Expertise is Associated with Higher Idiosyncratic Biases in Medical Image Perception

Julien Vignoud

Master's Thesis
for the obtention of a

Master of Science
(M. Sc.)

in
Data Science

submitted on the 1st of September 2023

Supervisors	Prof. David Whitney
	Ph.D. Candidate Zhihang Ren
Examiner	Prof. Dr. Mary-Anne Hartley
	Prof. Martin Jaggi

Abstract

Background. Expertise is an essential component of medical image perception tasks such as skin cancer screening. Previous studies have identified variations in medical image perception performance focusing on clinician-level measures of accuracy, sensitivity, or other diagnostic performance metrics to characterize expertise and analyze observer biases, for example concluding uneven diagnostic performance between clinicians.

Aim. In this study, we conducted analyses combining computer vision and human diagnostic data relying on skin lesion images to measure individual differences at the diagnosis-level, investigating clinician idiosyncratic biases, i.e., consistent diagnostic error patterns characteristic of an individual.

Methods. Our analysis relied on image categories formed via a deep learning encoding method novel to the study of diagnostic performance. To measure individual differences, we compared intra- and inter-participant diagnostic correlations over distinct image groups and analyzed the relationships between idiosyncratic biases, expertise and diagnostic ambiguity.

Findings. First, both visual and quantitative results indicated significant individual differences in skin cancer diagnostic performance. Additionally, visualization revealed more details about participants' individual differences in fine-grain image clusters. When comparing the responses of experts and amateurs, we found that not only, both amateurs and experts presented idiosyncratic biases, but also that experts displayed a significantly greater effect than amateurs in contentious diagnostic settings.

Conclusion. Our results suggest a potential systematic cause of diagnostic errors, deepening our understanding of the mechanisms underlying expertise, and we identify potential solutions to improve skin cancer screening capabilities.

Acknowledgments

I am deeply grateful for the opportunity to work at UC Berkeley that Prof. Whitney provided me within his lab. Dave's insightful comments and suggestions were essential at every stage of the project.

I would like to offer my special thanks to Peter, who was a pleasure to collaborate with throughout my thesis. I also thank Cindy for her valuable help and explanations. My gratitude extends to all the Whitney lab members, for their warm welcome, and the deliciously stimulating lab lunch discussions.

Finally, I would like to express my sincere thanks to Annie, who graciously accepted to co-supervise me and to take part in the defense jury.

Contents

1 Introduction

1.1 Background

Underlying clinicians' diagnostic and treatment decisions, medical images hold an essential position. The diagnostic process involves two fundamental steps: visually examining the image and rendering an interpretation. Unfortunately, the occurrence of errors in interpreting medical images is not negligible. Such errors can have significant impacts on patients' lives, emphasizing the importance of studying how medical professionals interact with image information during the interpretation process. By gaining a better understanding of this process, we can devise strategies to enhance decision-making and ultimately improve patient care.

Interpreting medical images is necessary for they are not self-explanatory. Medical images present substantial variability, even within the same examination type. Anatomical structures can obscure crucial clinical features, like a lung tumor partially concealed by a rib or hidden behind the heart. Moreover, lesions with very low prevalence affects the decision-making process. For instance, in skin cancer screening, there might be only one cancer detected for every 1,000 cases examined [Eis+14]. Consequently, each case presents notable differences, encompassing a range of abnormalities and normal characteristics that the interpreter must consider attentively.

The intricacies involved can result in interpretation errors, and clinicians are not exempt from making mistakes [Ber05; Ber07; Ber09]. In radiology alone, estimates suggest that certain areas might have up to a 30% miss rate and an equally high false positive rate. Errors can arise in identifying abnormalities, such as determining whether a skin lesion is benign or malignant, or discerning between pneumonia and an alveolar collapse. Undeniably, such errors can significantly impact patient care, leading to delays or fatal misdiagnoses.

A major area of research focus revolves around understanding the contribution of human perception's inherent limitations to these errors, though it remains not very well understood. Image perception likely stands as the most prominent yet underrated source of error in diagnostic imaging. The frequency of image reading errors in malpractice litigation is just one example of this lack of awareness.

Expertise in medical image perception related stimuli and tasks is understood to be critical. As one may expect, expertise largely determines whether clinicians can

perform well in clinical diagnostic tasks [Kru10; Kun06; SK18]. However, studies have repeatedly demonstrated that different clinicians can vary significantly in their diagnostic performance [BLS96; Elm+02; Elm+09; Elm+94; EWH98; Fel+95; Laz+06; Tan+06]. For example, a study on the prenatal detection of malformations using ultrasound images demonstrated that the sensitivity ranged from 27.5% to 96% among different medical institutes [Sal+08]. Similarly, skin cancer screening sensitivity can vary greatly between clinicians, particularly due to the range of professions performing screenings [KLP89]. It is of crucial importance to understand what accounts for these individual variations of performance, and subsequently update the specific training or selecting criteria to better improve clinicians' diagnostic accuracy.

One characteristic of clinicians' expertise in medical image perception, which has been intensively studied in the past, is the visual sensitivity of clinicians [Bir15; Cor11; Lan+15; SDG17; SDG18; Smo+84]. Visual sensitivity, or visual discriminability, here refers to the clinicians' visuospatial and object recognition skills, which contribute to the individual variations in diagnostic performance. Sensitivity differences could originate from genetic variations that affect basic visual perceptual abilities of human observers [Wan+18; Wil+10; Zhu+10; Zhu+21], as well as variability in clinician experience and training [Ber+02; EWH98; Lin+92; Man+06; Mol+08; Ros+16].

1.2 Related work

Another under-explored and non-exclusive reason accounting for perception variability are visual biases of individual clinicians. In the past decade, accumulating research has revealed that untrained observers can have many visual biases [FW14; OPG18; Tip85] and these biases can vary strongly from individual to individual [Cre+20; Cre+21; CW20; Grz+17; KR11; Sch14; Wan+12a; Wan+22; WDM15; Wil+10; Wil17; WMW20]. These idiosyncratic biases exist at every level of human visual perception, from the lowest level such as localization, motion, and color perception [Eme+19; Kan+18; KW17; Sch14; WDM15; WMW20], to higher-level objects [Cre+20; Cre+21; CW20; Ric+19; Wil+10]. For instance, despite extensive exposure to faces, human observers vary drastically in their face recognition abilities [BHB16; DN06; RCN12; RDN09; Wan+12a].

Some recent studies started to throw light on this topic and revealed that clinicians too, as human observers, have their own visual biases towards medical images [Man+21; Ren+23], which could serve as a non-exclusive, alternative origin for the substantial individual differences in diagnostic performance. For example,

clinicians may be subject to so-called serial dependency, due to which an image assessment is influenced by directly preceding images. Manassi et al. showed that form recognition of artificially generated tumors were biased towards previously observed tumors.

However, the precise relationship between individual visual biases and diagnostic performance remains unanswered in the existing literature, and the association between biases and expertise raises intriguing questions.

Intuitively, we would expect biases to diminish or even disappear among experts. This notion aligns with numerous studies suggesting that training has the capacity to reduce visual biases and improve visual perception [DL17; Gam+23; Hai+06; HBJ11; Her+06; NGL16; VO85]. Yet, a recent study hinted at the opposite on non-diagnostic tasks, noticing that radiologists were subject to stronger individual biases than untrained observers when participants were asked to recognize and match artificially generated tumors [Wan+22]. While Wang et al. studied form recognition, it is still unclear whether diagnostic performance in medical experts can be directly linked to perceptual biases, and further research is needed to explore and shed light on this aspect.

1.3 Aim and objectives

We analyzed a large dataset of dermatological judgments collected through a digital medical training application containing 758, 139 melanoma diagnoses from 1, 173 medical trainees, relying on 7, 818 quality-controlled dermoscopic images of skin lesions.

Skin cancer is the most prevalent type of cancer and melanoma, specifically, is responsible for 75% of skin cancer deaths, despite being the least common skin cancer. In 2023, it is estimated that 97,610 individuals in the United States will be diagnosed with invasive melanoma and 89,070 will be diagnosed with melanoma in situ [Cok+05]. Approximately 7,990 individuals will die of melanoma in the US during 2023. Visual inspection is usually the first of a series of 'tests' to diagnose melanoma. Not recognizing a melanoma when it is present delays surgery to remove it, risking cancer spreading to other organs in the body and possibly death [Din+96]. Therefore, the understanding of perceptual errors in skin cancer screening cannot be understated.

Accordingly, we studied clinician idiosyncratic biases when diagnosing skin cancer images and investigate the relationship between visual biases and expertise. In the context of our study, idiosyncratic biases refer to systematic diagnostic errors characteristic to an individual, that is, error patterns not reflected by the majority.

Dermatological judgments are ideal for addressing a possible association between expertise and perceptual biases because images of skin lesions are naturally limited to two-dimensions (non-volumetric) within the visual modality, and such images are available at a large scale. To isolate the nature of potential idiosyncratic biases, we characterized the individual stimulus-level effects using a deep computer vision model. This novel approach leverages deep learning to incorporate image content information in our analysis and break down clinicians' biases based on image semantic information.

