Title
Representational Smoothing to Improve Medical Image Decision Making

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors
Hasan, Eeshan
Trueblood, Jennifer

Publication Date
2022

Peer reviewed

Representational Smoothing to Improve Medical Image Decision Making

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

Abstract

We demonstrate how medical-image classification decisions can be denoised by aggregating decisions on similar images. In our algorithm, the final decision on a target image is cancerous if a percentage t of the k most similar images are cancerous, else it is not cancerous. Similarity between images is calculated as the distance between representations from an artificial neural network. We vary k and t for novice and expert participants using data from Trueblood et al. (2018) and Trueblood et al. (2021). We show that increasing k improves performance for novices, with their performance approaching that of experts. We also show that the algorithm is biased towards identifying cancerous cells, which is reflected in the representational space. The percentage t allows greater control over sensitivity and specificity and can be used to debias decisions. This algorithm is less effective for experts, partially explained by them giving similar responses on similar images.

Keywords: Medical Image Decision Making; Computational Modeling; Neural Networks; Representation; Concepts and Categories

Introduction

The identification and treatment of several diseases is contingent on the interpretations of medical images (e.g., images of blood cells in the diagnosis of leukemia) by doctors and other medical professionals. Despite advanced training, diagnostic mistakes occur. Some of these errors occur at random. In such cases, one might be able to use correct decisions made on similar images (e.g., blood cells with similar morphological characteristics) to overturn the original decision and fix it. Such a process would effectively "de-noise" decisions, leading to improvements in accuracy.

Artificial neural network representations trained on tasks such as categorization are similar to the ones measured in the visual cortex of the primate brain (Yamins & DiCarlo, 2016). They can also be used to determine the similarity between two images and as inputs to cognitive models (Sanders & Nosofsky, 2020; Peterson, Abbott, & Griffiths, 2018; Holmes, O'Daniels, & Trueblood, 2020). In exemplar cognitive models of categorization, one determines the label of a target based on the similarity between a target and other similar objects. Models based on such representations have been developed to model categorization beyond hand crafted lab stimuli to more naturalistic stimuli (Sanders & Nosofsky, 2020; Singh, Peterson, Battleday, & Griffiths, 2020). We consider the possibility of creating a 'hybrid' approach to categorization, where we boost the accuracy of an agent by instantiating such a process computationally after an agent has made their decisions.

In this paper, we build on our Similarity Based Aggregation (SBA) algorithm (Hasan, Eichbaum, Seegmiller, Stratton, & Trueblood, 2021b) based on the idea of aggregating decisions over similar images. In (Hasan et al., 2021b), for a given target image, we consider 'k' decisions made on the most similar images (including the decision made on the target image). We then consider the 'final aggregated response' on the target image to be the modal response in that set of images. This process is conducted separately for every individual. The algorithm is a de-noising procedure that smooths decisions in the representational space for that individual. Previously, we used data from (Hasan, Eichbaum, Seegmiller, Stratton, & Trueblood, 2021a) to show that aggregating over a small number of similar decisions can be used to improve performance in novices but not experts. We also showed that while small improvements were possible using general representations, using a representation that was obtained by training on cancer cell classification with task relevant information was especially effective in improving accuracy.

In this paper, we examine the consequences of varying the number of neighbors (k) used to generate the aggregated response. On the one hand, using a large k allows us to pool responses from more neighbors making the smoothing less noisy and possibly more accurate. On the other hand, using a large k amounts to using less similar neighbors and thereby potentially including neighbors that belong to another class.

In medical image classification, one might treat false alarms and misses differently. For example, while screening for cancer, a false alarm can be dealt with by conducting more tests (although this comes at additional cost). However, a missed diagnosis will stop further tests and might allow the cancer to metastasize, making future treatment tougher. In previous work, it was not clear whether the improvement in performance due to SBA was due to an improvement in specificity or sensitivity or both. This is pertinent to SBA since medical images may not be evenly distributed in the representational space. For example, cancerous white blood cells (called blast cells) might be closer to each other in the representational space while non-cancerous white blood cells (called non-blast cells) are composed of different kinds of cells and could be further apart in the representational space. Further, in humans, the trade-off between sensitiv-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

ity and specificity is affected by various factors such as the baseline prevalence of cancer in the dataset (Trueblood et al., 2021).

