IMPROVING MEDICAL IMAGE DECISION MAKING BY LEVERAGING METACOGNITIVE PROCESSES AND REPRESENTATIONAL SIMILARITY

By

Eeshan Hasan

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

Master of Science

in

Psychology

May 13, 2022

Nashville, Tennessee

Approved:

Jennifer S. Trueblood, Ph.D.

Geoffrey F. Woodman, Ph.D.

Copyright © 2022 Eeshan Hasan
All Rights Reserved

To all of those who suffered through the pandemic.

ACKNOWLEDGMENTS

I thank my advisor Dr. Jennifer Trueblood for all her guidance and encouragement. This project would have been possible without her support. I would also like to thank my co-authors Quentin Eichbaum, Adam Seegmiller and Charles Stratton for their contribution to the project and their medical expertise. I would also like to thank Payton O'Daniels for his excellent research assistance.
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.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Accurate interpretation and classification of medical images is an important step in the diagnosis and treatment of numerous diseases. Unfortunately, diagnostic errors occur despite advanced training and improvements in technology. One successful approach to reducing errors is through second opinions or multiple readings (Elmore et al., 2016; Kurvers et al., 2016; Wolf et al., 2015). For example, Elmore et al. (2016) showed that this approach reduced the misclassification rate from 24.7% to 18.1% in breast histopathology. 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 (Stroop, 1932; Vul and Pashler, 2008; Herzog and Hertwig, 2009; Koriat, 2012; Herzog and Hertwig, 2014; Litvinova et al., 2020; Hasan et al., tted; Litvinova et al., 2019). According to the wisdom of crowds principle, improvements in accuracy are obtained by combining the judgments of different individuals (Surowiecki, 2005). Previous studies have shown that the "wisdom of crowds" can be successfully applied to the domain of medical image decision-making (Kurvers et al., 2016; Wolf et al., 2015). Research on the "inner crowd" applies this same idea, but to a single individual who performs repeated judgments (Litvinova et al., 2019, 2020).

To test our approach, we used data from Hasan et al. (tted). In the task, participants categorized images of white blood cells as cancerous (i.e., "blast" cells) or non-cancerous (i.e., "non-blast" cells). Participants made two separate decisions for each image and reported their corresponding confidence. We examined both experts (i.e., medical professionals) as well as novices (i.e., undergraduate students). Using novices in addition to experts is important for two reasons. First, novice participants provide a point of comparison for evaluating expert performance. Second, novices or semi-professional humans can be used to make simple diagnostic decisions. These decisions can subsequently be used to generate large annotated datasets that are required to train data-hungry algorithms (Ørting et al., 2020).

We explored two algorithms for aggregating decisions with the aim of improving individual accuracy. One successful "wisdom of the crowd" algorithm for binary decision-making is the maximum confidence slating (MCS) algorithm (Koriat, 2012; Litvinova et al., 2019). 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 and Summerfield, 2012; Fleming et al., 2012; Griffin and Tversky, 1992). MCS has previously been shown to work for both novices and experts in medical image decision-making (Hasan et al., tted; Litvinova et al., 2019).

In addition to the maximum confidence slating algorithm, we also explored the similarity based aggregation algorithm (SBA) that leverages tools from machine learning and artificial intelligence to determine similarity. The algorithm determines the "final decision" by aggregating an individual"s decisions on similar images. We used the euclidean distance between latent representations obtained by convolutional neural networks to calculate the similarity between images. In this paper, we examined two representations, one with general visual features (He et al., 2015) and another one with features that are informative for white blood cell classification (Holmes et al., 2020).

Besides using these algorithms to improve accuracy, we compared the effectiveness of these algorithms with each other. Since the effectiveness of these algorithm hinges on different decision-making processes, comparing the algorithms gives valuable clues about these processes. For example, the metacognitive abilities of experts might be better than novices, possibly leading to larger improvements with the MCS algorithm for experts but not for 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.