It is likely that medical images vary in their ambiguity, and thus vary in difficulty and uncertainty [Son+15]. Because visual biases can be exaggerated when uncertainty increases [FW14; KW17], it is conceivable that idiosyncratic biases manifest under more difficult circumstances. We employed a novel image clustering technique to perform content-based image analysis and further investigated whether individual differences in diagnostic biases remains homogeneous across different types of lesion images.

Precisely, we aim to answer the following research questions:

- Do medical professionals present diagnostic idiosyncratic biases when assessing skin lesion images?
- Can expertise mitigate these individual biases?
- How do idiosyncratic biases vary in context of diagnostic uncertainty? Will participants show magnified biases, or will they make similar mistakes?
- Does expertise play a role in bias magnitude when assessing contentious images?

2 Methods

2.1 Datasets and participants

The data used in this research is composed of two main datasets: the skin lesion image dataset and the skin lesion diagnosis dataset.

The image dataset comprises 7,818 pigmented skin lesion images along with participant melanoma diagnoses. The pigmented skin lesion images originate from the International Skin Imaging Collaboration (ISIC) Archive [Cod+18; Com+19; TRK18] which is the largest publicly available collection of quality-controlled dermoscopic images of skin lesions. This set of images contains two types of lesion, nevus and melanoma, indicating benign and malignant cases respectively. The skin lesion images are dermoscopy images (i.e. collected via a dermatoscope) that underwent manual correction of color hue, luminance, and alignment and were collected by different devices using polarized and non-polarized dermatoscopy. Samples of skin cancer image stimuli are shown in Fig. 2.1.

Skin lesion diagnoses were collected through DiagnosUs, an app developed by Centaur Labs, a US medical Artificial Intelligence (AI) company based in Boston, MA. The diagnosis dataset contains skin lesion image ID references to the image dataset, participant anonymous IDs, diagnoses submitted by participants, and response times. The diagnosis dataset contained 758,139 diagnoses from 1,173 participants. Among all diagnoses, there were 434,089 benign lesions (57.3%), i.e., "nevus", and 324050 malignant lesions (42.7%), i.e., "melanoma". The 25th, 50th, and 75th percentiles of the number of diagnoses per participant were respectively 48, 136, and 403. Yet, the maximum number of diagnoses per participant was 33,786. Despite the relatively large number of participants, these statistics illustrate how only a minority of participants produced enough diagnoses to be reliably studied.

The participants were mostly composed of medical students, with some medical residents. Individual subject information such as age or sex is not known. All participants have normal or corrected-to-normal vision, and users must be located in the U.S. in order to use the app. Users receive earnings from a predefined money pool (around 50 USD) for each task they complete.

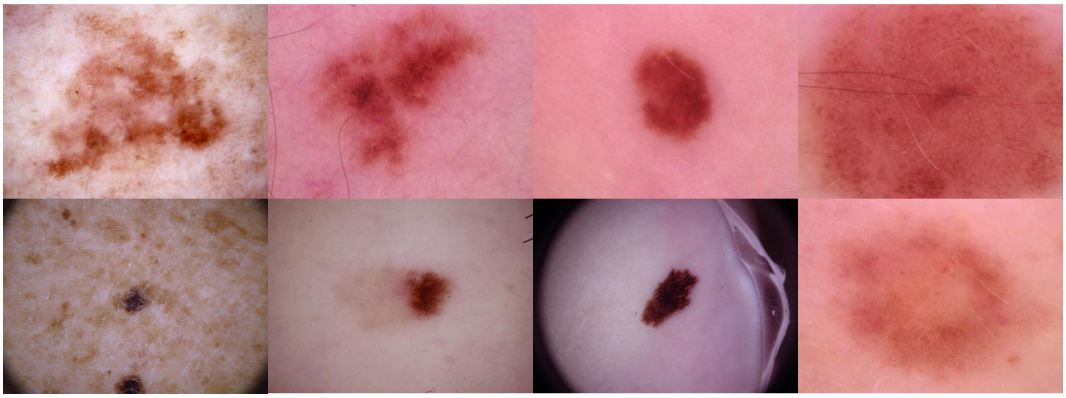

Figure 2.1: Examples of skin lesion images presented to participants

2.2 Task

After downloading the DiagnosUs app and giving consent to have Centaur Labs use the data they provide through app usage, users can choose between different tasks. For the dermatological classification task that was investigated in this study, users first completed a training session of 10 trials with 10 separate stimuli. This training explained the procedure of the task and prepared users for the actual diagnostic task. In each trial, a random skin lesion was drawn from the image dataset and presented to the participant. Below each image, participants were prompted to choose one of two possible responses, "benign" or "malignant". No time limitation was enforced. Immediate feedback was provided after every trial to inform users if their response was correct or incorrect. Afterward, users voluntarily moved on to the next trial at their own pace. Users were told they could end the task at any time.

Note that images being selected randomly, participants did not diagnose the exact same sets of images. This is a key challenge for our analysis, as we can not rely on a common set of images diagnosed by every participant.

2.3 Image embeddings

Computer vision model

In order to perform an image content-based analysis, we leveraged a deep computer vision model to learn new representations, also called embeddings, of the skin lesion images. Intuitively, we extracted deep learning embeddings to create a measure of

semantic similarity between images. Two images may be completely different at the pixel level and yet represent the same melanoma, for example with a shift of camera or with different lighting. Embeddings may also capture a degree of malignancy; even though images have binary labels, there is a physiological continuum between benign and malignant skin lesions. Embedding semantic properties are empirically explored in Section 3.1.

To learn image embeddings, which in practice are vectors of numbers, we relied on a deep learning classifier's last hidden layer. Deep computer vision classifiers extract image information throughout successive hidden layers, information that is then used for classification at the output layer. Thus, the state of the last hidden layer contains all the information necessary for the classification layer to predict the malignancy of an image [LBH15]. By using a deep learning model's internal representations, we expected image embeddings to capture semantic information such as a measure of malignancy necessary for a successful classification [BCV13]. As a result of this representation learning process, conceptually similar images according to the deep learning classifier can produce embeddings that are spatially adjacent.

We relied on a automated melanoma classification challenge organized by the SIIM-ISIC [Rot+21], a collaboration between the Society for Imaging Informatics in Medicine (SIIM) and the ISIC, aiming to improve and automate the diagnosis of melanoma. The image dataset used to train machine learning models within this competition was the same as our image dataset used to collect human diagnoses. That is, participants of the automated classification challenge trained their models on the same dataset we used to collect human diagnoses.

We leveraged the winning model of the competition [HLL20], reaching 0.95 of AUROC (area under the receiver-operator curve) on the challenge leaderboard. Similarly to other models on the leaderboard, it virtually perfectly classified all skin lesions. This deep learning model is an ensemble of sub-models with slightly different architectures, each trained to classify benign and malignant images. The final malignancy labels are computed by averaging the probability predictions of each sub-model. In order to preserve the model's representational power, i.e. potential information captured by image embeddings, we only used one sub-model (which was the best-performing). Alternatives such as aggregating all sub-model's embeddings do not guarantee the preservation of semantic information in the final image embeddings. Indeed, each sub-model may capture different information in the image embeddings, and aggregating may simply negate or cancel them out. Imagine one dimensional embeddings measuring image malignancy. While one sub-model may assign positive values to nevus, another may predict nevus through

negative values. Thus, averaging embeddings from different sub-models may result in a degradation of information.

Unsurprisingly, the model is a variant of convolutional neural networks (CNN). Briefly, the basic components of CNNs consists of three types of layers, namely convolutional, pooling, and fully-connected layers, represented in Fig. 2.2. The convolutional layer aims to learn feature representations of the inputs [Gu+18]. The pooling layer aims to achieve translation-invariance by reducing the resolution of the so-called 'feature maps' resulting from previous layers. As mentioned previously, two images different at the pixel-level may show the same skin lesion with a simple shift of camera. Pooling layers allow extracting similar internal representations between identical yet translated images (robust to translation). The typical pooling operations are average and max pooling, and pooling layers are usually placed between two convolutional layers [Wan+12b].

After several convolutional and pooling layers, there may be one or more fully-connected layers which aim to perform high-level reasoning [Hin+12; SZ14; ZF14]. They take all neurons in the previous layer and connect them to every single neuron of the current layer to generate global semantic information. As a matter of fact, the model used in this study only contained one fully-connected layer, and it is the state of this layer (numerical values of neurons) that we used as image embedding. Thus, the size of our image embeddings are defined by the size of the fully-connected layer (number of neurons), which was in fact 2048. The last layer of CNNs is an output layer. For classification tasks as in the present case, the softmax operator is commonly used [Den+09]. The classification layer is used to train and evaluate the model, and classify new images but was discarded when extracting image embeddings.

More specifically, the CNN used in this study belongs to the EfficientNet family of models [TL19]. EfficientNets are models whose dimensions are defined through a precise scaling method. Tan et al. show that better performance can be reached more efficiently by scaling the network's width (number of feature maps), depth (number of layers) and resolution (height and width of the feature maps) uniformly through a single compound coefficient.