In this paper, we examine a modified version of SBA to include a threshold parameter to manage the tradeoff between sensitivity and specificity. According to this parameter, instead of using the modal response, we decide that a cell is cancerous if a certain fraction (t) of its nearest neighbors are also cancerous. Hence, by setting a low threshold, one might make the algorithm more sensitive while reducing the specificity.

In this paper, we use data from Trueblood et al. (2018) and Trueblood et al. (2021) which contained classification decisions (cancer or not) made by novices (i.e., undergraduate students) and experts (i.e., medical professionals) on sets of white blood cell images (examples in Figure 1). We use novice participants in addition to medical experts for two important reasons. First, novices provide a baseline for comparing experts. Second, there is recent interest in using novices to assist with medical image diagnosis. Particularly relevant for this paper is the possibility of crowd-sourcing large numbers of untrained individuals to perform simple diagnostic tasks (Ørting et al., 2020; Press, 2021), which can later be used to train data hungry artificial intelligence algorithms. Third, we previously observed that our algorithm improved performance for novices but not experts, suggesting different decision making mechanisms for the two populations (Hasan et al., 2021b).

Methods

Datasets

All the experiments involved making binary decisions about Wright-Stained White Blood Cells. These cells were classified into 'blast' and 'non-blast' categories based on the mutual independent agreement of three hematopathologists at Vanderbilt University Medical Center. Example cell images can be seen in Panel (a) of Figure 1. More details of the image curation and experimental procedure can be found in Trueblood et al. (2018) and Trueblood et al. (2021).

Exp. 1 and Exp. 2 were from (Trueblood et al., 2018). Participants were trained to classify white blood cells using two tasks before the main trials. In the first task, they were exposed to blast and non-blast images along with their labels. In the second task, they had to pick the blast cell among three images. In the main task, participants made decisions under three conditions - speed, accuracy and bias. In this paper, we only analyze results from the speed and accuracy conditions. In the speed condition, they were asked to make decisions 'as fast as they can' and 'as accurately as they can' in the accuracy condition. The participants completed practice trials before the main task to familiarize themselves with the conditions and interface. Undergraduate students from Vanderbilt University participated as novice participants in Exp. 1. Pathologists with a range of experience from first year pathology residents to senior faculty pathologists from Vanderbilt University Medical Center participated as experts in Exp. 2. The procedure was identical for both the novice (Exp 1) and expert (Exp 2) participants.

Exp 3 and Exp 4 were from (Trueblood et al., 2021). The training phases for these experiments were similar to the ones described above with minor differences. In the main blocks of these experiments the prevalence rate of blast (i.e., cancer) cells was varied in different conditions. Exp. 3a and Exp. 3b used undergraduate students at Vanderbilt University as novice participants. Exp. 3a had three conditions with *50%*, *25%*, and *75%* blast prevalence. These conditions were varied within subject for Exp. 3a. Exp 3b had three conditions with blast prevalence *50%*, *10%*, and *90%*. All participants did the condition with *50%* but the *10%* blast prevalence and *90%* blast prevalence was varied between subjects. This was done to gain enough responses on blast cells in the *10%* condition and non blast responses in the *90%* blast prevalence condition. Exp 4 used expert participants with *50%* and *90%* blast prevalence. More details of the datasets can be found in (Trueblood et al., 2018) and (Trueblood et al., 2021).

In both (Trueblood et al., 2018) and (Trueblood et al., 2021), novice participants were trained on classifying white blood cells prior to starting the main task. On average, novices were above chance performance in categorizing the cell images. Full details on the training procedure can be found in the original papers.