CHAPTER 2

Behavioral Analysis

2.1 Methods

The data reported here were also reported in Hasan et al. (tted). In the present paper, we compared different "wisdom of the inner crowd" algorithms with a focus on algorithms that use representations from machine learning models. Hasan et al. (tted) examined confidence based aggregation algorithms at both the individual level and across participants (i.e., standard "wisdom of the crowd"). Below, we briefly review the experimental methods and results and refer the reader to Hasan et al. (tted) for additional details.

The experiments were designed to collect two classification decisions on a series of white blood cell images across two blocks. Experiment 1a used novice participants and used two manipulations to elicit slightly different responses in the two blocks. In the first manipulation, the prompt was changed from 'Is this a blast?' to 'Is this a non-blast?'. In the second manipulation, the images were rotated by 180 degrees. The orientation of an image is irrelevant to classification and hence the true classification of an image would remain unchanged with rotation. Experiment 1b was the same as Experiment 1a but did not use the two manipulations mentioned above. Instead, the two blocks used identical prompts and images. Experiment 2 was a shorter version of Experiment 1a but used experts instead of novices. The data for all of the experiments is available on the Open Science Framework at https://bit.ly/3p8LnM7.

2.1.1 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 Institutional Review Board at Vanderbilt University.

The sample size was based on similar studies examining image-based medical decision-making (Trueblood et al., 2018, 2021). 87 undergraduate students (novices) participated in our experiments for course credit. 45 students participated in Experiment 1a and 42 in Experiment 1b. 23 pathologists and laboratory professionals (experts) participated in Experiment 2. These participants were given a $10 Starbucks gift card for participating. The sample size for Experiment 2 was due to convenience.

The participants primarily identified as female in both experiments (Exp. 1a: 76%; Exp. 1b: 70%; Exp 2: 73%). The mean age was 18.9 years (SD=1.2) for Experiment 1a, 19.5 years (SD=2.5) for Experiment 1b, and 42.4 years (SD=13.5) for Experiment 2.

2.1.2 Materials

The stimuli were the same as in Trueblood et al. (2018); Hasan et al. (tted) and consisted of 300 digital images of Wright-stained white bloods cells. These images were taken from anonymized patient peripheral blood smears at Vanderbilt University Medical Center (VUMC) using the CellaVision DM96. See Figure 3.1 for examples of these images. 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. More details on the rating procedure and image curation can be found in Trueblood et al. (2018).

2.1.3 Procedure

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

2.1.3.1 Novices - Experiments 1a & 1b

Novice participants completed three blocks - familiarization, training, and practice blocks - before the main trials. In the familiarization block, participants viewed individual images with their corresponding category for 36 trials. In the training block, participants were asked which of two cell images matched the category label that was presented at the top of the screen for 60 trials. They were given feedback at the end of each trial. In the practice block, they were instructed to classify the image of a cell based on the prompt - "Is this a blast?". Once again, they received feedback at the end of each trial.

The main task was designed so that participants made two decisions on each of the 300 images. In the first block of the main trials, each of the 300 images was presented once. Participants had to decide the category of the image by responding to the prompt 'Is this a blast?'. After they indicated their decision, they were asked to rate their confidence on a scale of 50% (guess) to 100% (certain). In Experiment 1a, in the second block of the main trials, the prompt was changed to 'Is this a non-blast?' and each image was rotated by 180 degrees. Participants in Experiment 1a completed 20 practice trials with the new instructions before starting the second block of the main trials. In Experiment 1b, the second block of the main trials was identical to the first block. That is, there was no change in the prompt and the images were not rotated. In both Experiment 1a and Experiment 1b, the 4 different cell types (blast / non-blast x easy / hard) were counterbalanced in each block.

2.1.3.2 Expert - Experiment 2

Experiment 2 was a shorter version of Experiment 1a that used experts instead of novices. The shorter length was due to the time constraints at the ASCP conference. For this experiment, we only used the hard images to make the experiment challenging for the experts. All of the blocks in the experiment were counterbalanced across the two categories - blast and non-blast. Since experts were already familiar with white blood cells, we shortened their training phase so that it consisted of 20 trials with feedback.