Human participants and deep learning model respectively diagnosed and classified images from the same dataset, yet it is important to note that the deep learning model has been trained solely on skin lesion images and their associated gold standard diagnoses. That is, the model has been trained entirely independently of human diagnosis. Therefore, resulting image embeddings are obtained independently of human diagnosis, and more importantly, independently of human diagnostic performance.

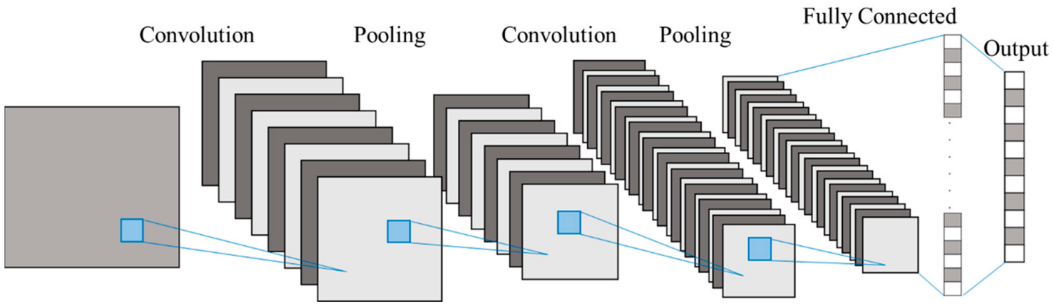

Figure 2.2: An example architecture of a convolutional neural network [Gu+18; LLH20]. Convolutional layers and pooling layers alternate before one or multiple fully-connected layers. Only one layer is depicted in this figure. We can see the number of feature maps increasing and the feature map resolution (height and width) decreasing throughout the pooling layers. The output layer predicts probabilities for the alternative categories, for example benign or malignant. Our model only had one fully-connected layer and has been trained to recognize 9 categories of skin lesions, such as nevus (benign), melanoma (malignant) or seborrheic keratosis (benign), which were then translated into benign or malignant labels.

Dimensionality reduction

Once image embeddings extracted, we leverage the dimensionality reduction technique t-SNE to map high-dimensional embeddings into a more convenient 2-D space while preserving as much structural information as possible.

Dimensionality reduction serves multiple purposes: it decreases noisy and redundant embedding information, it alleviates the curse of dimensionality occurring when measuring distances in a high-dimensional space [SG17] and allows us to visualize the embeddings [VPV+09].

t-SNE is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure, such as the presence of clusters at several scales [VPV+09]. We were interested in visualizing the relative location of embeddings and identifying potential clusters and patterns of images. Hence, why t-SNE is relevant to understanding the information captured through image embeddings.

To do so, pairwise distances between embeddings are modeled through two probability distributions: one represents embedding similarities in the initial high-dimensional space (domain of the function) and the other in a low-dimensional space (codomain). t-SNE then learns a mapping from high to low dimensions by minimizing the mismatch between the two distributions (using the Kullback-

Leibler divergence). In other words, t-SNE optimizes the mapping such that relative distances in low-dimensional are as faithful as possible to the high-dimensional ones, aiming to preserve local and global spacial structures. From now on, "embeddings" refer to image representations in the two-dimensional space.

We parameterized the t-SNE with 2 components (dimensions), a perplexity of 30, 1000 iterations and the Euclidean distance metric.

Clustering

In order to measure participants' idiosyncrasies in their diagnostic performance, we grouped images into semantic clusters defined using image embedding distances. Within image clusters, we then analyzed participants' diagnostic accuracy. Intuitively, image clusters serve as diagnostic performance evaluation axes, which we used to compare and analyze error patterns. The use of image clusters is also motivated by the lack of a common set of images diagnosed by every participant. Clusters aggregate image information and allow comparisons between participants who didn't diagnose the same exact set of images.

In practice, we used the K-means algorithm [Llo82] to cluster all image embeddings into 100 clusters. The algorithm is iterative. It successively assigns each embedding to the nearest cluster mean (centroid) and subsequently updates cluster means until assignments no longer change. We performed 10 algorithm runs with random assignment initialization, a maximum of 300 iterations and a convergence tolerance of 10^{-4}. The number of clusters is a trade-off between the number of images in each cluster, i.e. the reliability of measures within clusters, and the granularity of our analysis where too few clusters may 'over-smooth' important information. Choosing to form 100 clusters assigned around 80 images per cluster in average.

To summarize the overall method, we extracted deep learning image embeddings to identify groups of similar images useful to analyze human diagnostic idiosyncrasies. As stated previously, image clusters were obtained independently of how well human participants diagnosed images.

2.4 Evaluation metrics

The gold standard diagnostic test is used as ground truth, with melanoma diagnoses defined as positive instances and nevus diagnoses as negative instances. Table 2.1 defines the evaluation metrics used in our study.

Metric	Definition
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity, hit rate (H)	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
False alarm rate (F)	$\frac{FP}{TN+FP}$

Table 2.1: Evaluation metric formulas, where TP = True positive; FP = False positive; TN = True negative; FN = False negative.

Building on the hit rate H and the false alarm rate F defined in Table 2.1, we made use of two additional measures relevant to the field of clinical assessment: the discriminability index d' and decision criterion C [MC04].

$$d' = z(H) - z(F)$$
$$C = -\frac{z(H) + Z(F)}{2}$$

Both d' and C are defined in terms of z, the inverse of the cumulative Gaussian distribution. In simple terms, the z-score of a proportion measures how far above the mean the data point is in standard deviation units. Thus, a proportion of .5 (equal to the mean) is converted into a z-score of 0, larger proportions into positive z-scores, and smaller proportions into negative ones.

Intuitively, d' increases when the hit rate H increases and decreases when the false alarm rate F increases, giving each equal weight. When observers cannot discriminate at all, H = F and $d' = 0$. As long as $H \geq F$, d' must be greater than or equal to 0. However, perfect accuracy implies an infinite d', it is therefore common to apply ceilings to H and F: when $H = .99$ and $F = .01$, $d' = 4.65$ [MC04].

While d' stays constant if $z(H)$ and $z(F)$ shift up or down equally, this change is measured by the decision criterion C, the midpoint between $z(H)$ and $z(F)$. An increase in $z(H)$ and $z(F)$ (higher hit rate and false alarm rate) reflects a lower,

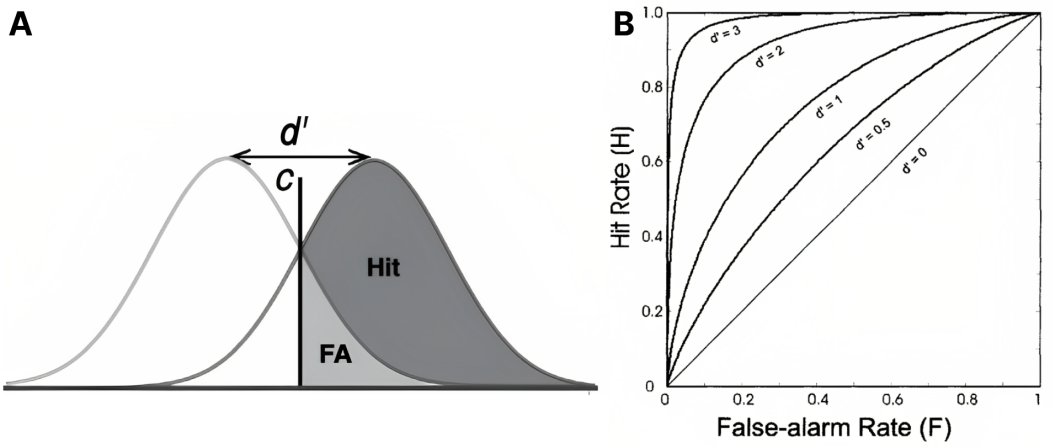

Figure 2.3: (A) Schema of the discrimination index d' and decision criterion C with respect to the Hit rate and the False Alarm rate. The left and right curves represent the correct rejection rate and hit rate probability densities respectively. **(B)** A particular value of d' can also be represented as a receiver-operator curve (ROC) where the (hit rate, false alarm) pairs yield a constant d', so-called iso-sensitive curves.

more relaxed criterion for a positive diagnosis; $C < 0$. If the observer uses a stricter criterion (lower H and F) then $C > 0$. When $C = 0$ the observer is said to be unbiased. Fig. 2.3 illustrates the relationship between discrimination index and decision criterion.

Finally, we measured clusters' standard deviation of participants' accuracy as a proxy for diagnostic "ambiguity". While some image clusters can be uniformly easy (high accuracy) or hard to diagnose (low accuracy), others may show a wide range of diagnostic accuracy (high standard deviation of accuracy), for example with some high performing participants while others showing low diagnostic accuracy (especially accuracy lower than the random accuracy being approximately 0.5). As discussed in following sections, participant's individual biases may be interesting to analyze in the context of such contentious clusters.