Representation

In this paper, following Hasan et al. (2021b), we use a representation from (Holmes et al., 2020). This representation was obtained by using a pre-trained GoogLeNet (Szegedy et al., 2015) on ImageNet to classify cancer cells using transfer learning. This representation had 1024 abstract dimensions that contained information relevant to classifying cancer cells. We visualised this representation in Figure 1 by using t-SNE, a dimensionality reduction technique. As is clear from the figure, this representation neatly separates blast and non-blast cells. Most of the neighbors for all of the cells belong to the same class. The accuracy of the network was 98% on the training dataset and 94% on the validation dataset. This shows that the network generalized to out of training sample images without overfitting the data too much. (Holmes et al., 2020).

Similarity Based Aggregation Algorithm (SBA)

As mentioned above, in our algorithm, for a given participant, for a given target image, we consider the responses made on the k most similar images (including the response made on the target image). Similarity between two images is determined as the inverse of the distance between their representations. If the percentage of cancer decisions on this set is greater than a threshold t, the algorithm selects 'cancer' as the 'final' response on that image for that participant.

In our analyses, we vary the number of neighbors (k) that are used in SBA. For $k = 3$, one might be able to overturn the original decision on the image if both the decisions made on

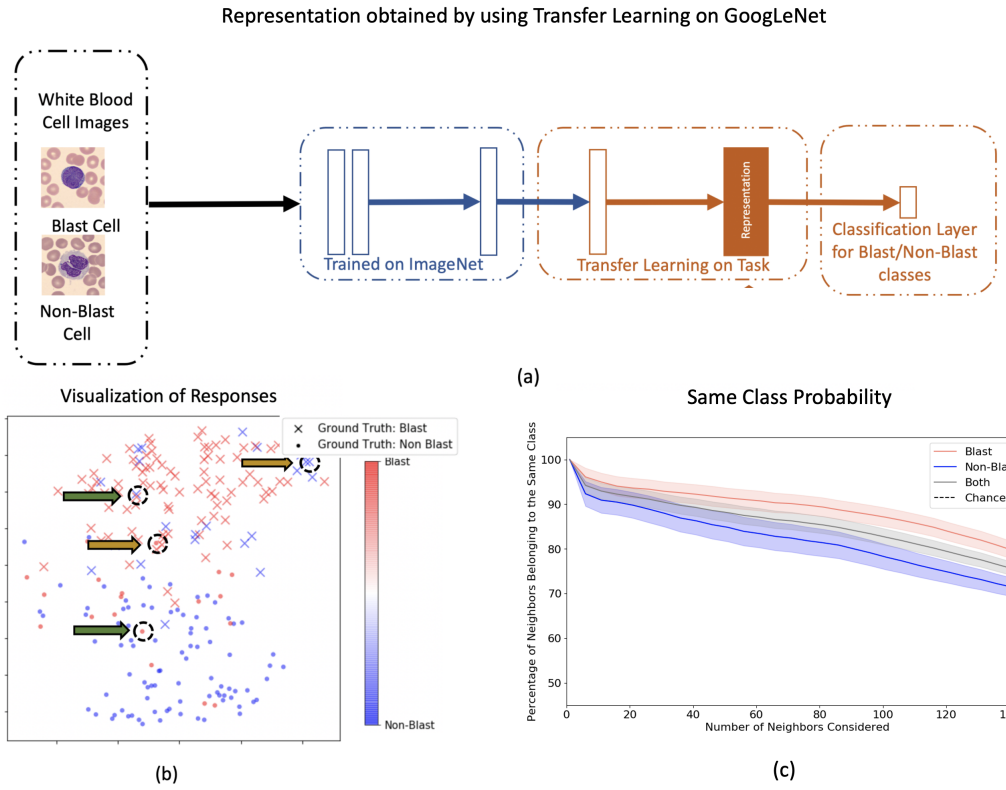

Figure 1: [(a) Top] Schematic of the Holmes et. al. (2020) representation. The representation was obtained by using transfer learning on a GoogLeNet trained on ImageNet to classify blast cells. The activations in the penultimate layer were used as our representation. [(b) Bottom Left] The decisions made by a typical individual in the experiment. The crosses (circles) are cells where the true class was blast (non-blast). Red (Blue) markers indicate that the decision made by the individual was cancerous (non-cancerous). The green arrows show where the algorithm is expected to work since the decisions on the neighbors were correct. The top yellow arrow is where the decisions made on the neighbors are also incorrect. The bottom yellow arrow shows an example of where the decisions made on the neighbors are correct but the images belong to another class. [(c) Bottom Right] This Figure shows the relationship between the number of neighbors k and the average percentage of the k closest neighbors belonging to the same class as a target image. We observe that as more images are considered, the probability that they belong to the same class as the target decreases. We also observe that the decrease is more stark for non-blast cells than for blast cells.

the two most similar images were different from the one made on that image. However, with $k = 15$, it is easier to overturn the original decision as it only requires that 8 out of the 14 decisions made on the most similar images to be different from the original one. Examples of when the algorithm might be successful or not can be found in Figure 1.

Signal Detection Theory

