9, p = 0.0004$). However, this is only marginally significant for Exp. 1b ($t(30) = -2.8$, $p = 0.0078$) using the Bonferroni correction to the p-value. Finally, it is significant for Exp. 2 ($t(21) = -3.5, p = 0.0020$).

Latent Representations versus Average Performance

Next, we compared the performance of the latent representation algorithms to average performance using the post-hoc tests mentioned above. As seen in Table 1, for the unsupervised representation at $k = 3$, we observe an improvement in performance for both of the novice experiments (Exp. 1a: M= 67.0%, Exp. 1b: M=67.1%). As shown in Table 2, the post-hoc tests show that this improvement in performance is significant for the first experiment but not the second (Exp. 1a: $(t(33) = -3.8, p = 0.0005$ Exp. 1b: $t(30) = -1.6$, $p = 0.1202$) with the Bonferroni correction to the p-value. We also see a slight increase in performance for the experts

Table 2: Results of the post-hoc t-tests comparing all the algorithms to each other. The values in bold are significant using the Bonferroni corrected p-value (p=0.003).

Algorithm 1	Algorithm 2	Experiment 1a		Experiment 1b		Experiment 2	
		t	p	t	p	t	p
Average Response	Max. Conf. Slating	**-3.9199**	**0.0004**	-2.8497	0.0078	**-3.5270**	**0.0020**
Average Response	Unsupervised k=3	**-3.8415**	**0.0005**	-1.5996	0.1202	-2.3375	0.0294
Average Response	Unsupervised k=7	**3.2456**	**0.0027**	**4.2717**	**0.0002**	**5.9194**	**0.0000**
Average Response	Supervised k=3	**-8.4720**	**0.0000**	**-5.3118**	**0.0000**	-1.4990	0.1488
Average Response	Supervised k=7	**-7.3830**	**0.0000**	**-5.8774**	**0.0000**	0.1880	0.8527
Max. Conf. Slating	Unsupervised k=3	0.8544	0.3990	0.8943	0.3783	0.7644	0.4531
Max. Conf. Slating	Unsupervised k=7	**4.2489**	**0.0002**	**5.0598**	**0.0000**	**6.8010**	**0.0000**
Max. Conf. Slating	Supervised k=3	**-4.0949**	**0.0003**	-2.5250	0.0171	0.7812	0.4434
Max. Conf. Slating	Supervised k=7	**-5.0500**	**0.0000**	**-4.4756**	**0.0001**	1.4466	0.1628
Unsupervised k=3	Unsupervised k=7	**5.2025**	**0.0000**	**5.4419**	**0.0000**	**6.2189**	**0.0000**
Unsupervised k=3	Supervised k=3	**-5.8200**	**0.0000**	**-4.4757**	**0.0001**	0.0712	0.9439
Unsupervised k=3	Supervised k=7	**-6.1784**	**0.0000**	**-5.6123**	**0.0000**	1.1530	0.2619
Unsupervised k=7	Supervised k=3	**-7.5460**	**0.0000**	**-6.8833**	**0.0000**	**-5.2149**	**0.0000**
Unsupervised k=7	Supervised k=7	**-7.8765**	**0.0000**	**-10.1618**	**0.0000**	**-3.9313**	**0.0008**
Supervised k=3	Supervised k=7	**-3.9525**	**0.0004**	**-4.1524**	**0.0003**	1.2251	0.2341

(Exp. 2: M=73.0%), which was not significantly different ($t(21) = -2.3$, $p = 0.0294$ from average performance with the Bonferoni correction. For the unsupervised representation at $k = 7$, there is a consistent significant decline in performance for all three experiments (Exp. 1a: 64.1%, $t(33) = 3.2$, $p = 0.0027$; Exp. 1b: 64.4%, $t(30) = 4.3$, $p = 0.0002$; Exp. 2: 62.1%, $t(21) = 5.9$, $p < 0.0001$). Note that this representation relied only on general visual features and not features specific to the task.