Like the novice experiments, the main task consisted of two parts. However, we only used 60 hard images instead of 300 images that were used in the novice experiments. Both parts of the main task used 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. In the first block of the main trials, the participants were presented with the prompt "Is this a blast?". In the second block of the main trials, this was changed to "Is this a non-blast?". Additionally, the images were rotated by 180 degrees. These were the same manipulations that were used in Experiment 1a.

2.2 Behavioral Results

We followed the exclusion criteria in Hasan et al. (tted) and report relevant results from the paper. We had three exclusion criteria. First, we excluded participants if their accuracy was worse than chance (50%) on the practice trials. Second, participants were excluded if their confidence ratings were outside the valid range of 50 - 100 for 50 or more trials. Third, we excluded participants for giving the same response for a large majority of the trials (¿95%) in a given block. This left us with 34 out of 45 participants in Experiment 1a and 31 out of 42 participants in Experiment 1b. Only 1 out of the 23 experts was excluded. This expert did not give any confidence ratings in the main blocks.

The mean accuracy was 66.1% (SD=8.8; IQR $60.1\% - 71.5\%$) and 66.5% (SD=10.7; IQR $59.5\% - 74.8\%$), for Experiments 1a and 1b respectively. For Experiment 2, the mean accuracy was 71.6% (SD=14.3; IQR $60.1\% - 83.9\%$). We also report the accuracy from Experiment 1a and 1b on the subset of stimuli seen by experts in Experiment 2 so that we could directly compare the performance of novices and experts. On this common 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 performed better than novices.

Since self-reported confidence judgments can have large individual differences, it is important to normalize the confidence ratings before applying MCS (Griffin and Tversky, 1992; Griffin and Brenner, 2004; Koriat, 2012). Additionally, participants could have changed the way they reported their confidence over the blocks. In Experiment 1a and Experiment 2, participants were also responding to different prompts in the two

blocks, which could have also affected the way the confidence scales were used. To determine if participants changed the way they used the confidence scales in the two parts of the experiment, we conducted a Kolmogorov Smirnoff test. We found that 18 out of 34 participants in Experiment 1a, 18 out of 31 participants in Experiment 1b and 6 out of 22 participants in Experiment 2 had significantly (p¡0.05) different distributions for confidence ratings in the two parts of the main task. Hence we normalized the confidence ratings for the two blocks of the main task separately. That is, for each person, we calculated the z-score of the confidence ratings separately for both blocks in the main task.

We were interested in understanding how accuracy, cell type and difficulty were related to the confidence ratings. We conducted a 2x2x2 (accuracy: correct vs incorrect) x (classification: blast vs non-blast) x (difficulty: easy vs hard) repeated measures ANOVA for novices. Since experts only categorized hard images, we conducted a 2x2 (accuracy: correct vs incorrect) x (classification: blast vs non-blast) repeated measures ANOVA for experts. We conducted our analysis only on the main trials since the practice trials were used in the exclusion criteria. Across the three experiments, we observed a significant main effect of accuracy (Exp. 1a: $F(1,33) = 97.9$ $p < 0.0001$; Exp. 1b: $F(1,30) = 71.8$, $p < 0.0001$; Exp. 2: $F(1,21) = 36.6$, $p < 0.0001$). We also found 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 Exp. 1b ($F(1,30) = 2.3$, $p = 0.1407$). We also found significant interactions between confidence and classification and difficulty in both the novice experiments. In sum, when participants were accurate, they were also more confident. This shows that confidence reflects accuracy, which is critical for the application of the maximum confidence slating algorithm.

CHAPTER 3

Modeling Analysis

3.1 Modeling Methods

As mentioned above, we 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.

3.1.1 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 (Koriat, 2012). First, we normalize 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. We note that the MCS algorithm was also examined for this data in Hasan et al. (tted). Here we include this algorithm as a point of comparison for the similarity based aggregation algorithms, described below.

3.1.2 Similarity based aggregation (SBA)

Similarity Based Aggregation (SBA) attempts to improve individual performance by aggregating the decisions made on similar images. In these algorithms, we first calculate the similarity between two images and then use a k Nearest Neighbor (kNN) imputation. Figure 3.2 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.