Signal Detection Theory (SDT) is used to study the categorization ability of individuals along with their bias towards making false alarms and misses (Stanislaw & Todorov, 1999). It calculates two parameters - discriminability and criterion. Discriminability is a measure of performance or the ability to distinguish between blast and non-blast cells. Criterion measures how a participant manages the trade-off between false alarms and misses. If a participant is biased towards false alarms (saying 'blast'), their criterion is negative. A bias towards misses (saying 'non-blast') is indicated by a positive criterion. In our experiment, to avoid a perfect hit rate of 1 or perfect false alarm of 0, we used Laplace smoothing of 1, where we added 4 responses to every participant (one blast and one non-blast response to a blast cell and one blast and non-blast response to a non-blast).

Results

Representation

We report the results of the basic analysis of the representation in Panel (c) of Figure 1. The rate at which an image and its closest neighbor belonged to the same class was 94.3%. Hence, in some cases, the closest neighbor was not of the same class. This number falls to 92.3% when considering the 7 closest neighbors. This further drops to 91.7% and 87.7% when considering the 15 and 51 closest neighbors respec-

tively. Therefore, as expected, as we look at more neighbors, the rate at which they belong to the same class as our target image, declines.

Interestingly, as shown in panel (c) of Figure 1, this decline is different for blast and non-blast cells, where the probability of the neighbors belonging to the same class is greater for blast cells. Initially, at $k = 3$, this gap is significant (blast:98.2% non-blast:94.9%, $t(298) = 2.27$, $p = 0.023$). For larger k, such as 51, this gap widens and becomes highly significant (blast:91.4% non-blast:84.8%, $t(298) = 3.32$, $p = 0.0001$). Hence, the algorithm might not work as effectively for non-blast cells as blast cells.

Varying the Number of Neighbors k

As mentioned above, using a large number of images for the smoothing process results in pooling responses from more images, making the smoothing less noisy. However, using a large number of responses also decreases the probability that the cell belongs to the same class and increases the chance of the original decision being overturned.

We present the results of SBA for different k values in Table 1 and Panel (a) of Figure 2. We observe that initially increasing the number of responses improves performance for all of the novice experiments for all conditions. However, this improvement flattens out and then declines. For expert participants, on the other hand, the improvement is slight and non-significant with Bonferroni corrected p-values. The results are similar for Exp. 3 and 4, as shown in Table 2.

Since the procedures for novices and experts were identical in Exp. 1 and Exp. 2, we used this dataset to compare the performance of the two sets of participants. As shown in the Table 1, initially, experts perform better than novices and have a significantly higher accuracy in both speed and accuracy conditions (Speed: $t(52) = 4.54$; $p < 0.0001$; Accuracy: $t(52) = 5.12$; $p < 0.0001$). We then used the best performing k-value ($k = 51$) to compare the performance of the novices after we apply the algorithm with the performance based on the responses made by the experts. We observe that there is no significant difference between the performance of the novices after the model and the experts (Speed: $t(52) = -0.30$; $p = 0.765$; Accuracy: $t(52) = -1.34$; $p = 0.187$). This difference is non-significant even after we apply the algorithm to the experts (at $k = 51$) (Speed: $t(52) = -0.61$; $p = 0.545$ Accuracy: $t(52) = -1.53$; $p = 0.133$). Hence, it seems that after the application of the algorithm, the performance of the novices is similar to that of the experts.