As seen in the Tables 1 and 2, for the supervised representation at $k = 3$ and $k = 7$, we see a pattern that is similar to the unsupervised representation at $k = 3$. The post-hoc tests show that there is a significant increase in performance for both novice experiments (Exp. 1a, $k = 3$: $M = 69.2\%$, $t(33) = -8.5$, $p < 0.0001$; Exp. 1a, $k = 7$: 71.0%,$t(33) = -7.4, p < 0.0001$; Exp. 1b, $k = 3$: $M = 68.4\%, t(30) = -5.3$, $p < 0.0001$; Exp. 1b, $k = 7$: $M = 70.6\%$, $t(30) = -5.9$, $p < 0.0001$) with the Bonferroni correction to the p-value. However, this improvement is small and insignificant for experts (Exp. 2, $k = 3$: $M = 72.9\%$ $t(21) = -1.5$, $p = 0.1488$; Exp. 2, $k = 7$: $M = 71.3\%$ $t(21) = 0.2$, $p = 0.8527$). These results indicate that the latent representation algorithms are effective for the novices but not for experts.

Next, we examine whether the quality of representation or number of neighbors affects the efficacy of the algorithm, especially for novices. We used the post-hoc tests to compare the supervised and unsupervised representation at $k = 3$. For both of the experiments, as shown in Table 2, we observe that the performance is significantly better for the supervised than the unsupervised representation (Exp. 1a: $t(33) = -5.8$, $p < 0.0001$; Exp. 1b: $t(30) = -4.5$, $p < 0.0001$), showing that supervised representations are better than unsupervised representations.

We will now compare the algorithms at $k = 3$ and $k = 7$. We already know that the unsupervised representation at $k = 7$ is worse than average performance. However, the pattern is reversed for the supervised representation at $k = 7$. As shown in Table 2, the improvement in performance with $k = 7$ was significant for both of the experiments with novices (Exp. 1a: $t(33) = -4.0$, $p = 0.0004$; Exp. 1b: $t(30) = -4.2$, $p = 0.0003$). These results show that it is particularly useful to aggregate over several responses and possibly overturn the original decision only when the representation is well tuned to the task.

In sum, for the latent representations applied to the novice experiment 1a, we observe that the supervised representations are the best with $k = 7$, outperforming $k = 3$. After the supervised representation algorithms, we observe that the unsupervised representation at $k = 3$ still outperforms average performance. Finally, we see that the unsupervised representation performs the worst at $k = 7$. The pattern is similar for novice Experiment 1b. Most of these comparisons are not significant for experts.

Comparing MCS to Latent Representations

We now compare the latent representations to MCS. For the unsupervised representations, at $k = 3$, the performance is similar to MCS for all experiments. The post-hoc tests indicate that the difference is not significant (Exp. 1a: $t(33) = 0.9$, $p = 0.3990$; Exp. 1b: $t(30) = 0.9$, $p = 0.3783$; Exp. 2: $t(21) = 0.8p = 0.4531$). The supervised representation at $k = 3$ outperforms MCS for the novices, but not for the experts, where the difference is insignificant (Exp. 1a: $t(33) = -4.1p = 0.0003$; Exp. 1b: $t(30) = -2.5$, $p = 0.0171$; Exp. 2: $t(21) = 0.8$, $p = 0.4438$). The pattern is the same for $k = 7$.

It is especially of interest to compare the algorithms using latent representations at $k = 3$ with MCS. This is because both algorithms use different ways of resolving the conflict

when decisions for the same image differ, but have no effect when responses are consistent. MCS relies on metacognitive judgments (i.e., response confidence) whereas the latent representation algorithms use the similarity structure of the underlying problem.

Comparing Novices to Experts

As mentioned in the Methods, the experts provided judgments for 60 hard images compared to the 300 easy and hard images for novices. This might influence the efficacy of the latent representation algorithms that depend on image similarity. Because there were fewer images in the expert experiment, the average nearest neighbor is necessarily less similar than the novice experiments. Since we are interested in comparing the results for novices and experts, we also apply the best algorithm on the novice experiments (i.e., supervised representation with $k = 7$) to the restricted set of 60 images seen by experts.

On these images, the supervised representation with $k = 7$ resulted in mean accuracy of 67.5% for Exp. 1a and 63.2% for Exp. 1b, which was greater than average performance of 61.8% and 59.0%, respectively. Pairwise t-tests, showed this increase was significant (Exp. 1a: $t(33) = -4.8$, $p < 0.0001$; Exp. 1b: $t(30) = -5.4$, $p < 0.0001$). Hence, the algorithms with latent representations seem to be effective for novices, but not experts even when restricted to exactly the same image set.