2.5 Preprocessing

We first filtered out diagnoses with negative response times as well as diagnoses without an associated gold standard test (1055 diagnoses). Given that response times spanned up to multiple hours, we identified outlier data points using interquartile range, defined as $IQR = Q_3 - Q_1$, for its robustness to extreme values [Dod08;

Wan+14]. Q_1 is the 25th percentile of the response times, and Q_3 the 75th percentile. Outliers were identified as diagnoses with response time lower than $Q_1 - 1.5 * IQR$ or higher than $Q_3 + 1.5 * IQR$, removing 76,051 diagnoses.

To estimate image cluster standard deviations and participants' expertise, we randomly sampled and set aside 25% of each participant's diagnoses. With the remaining 75% data, we filtered out participants without at least 2 trials in each of the image clusters. In other words, we only used diagnoses of participants with at least 2 diagnoses in each of the 100 clusters. The preprocessing left 81 participants and 333,600 diagnoses remaining for analysis. The first subset of diagnoses was solely used to estimate participants' diagnostic performance (accuracy) and cluster's ambiguity, or 'contentiousness' (standard deviation of accuracy) while the second larger subset was leveraged during the diagnosis analysis, thus preventing potential circularity in our study.

2.6 Experiment 1

To quantify individual differences, we compared a participant's internal consistency with the general participant agreement. Participants show idiosyncratic biases if their diagnostic errors are significantly more correlated with themselves than overall participant errors. That is, a participant exhibiting idiosyncratic biases make systematic mistakes, which are characteristic to herself.

The main idea underlying diagnostic pattern analysis consists in computing a participant's diagnostic accuracy (and other metrics) within each of the 100 image clusters, resulting in what we call a "fingerprint": 100 diagnostic measures characterizing each participant. For example, one participant may excel in clusters where others show poorer performance, and vice versa. Given that clusters rely on deep learning embeddings, we expected diagnostic patterns to arise across different image groups.

Before any finer analysis, we wanted to assess whether participants showed any individual biases at all. We conjectured that participants would show higher self-consistency than general agreement. In statistical terms, we tested whether participants' average within-subject correlation was significantly greater than the between-subject correlation.

Internal participant consistency was measured through split-half within-subject correlation of diagnostic accuracy [HT15; Str03]. For each participant, we split cluster's diagnoses in halves, resulting in two 100-value fingerprints per participant. By measuring the correlation between the two fingerprints, we obtained a within-subject correlation. Note that we split participant's data at the cluster level, such

that each half contains approximately the same number of diagnoses per cluster. Splitting data without considering clusters could result in halves containing empty clusters or highly uneven distribution of diagnoses across clusters between two halves. The split-half procedure is illustrated in Fig. 2.4.

In practice, we used a non-parametric bootstrap method to estimate split-half correlations [ET94] where we split halves randomly at each iteration. We measured the Pearson's r correlation coefficient between the two halves. We repeated this procedure 1,000 times and averaged the correlation values across all participants using Fisher's z transformation [SD87] (different from the z-score previously mentioned) to estimate the mean within-subject correlations and 95% bootstrapped confidence intervals.

Between-subject consistency, i.e., inter-participant agreement, was calculated similarly. After splitting every participant's data into two random halves, we correlated halves from different participants. At each iteration, 200 random pairs of participants were sampled, and the pairwise correlations were averaged to estimate the between-subject consistency. By repeating the procedure 1,000 times, we obtained the mean between-subject correlations and 95% bootstrapped confidence intervals.

Next, we estimated the expected chance-level within- and between-subject correlations by calculating permuted null distributions. The idea behind a permutation test is to measure the likelihood of achieving the same results by chance, in this case a certain correlation. By counting the number of times we obtained similar correlation levels via random permutation of data, we estimated the significance of our results. [Dwa57; EO07]

At each iteration, and for each participant's clusters, we again split the diagnoses into two halves, as we did in the bootstrap procedure. We then randomly shuffled the accuracy values across clusters. The resulting correlations from individual participants (within-subject) or different pairs of participants (between-subject) were averaged to get the permuted within-subject or between-subject correlations respectively. This permutation procedure was repeated 1,000 times to estimate permuted null distributions for within-subject and between-subject consistency. The mean empirical bootstrapped correlations were then compared to their corresponding permuted null distributions to estimate the statistical significance of the mean bootstrapped within and between-subject correlations. Schematically, permuted correlations are obtained by simply shuffling each accuracy vector in Fig. 2.4 before computing within and between-subject correlations.

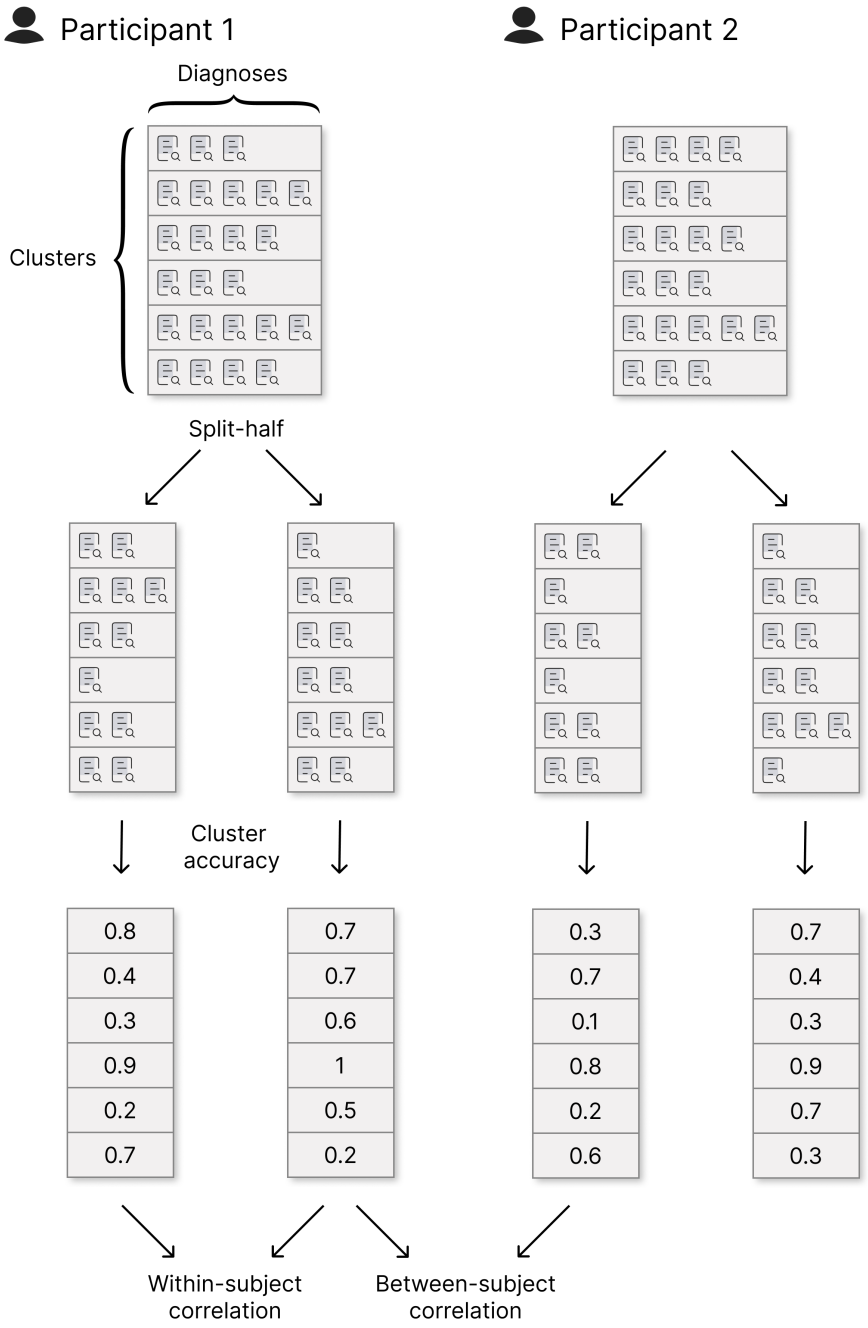

Figure 2.4: Split-half correlation computations. For a given participant, we group her diagnoses into the associated image clusters (6 clusters in this schema). For each cluster, we split the participant's diagnoses in half and computed the cluster accuracy for each half. This process is applied to every participant, resulting in two fingerprint vectors per participant. By averaging the correlations of each participant's own two halves, we obtained an average within-subject correlation. Correlating halves of different participants yields an average between-subject correlation. At each bootstrap iteration, we randomly split-half participants' data and estimate new average within and between-subject correlations.

2.7 Experiment 2

To better grasp the intricacies of idiosyncratic biases, we investigated individual differences as a function of expertise. As a measure of participant expertise, we estimated their diagnostic accuracy using a random sample of each participant's response diagnoses, as described in Section 2.5. Participants were then split into two halves, a "high-performance group" and a "low-performance group". We proceeded with an analysis similar to Experiment 1. For each performance group, we estimated the within-subject and between-subject correlations within the group. Thus, only subjects of the corresponding group were used to compute the group within-subject and between-subject correlations.

In doing so, we tested whether performance groups showed significant idiosyncratic biases and we investigated potential differences in effect magnitude, assessing whether one performance group would show significantly different within-subject or between-subject correlations than the other.