Since we were interested in understanding if the algorithm improved performance for blast and non-blast cells differently, we conducted an SDT analysis of our data. For these analyses, we calculated the discriminability and criterion of individual participants before and after the application of algorithm. Figure 2 plots the mean difference in discriminability and criterion before and after application of the algorithm. As shown in Panel (a) of Figure 2, we observe that the algorithm significantly lowers the criterion. This indicates that our algorithm is biased towards identifying blast cells. As for the discriminability, we also observe a significant increase in the discriminability followed by a decline, which mirrors the accuracy results (except for at k=101, where the discriminability is high). In fact, even for experts, the improvement in discriminability is significant for k=3,7,15,31, and 51 (although smaller than novices).

Varying the Threshold t

For the following analyses, we set $k = 15$ and varied the threshold to see whether it could be used to manage the tradeoff between false alarms and misses. Since accuracy does not measure the tradeoff between false alarms and misses, we used SDT to evaluate our results. We were also interested in seeing how the algorithm would respond to biased datasets as in (Trueblood et al., 2021). Hence, we report our results by varying the threshold and applying it to Exp. 3 and 4.

We show our results in Panels (b), (c) and (d) in Figure 2. For Exp. 3a and 3b, we observe that across all conditions, when the threshold is small (less than 50%), the criterion is lower after SBA is applied (indicated by the negative difference). This means that SBA increases the rate at which images are classified as blast cells. This is because it needs a smaller number of neighbors to have been labelled cancerous before it decides that a given cell is cancerous. Similarly, when the threshold is high, the criterion is higher. It is also important to note that, for novices, except for in cases of extreme values of t, the discriminability is improved by using the algorithm (difference is positive). For experts, the improvement in discriminability is much smaller.

Similarity Consistency Rate (SCR)

We wanted to understand why SBA was more effective for novices than for experts. Since experts were more experienced with these images, it was possible that they made similar decisions on similar cells. In this case, one would not be able to overturn incorrect decisions, since the experts would have also incorrectly judged similar images to be of the same class. We calculate the Similarity Consistency Rate (SCR) as the rate at which the same response was made on a given image and its most similar neighbor.

If an individual has a really high similarity consistency rate, then the algorithm may not be very effective since the responses on an image and its most similar images will often be the same. Hence, SBA will not be able to change many of the responses. We compare this rate for novices and experts using a independent measures t-test. The SCR was lower for novices in Exp.1 than for experts in Exp. 2 in both the conditions (Accuracy: Exp. 1 - $M = 71.1\%$, Exp. 2 - $M = 80.8\%$ $t(52) = 4.2$; $p < 0.0001$) (Speed - Exp. 1 : $M = 64.0\%$, Exp. 2 : $M = 75.8\%$; $t(52) = 4.7$; $p < 0.0001$). Similarly the SCR was significantly lower for novices in Exp 3a. ($M = 65.5\%$) and Exp. 3b (63.3%) than experts in Exp 4 (82.2%) at 50% prevalence ($t(56) = 7.8$; $p < 0.0001$ and $t(74) = 8.3$; $p < 0.0001$) and Exp. 3b (70.1%) and Exp 4. (82.0%) at 90% prevalence ($t(45) = 4.6; p < 0.0001$). These