3.1.2.1 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 (Weiss et al., 2016). To this end, we use a GoogLeNet that was trained on ImageNet, 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 3.1 top row). The model was not trained on the blast task. As shown in Figure 3.2, the classes are slightly separated but also overlap in this representation.

3.1.2.2 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 3.1 bottom row). A

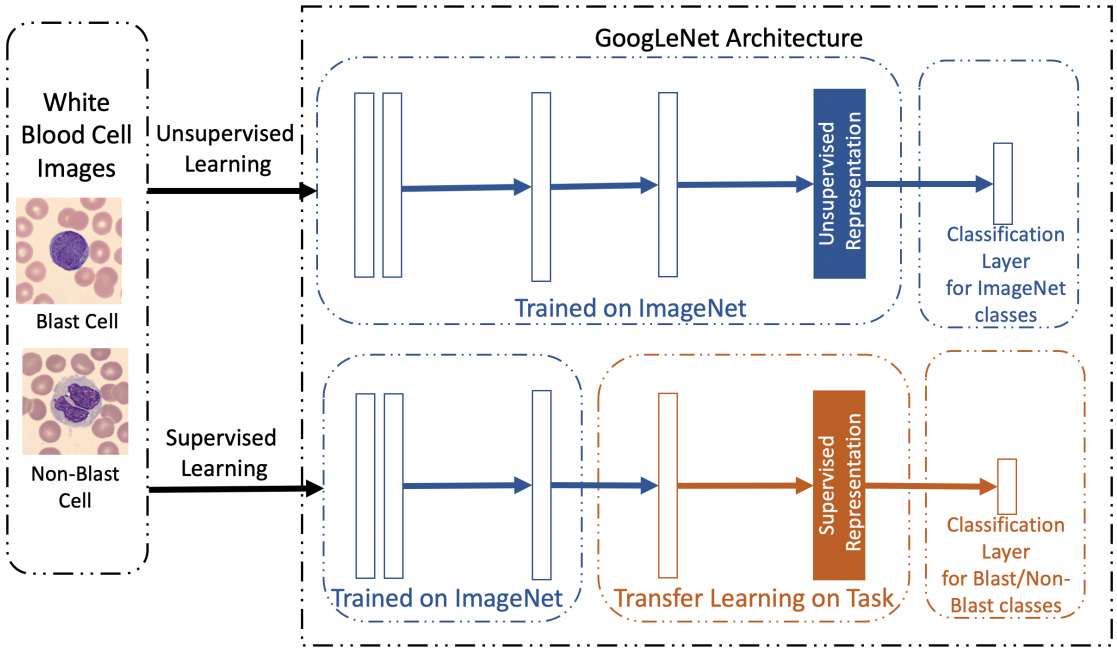

Figure 3.1: Schematic of the supervised and unsupervised representations. The unsupervised representation was obtained by using GoogLeNet trained on ImageNet. The supervised representation was obtained by using transfer learning on a GoogLeNet trained on ImageNet.

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 3.2, the classes are distinctly separated in the representation.

3.1.2.3 k Nearest Neighbor imputation

For every image, we use the k nearest neighbors to calculate the final response. That is, we examine 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 includes 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 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.