Discussion

In this paper, we explored different methods for aggregating repeated decisions from the same individual with the aim of improving medical image decision-making. To evaluate the accuracy of these algorithms, we used the stimuli that three sub-specialists agreed upon. Since these experts specialize in interpreting white blood cells, we expect their judgments to be more accurate than the expert participants used in Experiment 2, who were laboratory professionals and pathologists from many different areas of pathology.

The MCS algorithm works by exploiting people's metacognitive processes, namely their ability to judge the accuracy of their responses (Yeung & Summerfield, 2012). For the MCS algorithm to be successful, we need the differences in metacognitve information obtained at different times or through different question framings to be indicative of accuracy. We found that the MCS algorithm improved performance in all of our experiments, suggesting that confidence judgments can meaningfully solve the conflict of inconsistent decisions. We note that the effect is more prominent in Experiment 1a than Experiment 1b, suggesting that changing the question framing might result in more diverse confidence judgments, which is a necessary condition for wisdom of the crowds (Surowiecki, 2005). Beyond decision aggregation, our results suggest that metacognitive processes might be useful aids in decision making. Awareness of these processes might change and improve the quality of decision making even without a MCS algorithm (Boldt, Schiffer, Waszak, & Yeung, 2019).

Regarding the latent representations, we observed that aggregating decisions based on image similarity improved performance for novices. This was true for representations derived from both unsupervised and supervised neural network models with the supervised models providing the best performance of all algorithms tested for novice Experiment 1a. The novice Experiment 1b had a similar pattern but with smaller improvements, suggesting once again that changing the question framing might result in more diverse responses and interact with the aggregation algorithms. However, aggregating similar responses resulted in no improvement in the performance of experts even when the latent representation was informative and well tuned. This suggests that experts are more likely to make the same decision on similar images. That is, their decision might be biased towards the wrong answer in that portion of the latent space. On the other hand, for novices, we see substantial improvement with the latent representations suggesting that novices might be making decisions using a more random and noisy process as observed in Trueblood et al. (2018). These results suggest that using image similarity is a meaningful way to de-noise the decisions of novices.

Acknowledgement

This work was supported by a Clinical and Translational Research Enhancement Award from the Department of Pathology, Microbiology, and Immunology, Vanderbilt University Medical Center. This work was also supported by NSF grant 1846764. We thank Payton O'Daniels for his excellent research assistance.

References

Boldt, A., Schiffer, A.-M., Waszak, F., & Yeung, N. (2019). Confidence predictions affect performance confidence and neural preparation in perceptual decision making. *Scientific reports*, *9*(1), 1–17.

Elmore, J. G., Nelson, H. D., Pepe, M. S., Longton, G. M., Tosteson, A. N., Geller, B., ... Weaver, D. L. (2016). Variability in pathologists' interpretations of individual breast biopsy slides: A population perspective. *Annals of internal medicine*, *164*(10), 649–655.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026–1034).

Holmes, W. R., O'Daniels, P., & Trueblood, J. S. (2020). A joint deep neural network and evidence accumulation modeling approach to human decision-making with naturalistic images. *Computational Brain & Behavior*, *3*(1), 1–12.

Koriat, A. (2012). When are two heads better than one and why? *Science*, *336*(6079), 360–362.

Nelson, A. M., Hale, M., Diomande, M. I. J.-M., Eichbaum, Q., Iliyasu, Y., Kalengayi, R. M., ... Sayed, S. (2018). Training the next generation of african pathologists. *Clinics in Laboratory Medicine*, *38*(1), 37–51.

Ørting, S. N., Doyle, A., van Hilten, A., Hirth, M., Inel, O., Madan, C. R., … Cheplygina, V. (2020). A survey of crowdsourcing in medical image analysis. *Human Computation*, *7*(1), 1–26.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

Trueblood, J. S., Holmes, W. R., Seegmiller, A. C., Douds, J., Compton, M., Szentirmai, E., … Eichbaum, Q. (2018). The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making. *Cognitive Research: Principles and Implications*, *3*(1), 1–14.

Waite, S., Scott, J., Gale, B., Fuchs, T., Kolla, S., & Reede, D. (2017). Interpretive error in radiology. *American Journal of Roentgenology*, *208*(4), 739–749.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321.