2.8 Experiment 3

Finally, we studied how group idiosyncratic patterns were impacted by image ambiguity. In particular, because a large number of clusters were almost perfectly classified, we were interested in assessing how idiosyncratic biases varied between the groups when inter-participant agreement decreased. To measure disagreement, we identified contentious clusters via the standard deviation of participant accuracy, looking for clusters in which participants showed a wide range of accuracy.

Alternative measures seemingly intuitive, such as low accuracy or high diversity of diagnoses, do not measure disagreement altogether. Indeed, while clusters with lower accuracy can present more disagreement, participants may still unanimously agree with each other while being mistaken. Similarly, measuring the diversity of answers (diagnoses) within clusters does not measure disagreement when clusters contain both benign and malignant images, in which case a diversity of diagnoses is expected. In opposition, a high standard deviation of accuracy entails an extent of disagreement necessary for some diversity of accuracy, while a low standard deviation of accuracy entails a narrow range of accuracy, whether participants are right or wrong. For the sake of simplicity, we will refer to the standard deviation of diagnostic accuracy as "participant disagreement".

Analyzing contentious images was motivated by multiple research questions. How is the internal consistency of participants affected by image ambiguity? Will

the effect differ between low-performers and high-performers? Will there be any group trend as we analyze more and more contentious images?

We conducted the same split-half correlation analysis within each performance group over successive subsets of clusters. Clusters were subsampled based on participant disagreement. We started with all 100 clusters and successively removed clusters with the lowest disagreement by using a lower bound threshold, which we call "disagreement threshold". In other words, we only kept clusters with higher disagreement levels than the threshold. The first batch of clusters filtered out were the image clusters containing skin lesions that were perfectly diagnosed (i.e., easy diagnoses for all observers) or misdiagnosed (in fact this doesn't occur in our case), and the last remaining clusters contained the contentious skin cancer images on which diagnostic accuracy varied the most. We incremented the disagreement threshold from the minimum standard deviation, 0.175 to the maximum 0.49 with 0.025 increments, resulting in 14 measures of idiosyncratic biases.

We tested whether each group presented idiosyncratic biases over the different subsets, and subsequently compared the magnitude of the effect between the two groups. The p-values were adjusted with Bonferroni correction.

3 Results

3.1 Image embeddings

Fig. 3.1A shows skin lesion image samples along with the corresponding embeddings, color-coded by malignancy. Each dot in the figure represents a skin lesion image, and their relative location are defined by the embedding extraction and dimensionality reduction processes. Given that the deep learning model has been trained to classify malignancy and reaches an almost perfect accuracy [HLL20], it is not surprising that benign and malignant images embeddings are easily separable. This is one aspect of semantic similarity captured by these embeddings, they seem to be spatially located according to their image malignancy.

Fig. 3.1B shows the 100 clusters formed by the image embeddings via the K-Means clustering algorithm. Each cluster contains 78 skin lesion images in average. Participants' skin lesion diagnostic performance metrics were evaluated within these clusters. By grouping neighboring embeddings into clusters, we expected images within one cluster to be semantically similar. One aspect of similarity seems to be the malignancy of neighboring images, as illustrated in Fig. 3.1A.

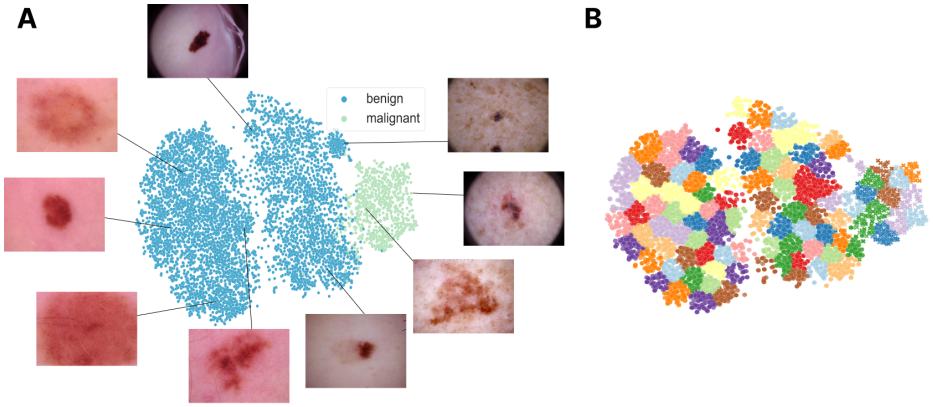

Figure 3.1: (A) Skin lesion samples and their corresponding embeddings. Each dot represents one of the 7,818 skin lesion images. The position of each dot is defined by the internal image representation of the computer vision model. We can see that embeddings of benign and malignant images can be spatially separated. **(B)** The 100 image clusters, represented with different colors. Due to the large number of clusters, some colors occur multiple times.

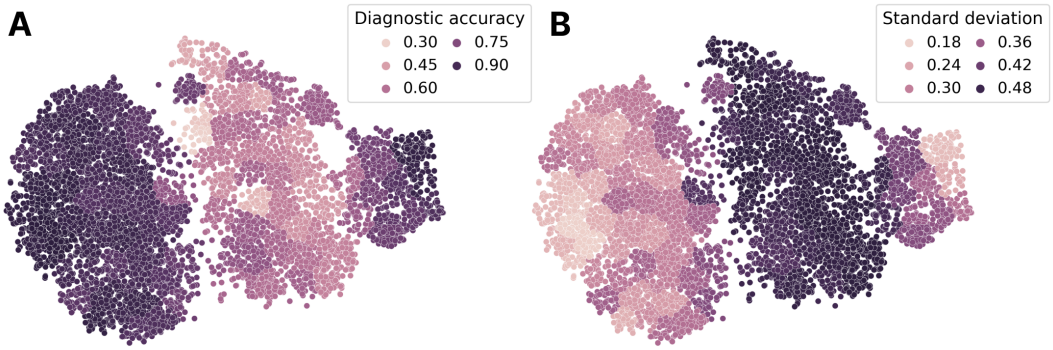

Figure 3.2: (A) Diagnostic test accuracy per cluster across all participants. While dots represent skin lesion images, metrics are color-coded by cluster, i.e., every image dot within a cluster is color-coded according to the cluster accuracy. **(B)** Standard deviation of accuracy per cluster across all participants. Visually, diagnostic accuracy and standard deviation seem to complement each other. This is not unexpected, for a high overall accuracy requires a lower standard deviation of participant accuracy due to a ceiling effect.

Leveraging image clusters, we can now analyze diagnostic performance patterns. Fig. 3.2 depicts average participant diagnostic accuracy per cluster and their standard deviation. Another facet of semantic similarity seems to arise from the image clusters. Three main groups of embeddings (dots) appear in the 2D space, aligned over the x-axis. While the right-most group may be linked to malignant images, the left and middle groups both contain mostly benign images. As shown in Fig. 3.2A and B, they can be associated with different range of accuracy and standard deviations. The left-most group of images seems to contain clusters in which participants show a high accuracy and low standard deviation, while the central group of image contains lower accuracy and higher standard deviation clusters. An observation highlighted when reminded that embeddings are created independently of human performance: it appears that deep learning embeddings also encompass a measure of human diagnostic difficulty and ambiguity. Indeed, while considering the left and central groups only, the embeddings' x coordinate is highly correlated with accuracy and standard deviation; negatively in the former case (Pearson's $r = -0.84$, $p < 0.001$), positively in the latter case (Pearson's $r = 0.91$, $p < 0.001$). Furthermore, the means of accuracy and standard deviation are significantly different between the two groups (t-test, $p < 0.001$ for both accuracy and standard deviation).

To better visualize individual differences, we compared participants' diagnostic accuracy to the average of each cluster. More precisely, within each cluster, we computed the difference between one participant's performance and the average of

all participants. Fig. 3.3 illustrates the accuracy of 5 participants relative to cluster averages. We can clearly see various patterns across clusters between participants. Visually, idiosyncratic biases are denoted by different color distributions between participants, one participant may be particularly good at diagnosing some lesions where another would make more mistakes and vice versa. One objective of our study was to assess whether these patterns of differences were consistent and idiosyncratic enough to be significant. In addition to noticeable deviations from the group performance, and unique patterns between individual observers, it is also clear that there are notably many individual differences in the central group of images, where standard deviation values are high (Fig. 3.2B).

In the Appendix 5, Fig. 5.6 shows individual differences of 30 participants for a broader comparison. Additional performance metrics, such as sensitivity, specificity, d', and the criterion are depicted in Fig. 5.2, 5.3, 5.4, and 5.5 along with the average of each metric across all participants in Fig. 5.1.