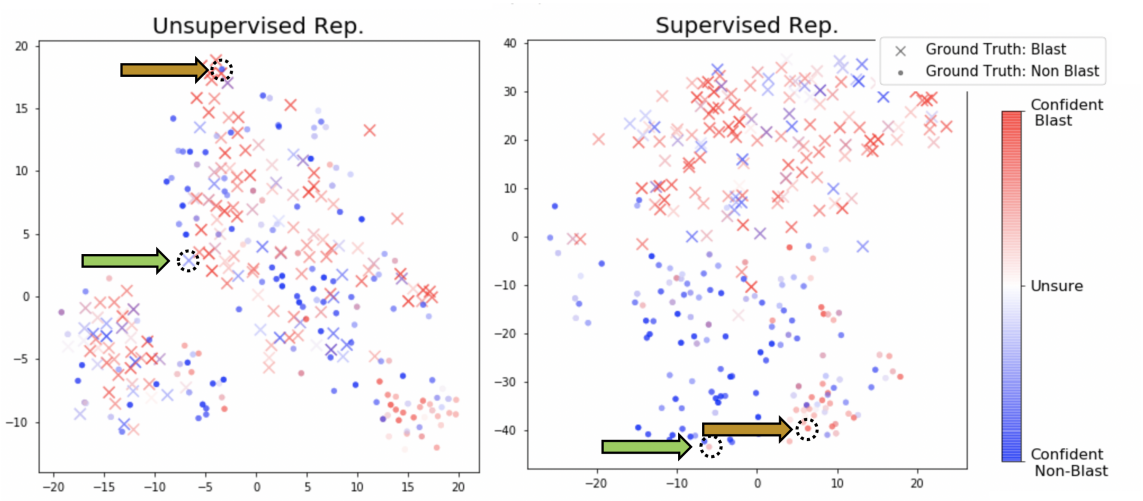

Figure 3.2: Example of a representative participant's judgments in the unsupervised (left) and supervised (right) representational spaces. The green (gold) arrows illustrate situations where the neighbors point to the correct (incorrect) answer. For example, in the right panel, both the arrows point 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).

3.2 Modeling Results

We applied the 5 algorithms to the data. The average performance of each algorithm is in Table 3.1. 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 3.2.

Table 3.1: 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%
MCS	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%

Table 3.2: 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	MCS	**-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
MCS	Unsupervised k=3	0.8544	0.3990	0.8943	0.3783	0.7644	0.4531
MCS	Unsupervised k=7	**4.2489**	**0.0002**	**5.0598**	**0.0000**	**6.8010**	**0.0000**
MCS	Supervised k=3	**-4.0949**	**0.0003**	-2.5250	0.0171	0.7812	0.4434
MCS	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

3.2.1 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 3.1, the mean MCS is 67.4% both for Exp. 1a and 1b, and 73.8% for Exp. 2, which is higher than the average response in these experiments. As shown in Table 3.2, 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 ($t(21) = -3.5, p = 0.0020$). These results are also reported in Hasan et al. (tted).

3.2.2 SBA versus Average Performance

Next, we compared the performance of the SBA algorithms to average performance using the post-hoc tests mentioned above. As seen in Table 3.1, 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 3.2, the post-hoc tests show that this improvement in performance is marginally significant for the first experiment with novices but not the second experiment (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 (Exp. 2: M=73.0%), which is 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 on features specific to the task.

As seen in Tables 3.1 and 3.2, 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 SBA algorithms are effective for the novices but not for the 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 3.2, we observe that the performance is significantly better for SBA with the supervised than the unsupervised representation (Exp. 1a: $t(33) = -5.8$, $p < 0.0001$; Exp. 1b: $t(30) = -4.5$, $p < 0.0001$).

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 3.2, 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 SBA algorithms 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.

3.2.3 Comparing MCS to SBA

We now compare the SBA and MCS algorithms. It is especially of interest to compare the SBA algorithm with $k = 3$ to 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 SBA algorithms use the similarity structure of the underlying problem. 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: $t(21) = 0.8 p = 0.4531$). The supervised representation at $k = 3$ outperforms MCS for the novices, but not for the experts, where the difference is not significant (Exp. 1a: $t(33) = -4.1 p = 0.0003$; Exp. 1b: $t(30) = -2.5$, $p = 0.0171$; Exp. 2: $t(21) = 0.8$, $p = 0.4438$). The pattern is the same for $k = 7$. Our results suggest that for experts, it might be better to rely on their metacognitive judgments, but for novices to use their decisions on similar images, especially with a well tuned representation.

3.2.4 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 SBA algorithms. With 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 a 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 SBA seems to be effective for novices, but not for experts even when restricted to the exact same image set.