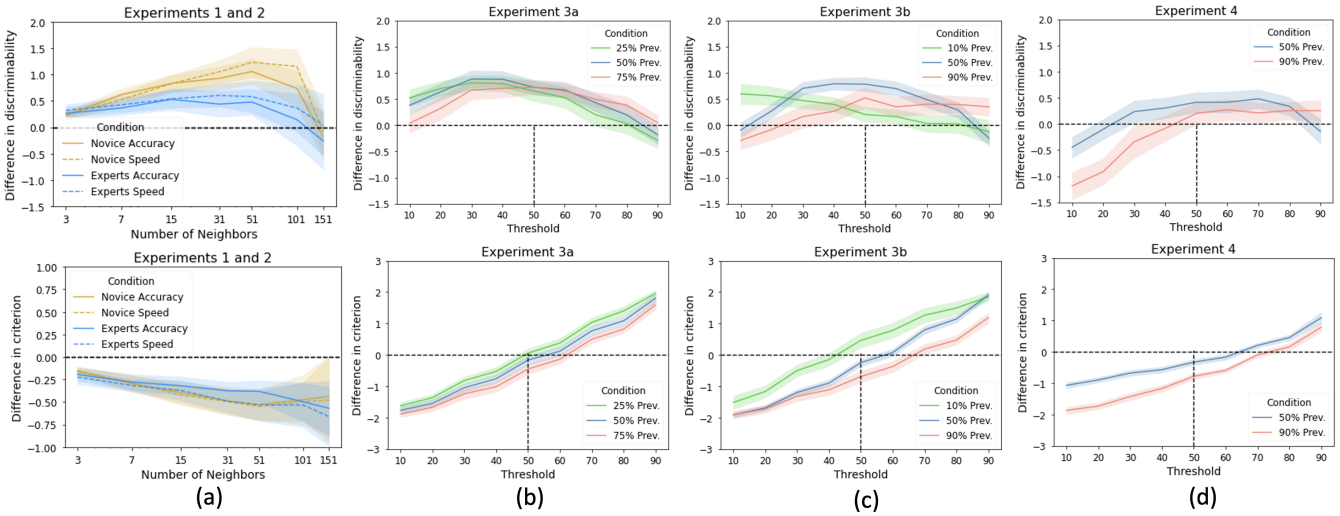

Figure 2: The results of the SDT analysis. The top row shows the difference in discriminability before and after applying the algorithm. For discriminability, a difference greater than 0 suggests an improvement. For criterion, a criterion less than 0 shows an increased tendency to choose blast cells (leading to an increase in hits and false alarms). Column (a) shows the results of varying the number of neighbors (k) in Exp. 1 (novice participants) and 2 (expert participants). The golden (blue) line is from Exp. 1 (2). The dashed line represents the speed condition while the solid line represents the accuracy condition. We observe that for Exp. 1 (novice), there is a much larger improvement in discriminability than for Exp. 2 (experts). We also observe that increasing the number of neighbors slightly shifts the criterion lower. Columns (b),(c) and (d) show the effect of varying the threshold (t) (i.e., the percentage of blast cells for a blast decision) in Exp. 3a, 3b (novice participants) and 4 (expert participants) respectively. Each of the lines represent one condition in each of the experiments.

results show that experts are more likely than novices to give similar responses on similar cell images.

Table 1: This table contains the accuracy results when SBA is applied to Exp 1 and Exp 2 from Trueblood et al. (2018). The data contained responses from novice and expert participants in speed and accuracy conditions. The bold values show a significant improvement compared to average accuracy at the Bonferroni corrected p-value of $p<0.05/7=0.0071$. The algorithm successfully improved the performance for novice participants across the task conditions. It did not improve the accuracy for the expert participants significantly ($p>0.0071$).

	Exp. 1		Exp. 2	
Condition	Accuracy	Speed	Accuracy	Speed
Avg. Acc.	73.8%	71.9%	85.7%	83.6%
3	**75.8%**	**74.0%**	86.6%	84.6%
7	**78.8%**	**76.4%**	86.9%	84.8%
15	**80.1%**	**78.8%**	87.9%	85.4%
31	**80.7%**	**80.8%**	86.7%	85.1%
51	**81.3%**	**82.5%**	86.9%	84.9%
101	76.6%	**80.3%**	84.2%	83.9%
151	**63.8%**	64.4%	78.6%	79.3%

Discussion

In this paper, we leveraged the similarity between images to improve medical image decision making. We considered the similarity based aggregation (SBA) algorithm that pools responses made by an individual on similar stimuli (in our case, cell images) in order to improve performance. We show that the SBA algorithm can be used to boost accuracy across different task conditions for novice participants performing a medical image classification task. For experts, the algorithm works only in limited settings with it failing to improve accuracy or even hurting performance in some conditions.