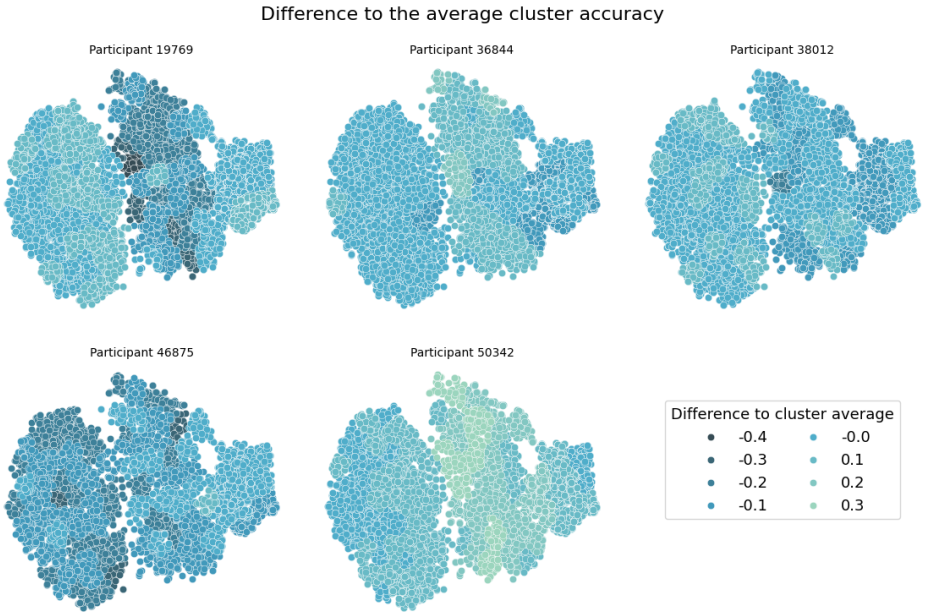

Figure 3.3: Relative diagnostic accuracy per cluster for 5 participants. Each dot represents a skin lesion image. Color coding illustrates participants' diagnostic accuracy compared to the mean performance of all participants within each cluster. Thus, colors highlight where participants diverge from the group average, which is represented in Fig. 3.2A.

3.2 Experiment 1

To formally investigate the diagnostic individual differences, we compared the within-subject correlation and between-subject correlation based on participants' diagnostic accuracy within each image cluster. As represented in Fig. 3.4, we obtained a significant within-participant correlation (orange bar; mean $r = 0.72$, permutation test, $p < 0.001$) and between-participant correlation (blue bar; mean $r = 0.62$, permutation test, $p < 0.001$), rejecting the null hypotheses that the correlation coefficients were a byproduct of chance only.

Importantly, the within-subject correlation was significantly higher than the between-subject correlation (permutation test, $p < 0.001$). This difference denotes that the participants presented individual differences in diagnostic performance, being significantly more consistent in their errors with themselves (intra-participant) than between each other (inter-participant). As we expected, this aligns with previous findings on medical image perception tasks [Wan+22].

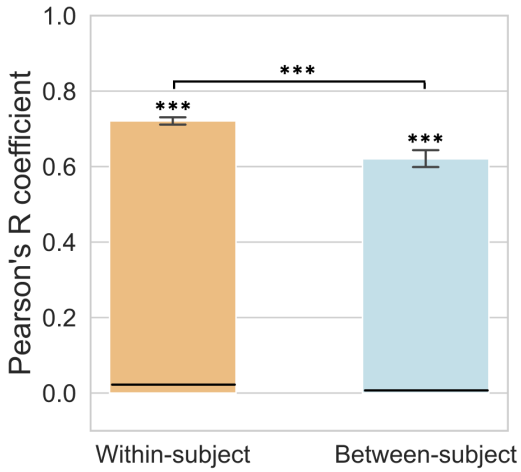

Figure 3.4: Individual differences analysis across all participants. Within-subject correlation and between-subject correlation were averaged across all participants. The within-subject correlation was significantly higher than the between-subject correlation, represented by the horizontal square bracket. Error bars represent the 95% bootstrapped confidence intervals, and the horizontal black lines mark the 97.5% upper bounds of the permuted null distributions for the within-subject and between-subject correlations. $^{***}p < 0.001$.

3.3 Experiment 2

Secondly, we analyzed individual differences as a function of expertise. To measure participant expertise, we randomly sampled a fifth of each participant's response diagnoses to estimate their individual diagnostic accuracy. Participants were then split into two halves, a "high-performance group" and a "low-performance group", each composed of 40 participants. The low-performers showed an average accuracy of 0.74 while the high-performers reached 0.81, the average cluster accuracy of the two groups being significantly different (t-test, $p < 0.01$).

We conducted the same individual differences analysis as performed in Experiment 1 (Fig. 3.5) for each expertise group and found that similar results also hold for both groups, represented in Fig. 3.5A (Fig. 3.5B will be discussed in the following section). For both low- and high-performance groups, within-participant and between-participant correlations were significant (Fig. 3.5A, permutation tests, $p < 0.001$). Moreover, the within-subject correlation was significantly higher than the between-subject correlation for each group (permutation tests, $p < 0.001$). Thus, not only low-performers but also high-performers exhibited idiosyncratic biases in diagnostic accuracy. Nonetheless, the high performance group showed both higher self-consistency and group agreement than the low performance group (permutation tests, $p < 0.001$). In other terms, despite higher performance and consistency, experts still displayed significant diagnostic idiosyncratic biases.

At this point, our analysis was limited by a ceiling effect. Numerous clusters were virtually perfectly classified by participants (see Fig. 3.2A), thus confining the magnitude of potential individual differences (an equivalent flooring effect may have arisen were some clusters systematically misclassified). Indeed, if all participants agree, there is less room to compare potential biases between groups. Accordingly, we subsequently focused on images causing more disagreement by incorporating a third variable: diversity of diagnostic performance.

3.4 Experiment 3

We conjectured that individual differences would increase in clusters showing higher participant disagreement. Moreover, we wondered whether high performers and low performers would exhibit similar individual differences in such context. Accordingly, we conducted idiosyncrasy analyses over different subsets of image clusters according to their range of diagnostic performance. The clusters were subsampled based on their participant disagreement. Initially starting with all 100 clusters, we successively removed clusters with little disagreement by using

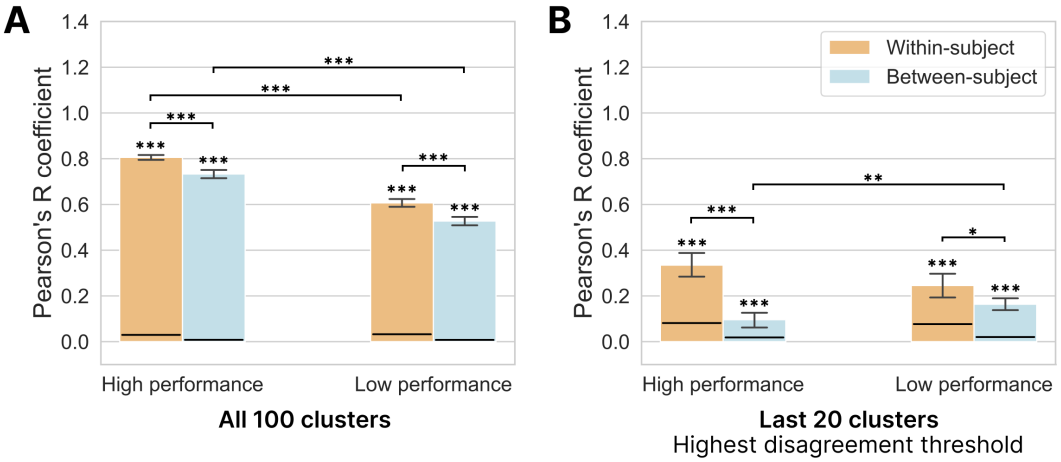

Figure 3.5: (A) Within-subject and between-subject correlations of the low- and high-performance groups. Correlation coefficients were significant (permutation tests, $p < 0.001$), visually represented by the distance between the error bars and their respective horizontal black lines marking the 97.5% upper bounds of the permuted null distributions. For both groups, within-subject correlations were also significantly higher than between-subject correlations, denoting that both low and high performers exhibited idiosyncratic biases. **(B)** Given a disagreement threshold, we filtered out image clusters with lower levels of participant disagreement. Using the remaining clusters, we computed the within-subject and between-subject correlations of each group. Here, we showed the 0th (A) and 100th (B) percentiles, respectively the lowest and highest disagreement threshold used. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

a lower bound threshold. The first batch of clusters filtered out were the image clusters containing skin lesions that were perfectly diagnosed (i.e., easy diagnoses for all observers), and the last remaining clusters contained the most contentious skin lesion images, with the widest range of participant accuracy, where some participants performed well and others worse than random (accuracy lower than 0.5).

Fig. 3.5 showed the first and last cases of the 14 different disagreement thresholds tested. Fig. 3.5A illustrates correlations using all 100 clusters, while Fig. 3.5B analyzes the most contentious image clusters with the highest disagreement threshold. The in-between thresholds will be discussed with Fig. 3.6. Fig. 3.5B illustrates that even over the most polarizing clusters, the within-participant and between-participant correlations were significant (permutation tests, $p < 0.001$), and within-subject correlations were significantly higher than the between-subject correlations (permutation tests, $p < 0.05$). This confirms that despite some group-

wide agreement, there were also significant individual differences (more consistency in the within-observer errors than the between-observer errors). Interestingly, when considering the most contentious clusters (Fig. 3.5B) we found that high-performance participants had a significantly lower between-subject correlation than the low-performance group (permutation test, $p < 0.01$). That is, the high-performers showed more disagreement than the low-performance group. This is a potential sign of higher idiosyncratic biases in the high-performance group. In contrast, for less ambiguous clusters (smaller thresholds), we found that the high-performance group showed higher within-participant and between-participant correlations than the low-performance group (permutation tests, $p < 0.001$), thus both higher self-consistency and group-agreement than low-performers.