CHAPTER 4

General 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 and Summerfield, 2012; Koriat, 2012; Fleming et al., 2012; Griffin and Tversky, 1992). 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 MCS algorithm improved performance in all of the experiments, suggesting that confidence judgments can meaningfully solve the conflict of inconsistent decisions (Hasan et al., tted). 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; Herzog and Hertwig, 2009, 2014). 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 et al., 2019).

Regarding the SBA algorithms, 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. We note that the improvements made to the accuracy using the unsupervised representation at k=3 are similar in magnitude to the improvements made by MCS for Experiment 1a. This is interesting because it had no information about the participants' confidence and also had no information about the task at hand. Instead it solved inconsistencies by leveraging information about general visual features relevant to the classification of unrelated images in ImageNet. Aggregating responses using the supervised neural network allows one to make larger improvements in accuracy in two different ways. At k=3, it allows one to break inconsistent responses at a better rate than the unsupervised representation. The efficacy of SBA with the supervised representation is further boosted by pooling responses from even more images. However, using a representation that is not as informative can hurt performance when pooling over a large number of responses, as was the case with the unsupervised representation.

In contrast, there was no improvement in the performance of experts with SBA even when the represen-

tation was informative and well tuned. This means that aggregating responses on similar images might not be useful for experts. This might be because experts are more likely to make the same decision on similar images. That is, their decision might be biased towards the wrong answer in certain parts of the representational space. On the other hand, for novices, we see substantial improvement with SBA 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. Our results where algorithm efficacy depended on the population, re-affirm the cautionary tale that it is important to test algorithms on the specific population of interest.

References

Boldt, A., Schiffer, A.-M., Waszak, F., and 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., Onega, T., Carney, P. A., Jackson, S. L., Allison, K. H., and Weaver, D. L. (2016). Variability in pathologists' interpretations of individual breast biopsy slides: A population perspective. *Annals of internal medicine*, 164(10):649–655.

Fleming, S. M., Dolan, R. J., and Frith, C. D. (2012). Metacognition: computation, biology and function.

Griffin, D. and Brenner, L. (2004). Perspectives on probability judgment calibration. *Blackwell handbook of judgment and decision making*, pages 177–199.

Griffin, D. and Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive psychology*, 24(3):411–435.

Hasan, E., Eichbaum, Q., Seegmiller, A. C., Stratton, C., and Trueblood, J. S. (submitted). Harnessing the wisdom of the confident crowd in medical image decision-making. *Decision*.

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

Herzog, S. M. and Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20(2):231–237.

Herzog, S. M. and Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in cognitive sciences*, 18(10):504–506.

Holmes, W. R., O'Daniels, P., and 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.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Argenziano, G., Zalaudek, I., and Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, 113(31):8777–8782.

Litvinova, A., Herzog, S. M., Kall, A. A., Pleskac, T. J., and Hertwig, R. (2020). How the "wisdom of the inner crowd" can boost accuracy of confidence judgments. *Decision*, 7(3):183.

Litvinova, A., Kurvers, R. H., Hertwig, R., and Herzog, S. M. (2019). When experts make inconsistent decisions.

Nelson, A. M., Hale, M., Diomande, M. I. J.-M., Eichbaum, Q., Iliyasu, Y., Kalengayi, R. M., Rugwizangoga, B., and 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., Mavridis, P., Spiers, H., and Cheplygina, V. (2020). A survey of crowdsourcing in medical image analysis. *Human Computation*, 7(1):1–26.

Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of experimental Psychology*, 15(5):550.

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

Trueblood, J. S., Eichbaum, Q., Seegmiller, A. C., Stratton, C., O'Daniels, P., and Holmes, W. R. (2021). Disentangling prevalence induced biases in medical image decision-making. *Cognition*, 212:104713.

Trueblood, J. S., Holmes, W. R., Seegmiller, A. C., Douds, J., Compton, M., Szentirmai, E., Woodruff, M., Huang, W., Stratton, C., and 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.

Vul, E. and Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7):645–647.

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

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

Wolf, M., Krause, J., Carney, P. A., Bogart, A., and Kurvers, R. H. (2015). Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. *PloS one*, 10(8):e0134269.

Yeung, N. and 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.