In our approach, we use representations obtained from neural networks to determine the similarity between two cells. The distance between two stimuli has been shown to correlate with human judgments of similarity in a wide range of tasks, metrics, and representations (Richie & Bhatia, 2021; Peterson et al., 2018). In our previous work, we compared the results obtained from a GoogLeNet trained only on ImageNet to the representation obtained by using the same GoogLeNet trained on cancer cell classification through transfer learning (Holmes et al., 2020). We demonstrated that without transfer learning, the improvement by SBA was limited because all of the neighbors did not belong to the same class. The transfer learning approach learned a representation such that images of the same class were close to each other in the representational space. In other metric learning approaches, representa-

Table 2: This table contains the accuracy results when SBA is applied to Exp 3 and Exp 4 from Trueblood et al. (2021) where the blast prevalence was varied. The data contained responses from novice and expert participants. The bold values show a significant improvement compared to average accuracy at $p<0.05/7=0.007$. The algorithm successfully improved the performance for novice participants across the task conditions. There was no improvement for experts.

	Experiment 3a (Novice Part.)			Experiment 3b (Novice Part.)			Experiment 4 (Expert Part.)	
Condition	50% Prev.	25% Prev.	75% Prev.	50% Prev.	10% Prev.	90% Prev.	50% Prev.	90% Prev.
Avg. Acc.	68.3%	70.0%	71.1%	67.7%	74.1%	79.4%	90.3%	91.8%
3	**70.8%**	**72.1%**	**74.0%**	**70.6%**	**76.6%**	**82.0%**	91.1%	92.6%
7	**73.4%**	**74.9%**	**78.0%**	**73.3%**	**78.6%**	**83.7%**	91.5%	93.3%
15	**75.2%**	**77.1%**	**79.8%**	**75.6%**	**80.4%**	**87.0%**	91.9%	94.2%
31	**76.6%**	**78.2%**	**81.2%**	**76.8%**	**82.3%**	**87.6%**	91.6%	93.2%
51	**77.1%**	**78.7%**	**81.6%**	**77.8%**	**82.9%**	**88.1%**	91.9%	91.8%
101	**78.6%**	**80.2%**	**79.6%**	**76.8%**	**82.9%**	**88.5%**	92.5%	90.4%
151	**78.7%**	**79.6%**	**77.9%**	**74.7%**	**83.2%**	**88.0%**	89.2%	90.0%

tions where neighbors belong the same class can be obtained (Zhuang, Cai, Wang, Zhang, & Zheng, 2020). For SBA to be effective, it is sufficient for the images from the same class to neighbor each other in the representational space. However, in future work, one might try to obtain a representation that corresponds more closely to mental representations (Peterson et al., 2018; Richie & Bhatia, 2021; Nosofsky, Sanders, & McDaniel, 2018; Sanders & Nosofsky, 2020).

We notice that increasing the number of neighbors can dramatically improve performance for novice participants. However, there is a limit to this process, where using a very large number of neighbors might hurt performance. This improvement is significant for novices but not for experts. After the application of the algorithm, the performance of the novices from Exp. 1 is similar to that of the experts in Exp. 2. This shows the power of the SBA in practical applications, where novices can be used to label images to train medical artificial intelligence systems (Ørting et al., 2020; Press, 2021), which in turn can be used to improve artificial neural network representations, which can then be used to boost the accuracy of the novices.

It is interesting to note that blast cells have a higher probability of having neighbors that belong to the same class than non-blast cells. This likely occurs because non-blast cells are composed of several different cell types as compared to blast cells (Al-Dulaimi, Banks, Chandran, Tomeo-Reyes, & Nguyen Thanh, 2018; Nissim, Dudaie, Barnea, & Shaked, 2021). As a result, the algorithm is biased towards responding blast than non-blast. This indicates that the geometrical properties or the way categories are distributed in the representational space might influence its efficacy on a dataset. Future work could investigate this for classification tasks with a greater number of classes, which are distributed in more non-homogenous ways in the representational space. In this case, the structure of the representational space may play a larger role in the kind of errors that SBA can resolve, and the ways in which it might bias the results.