In order to analyze the effects of increasing ambiguity, we measured the average difference between within-participant correlations and between-participant correlations, which we call "idiosyncratic bias magnitude". Intuitively, a highly positive idiosyncratic bias magnitude denotes that participants each make their own systematic and individual mistakes, while a highly negative magnitude arises when all participants make inconsistent (no self-consistency) yet similar mistakes (strong group-agreement). We repeated this measure at different disagreement thresholds (different cluster subsets) for both low- and high-performance groups. As a result, we found that the high-performers' idiosyncratic bias magnitude increased more than for the low-performers as images became more contentious. In other words, as the images got more ambiguous to diagnose, the high-performance group showed stronger idiosyncratic biases than the low-performance group, showing less group agreement and more idiosyncratic errors than the low-performers. Fig. 3.6 illustrates this effect. Fig. 3.6A and B explains how we computed the idiosyncratic bias magnitude difference between the two groups, which is then plotted against the increasing disagreement thresholds in 3.6C as the increasing blue line. Thus, the blue line depicts the increasing difference of bias magnitude between the high-performers and the low performers. In contrast, the pink line marks the permuted null distributions. As the gap increases between the magnitude difference (blue line) and the permuted values (pink line), the effect gets more and more significant. High performing individuals therefore have more stable individual perceptual biases. For the cluster subsets with a standard deviation (disagreement) higher than 0.375, we found that the high-performance group showed a significantly higher idiosyncratic bias magnitude than the low-performance group (Bonferroni-adjusted p-value levels can be found in 3.6C). Additionally, we observed a statistically significant positive correlation between the difference of idiosyncratic bias magnitudes and the disagreement threshold (Pearson's $r = 0.91$, $p < 0.001$), further highlighting an association between expertise and higher idiosyncratic biases.

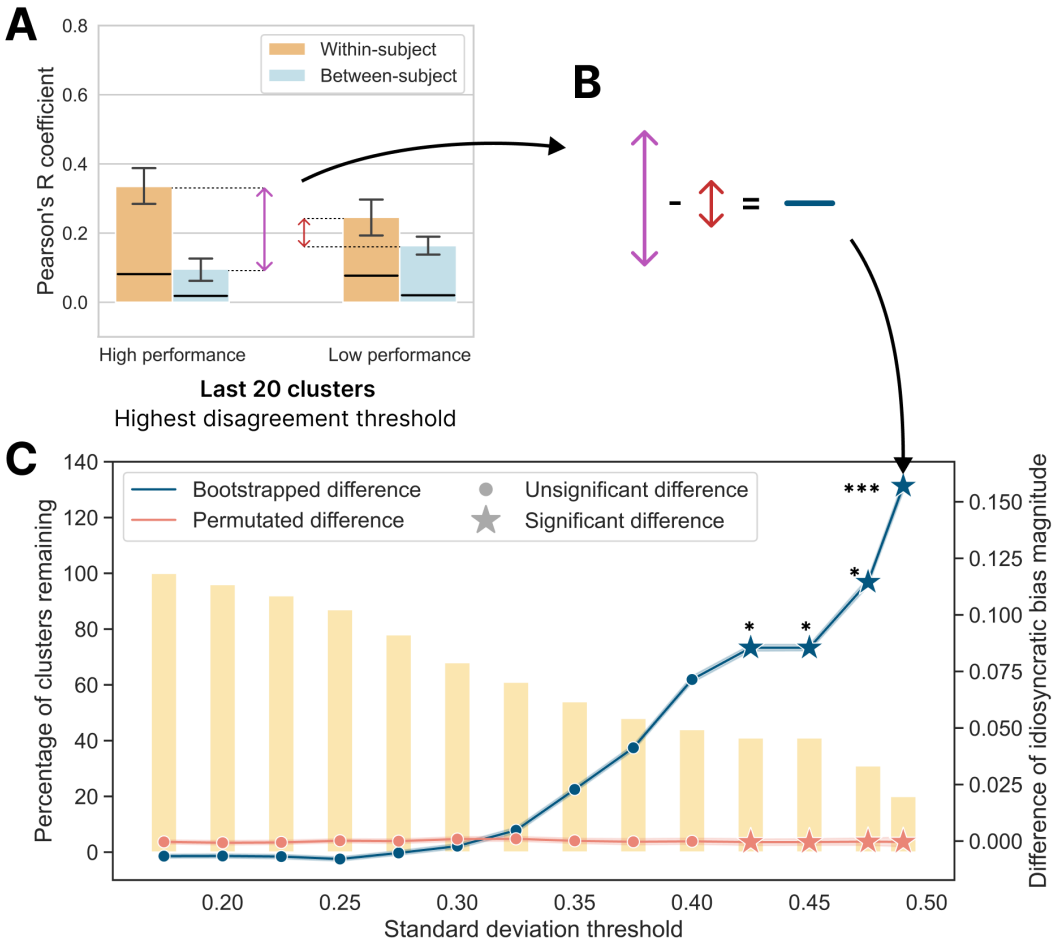

Figure 3.6: Idiosyncratic bias magnitude difference between the two groups with respect to cluster standard deviation thresholds (participant disagreement thresholds). Using the remaining clusters, we computed the idiosyncratic bias magnitude difference. **(A)** Given one subset of clusters (i.e. disagreement threshold) we measured the difference between within-participant correlation and between-participant correlation (idiosyncratic bias magnitude) for each group. **(B)** We then computed the difference of magnitude between the two performance group. Note that **(A)** is the same figure as Fig. 3.5B with a different y-axis range. **(C)** We repeated this procedure for increasing disagreement thresholds. The bootstrapped idiosyncratic bias magnitude and the permutation test values are represented by the dark blue line and the pink line respectively. The yellow columns represent the percentage of remaining image clusters after apply thresholds. Star markers denote where the permutation test is statistically significant, i.e. when the difference between the blue line and pink line is significant. Asterisks represent Bonferroni-adjusted p-value significance with $^{*}p < 0.05$ and $^{***}p < 0.001$.

4 Discussion

In this study, we used a large dataset of teledermatology records to isolate and identify the presence of individual observer-specific biases in the perception of skin lesions. Our results demonstrated that, counterintuitively, expertise is associated with increased idiosyncrasy within individual observers. Rather than becoming more alike and homogeneous, experts tend to have more unique patterns of perceptual bias. The results confirm the importance of expertise, but, more importantly, they reveal the growing importance of individual differences with expertise. The results have important implications for individualized training, paired-reader performance and optimization, bias-mitigation strategies, and the use of computer vision in assessing clinician performance.

To identify and measure individual differences in observer performance, we harnessed a computer vision model in conjunction with 758, 139 skin cancer diagnostic judgements collected from 1, 173 medical trainees. The computer vision model sorted images into nearby clusters based on semantic or content-based similarity, an essential step enabling us to analyze diagnostic performance on a fine granularity. By comparing the high-performers to the low-performers, we found that expertise remains associated with idiosyncratic biases, and medical trainees with better expertise tend to demonstrate more idiosyncratic biases. A counter-intuitive finding, as one may easily assume that biases would decrease with the growing expertise from extensive training and experience. Thus, combining computer vision (establishing skin lesion semantic categories) with behavioral (human diagnostic performance) approaches can lead to novel insights otherwise beyond reach of either individual approaches.

Our results may raise a number of questions that we address in the following discussion. First, it might be argued that stronger idiosyncratic biases exhibited by experts could simply result from the high-performance group being more attentive to the task or lapsing less frequently. By comparing participant reaction times, we found that the time taken to submit diagnoses between the two groups was comparable and not significantly different (t-test, $p > 0.05$). Hence, it is unlikely that the stronger individual differences within the high-performers simply arose due to difference of attentiveness. Furthermore, while higher levels of attention could account for the higher within- and between-correlations of high-performers

in some settings (Fig 3.5A), it may not explain altogether the increased difference of self-consistency and group-agreement displayed by experts.

One might be concerned about the internal consistency of these idiosyncratic biases, that these biases are not systematic. Using the split-half Pearson's correlation, participants had a significant internal reliability of 0.68 (permutation test, $p < 0.001$). When measuring the internal reliability of each group, we found that high-performers reached a correlation coefficient of 0.77 (permutation test, $p < 0.001$) and the low-performers 0.58 (permutation test, $p < 0.001$). Furthermore, we measured a Cronbach's alpha of 0.95, underscoring the high internal consistency of participants' answers.

Leveraging that images were sometimes diagnosed multiple times by one participant, we evaluated participants' test-retest reliability. That is, whether participants answer similar diagnoses when assessing the same image at different times? Via Pearson's correlations, we found that participants showed significant reliability (permutation tests, $p < 0.001$), with r coefficients of 0.44 for all participants, 0.46 when considering only the high-performance group, and 0.40 for low-performers.