We used SDT to evaluate the effect of changing the threshold parameter (i.e., percentage of blast cells needed to make a blast decision). We observe that it allows us to tradeoff between blast and non-blast cells while maintaining a similar and improved discriminability. Hence, we show that smoothing the responses in an uneven way can control the tradeoff between sensitivity and specificity. In situations with unequal prevalence rates, such as the one in (Trueblood et al., 2021), where changing the prevalence of blast cells causes participants to give biased responses, the algorithm can be used to de-bias responses depending on the requirement. For example, if the algorithm was being used to screen for cancer, one might desire more sensitivity and use a lower threshold. However, if it was used for confirmatory testing, it could be made more specific by using a higher threshold.

We observed the similarity consistency score was higher for experts than for novices. This suggests that experts are more likely to give similar responses on similar cells. This suggests that the mistakes made by experts are less random and are more biased. This is consistent with other analysis made on the same data-set in Trueblood et al. (2018). Therefore, aggregating responses may not be as beneficial for experts. Hence, the efficacy of the algorithm depends on the decision making mechanisms of the underlying population. This indicates that it is important to study the efficacy of various algorithms on the population for which it is intended. However, the higher SCR might also be due to their higher accuracy. Future simulations could vary the similarity consistency score while maintaining the same accuracy to test whether similarity consistency score can predict improvement.

In addition to applications related to developing image sets for training medical AI, the approach discussed in this paper might also have applications to medical education and training. For example, one might use the representational space to design training procedures for medical students and laboratory professionals where example images are sampled intelligently from the representational space. Further, SBA could be used to develop de-biasing procedures where one identifies a certain area of the representational space that an observer consistently gets wrong.

Acknowledgement

This work was supported by NSF grant 1846764.

References

Al-Dulaimi, K. A. K., Banks, J., Chandran, V., Tomeo-Reyes, I., & Nguyen Thanh, K. (2018). Classification of white blood cell types from microscope images: Techniques and challenges. *Microscopy science: Last approaches on educational programs and applied research (Microscopy Book Series, 8)*, 17–25.

Hasan, E., Eichbaum, Q., Seegmiller, A. C., Stratton, C., & Trueblood, J. (2021a). Harnessing the wisdom of the confident crowd in medical image decision-making.

Hasan, E., Eichbaum, Q., Seegmiller, A. C., Stratton, C., & Trueblood, J. S. (2021b). Improving medical image decision-making by leveraging metacognitive processes and representational similarity. *Topics in Cognitive Science*.

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.

Nissim, N., Dudaie, M., Barnea, I., & Shaked, N. T. (2021). Real-time stain-free classification of cancer cells and blood cells using interferometric phase microscopy and machine learning. *Cytometry Part A*, *99*(5), 511–523.

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018). Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. *Journal of Experimental Psychology: General*, *147*(3), 328.

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.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, *42*(8), 2648–2669.

Press, G. (2021). Centaur labs gets $15 million to improve data for healthcare ai. *Forbes*.

Richie, R., & Bhatia, S. (2021). Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science*, *45*(8), e13030.

Sanders, C. A., & Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, 1–23.

Singh, P., Peterson, J. C., Battleday, R. M., & Griffiths, T. L. (2020). End-to-end deep prototype and exemplar models for predicting human behavior. *arXiv preprint arXiv:2007.08723*.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, *31*(1), 137–149.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Trueblood, J. S., Eichbaum, Q., Seegmiller, A. C., Stratton, C., O'Daniels, P., & 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., ... 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.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

Zhuang, J., Cai, J., Wang, R., Zhang, J., & Zheng, W.-S. (2020). Deep knn for medical image classification. In *International conference on medical image computing and computer-assisted intervention* (pp. 127–136).