One may worry that the skin lesion images included in the experiment encompass only two types of lesions, nevus (benign) and melanoma (malignant), which do not adequately represent the full range of skin lesions. Additionally, among the images presented to participants, a balanced distribution was observed between benign lesions (57.3%) and malignant lesions(42.7%) which contrasts with the true prevalence of melanoma, being much rarer than benign lesions in actual skin cancer screenings. Although the skin lesion types and their prevalence within the dataset deviate from real-world scenarios, the skin lesion images were directly extracted from real diagnostic records, and they include a diverse array of lesions, textures, and collection methods, spanning multiple subtypes of skin lesions. Thus, the size and scope of the dataset is a strength, offering a means to capture the biases that medical professionals exhibit in their day-to-day diagnostic practice. Future studies can expand the lesion categories and investigate the effect of more lesion types in typical skin cancer diagnostic scenarios. Whether disease prevalence may influence individual differences is another interesting question to investigate in future studies.

All the diagnostic data in this study were collected online because our goal was to investigate remote store-and-forward teledermatology. Whether in-person dermatologists might exhibit the same sorts of idiosyncratic biases as a function of expertise remains unclear and should be investigated in future work. With the increasing prominence and adoption of teledermatology, our findings hold significant value in understanding the relationship between bias and expertise in remote medicine. However, we acknowledge that the results here should be

extrapolated to in-person settings because telemedicine is not directly comparable to in-person diagnosis. For example, in the clinic, dermatologists have access to much more information, including tactile cues, larger field of view, and a variety of other sources of information

Another question that might arise is whether time pressure or constraints may have imposed a burden on observers that led to the biases. This is not a likely explanation because when performing the diagnostic task, there was no restriction on the responding time and participants had unlimited time to submit a diagnosis. As such, it is unlikely that individual biases stemmed from time pressure or the need to balance speed and accuracy. Additionally, the stimuli employed in this study consisted of 2D skin cancer images, and it remains uncertain whether the observed effects would hold true for 3D volumetric data. Nevertheless, considering that these individual biases are linked to our visual perception system, it is plausible that comparable outcomes could be obtained in other medical image modalities and imaging techniques. Furthermore, it is important to note that the lesions used in this study are relatively small when viewed on devices, and are primarily located within the fovea and perifovea regions of the retina. As a result, where participants foveate (where they focus their gaze) is unlikely to contribute to the individual biases observed in our study. Nonetheless, future works should explore whether the foveation pattern can contribute to individual biases.

Via visualizations, we explicitly showed how individuals displayed distinct diagnostic "fingerprints". These idiosyncratic fingerprints in medical image perception can be a potential alternative explanation to the variations in medical image diagnosis, besides the visual sensitivity stemming from training or experience mostly discussed in previous studies. Our findings also point towards a new direction for improving clinician' diagnostic performance based on medical image perception. Unlike most other studies that focused on the potential benefit of reducing biases to improve performance [DL17; Gam+23; Hai+06; HBJ11; Her+06; NGL16; VO85], our study showed that the biases displayed by experts can instead distinguish them, motivating the need to further study and possibly leverage individual biases. These increased image-specific biases suggest that experts could have unique subjective perceptions of medical images and their own perceptual statistics and cognitive knowledge underlying higher diagnostic accuracy.

For example, individual bias characteristics may be used to identify personal diagnostic weaknesses over different image categories and adjust clinicians' training accordingly. Another approach to harness individual biases and improve diagnostic capacity may reside in pairing clinicians according to their individual biases. One may expect that two clinicians with different idiosyncratic biases may discuss and produce more accurate and educated diagnoses than two similarly-biased clinicians.

In summary, we found that medical trainees have significant idiosyncratic biases when diagnosis skin cancer via visual inspection, and higher diagnostic accuracy is associated with higher idiosyncratic biases. Our study, together with more and more recent findings that cumulatively demonstrated various visual biases among medical doctors or radiologists, should hopefully draw more attention to this intriguing yet under-investigated research area with high potentials for understanding and improving the diagnostic performance relying on medical image perception.

5 Appendix

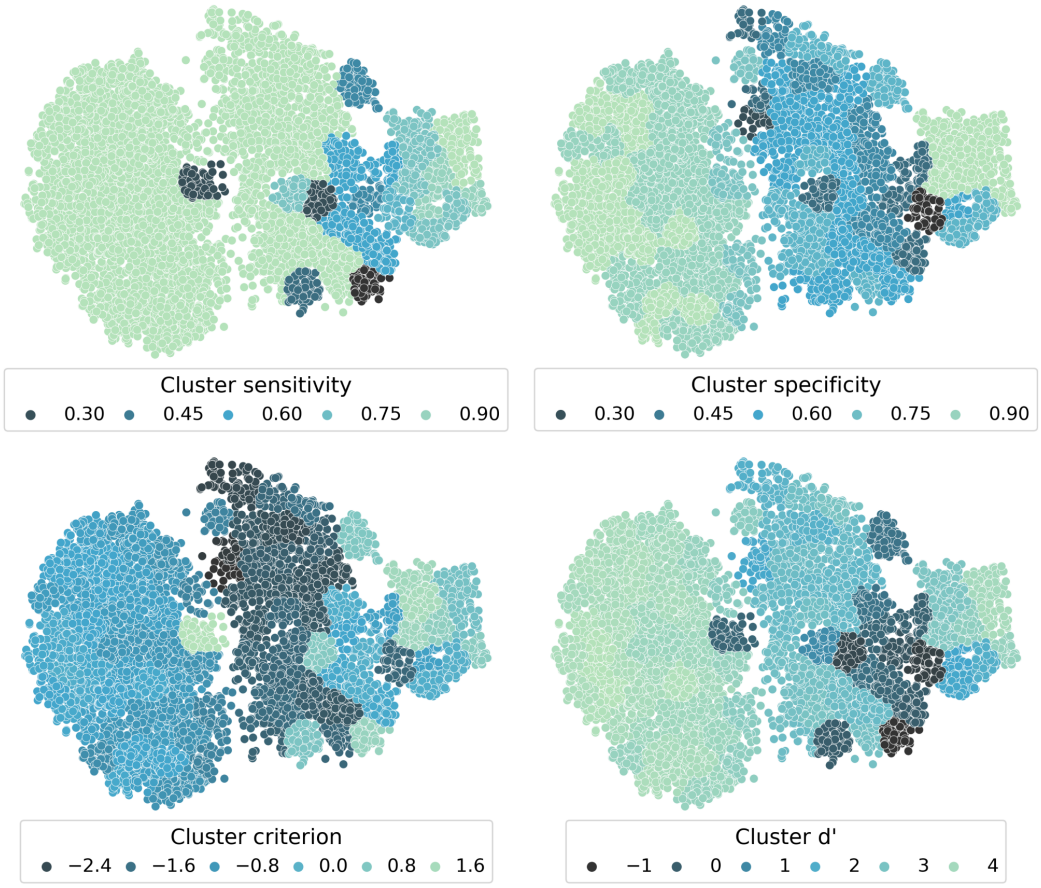

Figure 5.1: Cluster evaluation metrics averaged across all participants. Each dot represents a skin lesion image. Colors encode diagnostic metrics evaluated at the cluster-level, when considering all participants' diagnoses.

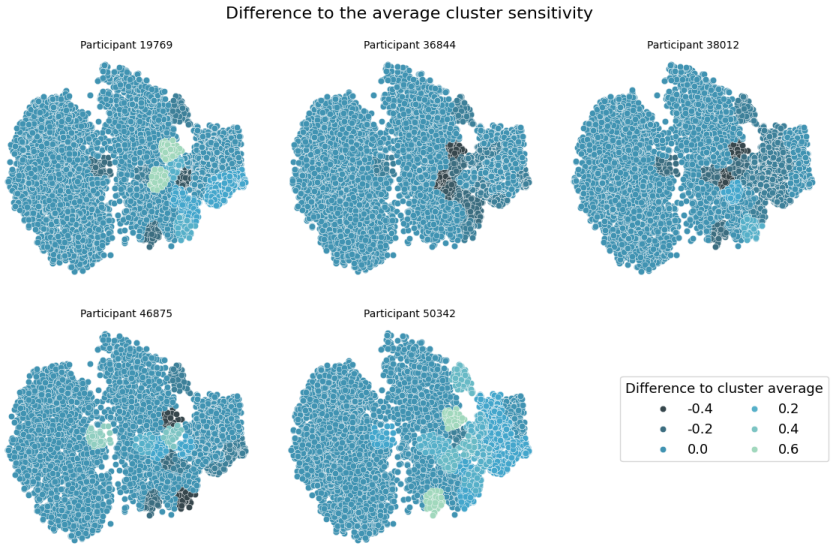

Figure 5.2: Diagnostic sensitivity per cluster for 5 participants compared to the cluster average across all participants. Because we interested in individual differences, after computing diagnostic metrics for one participant at the cluster-level, we compute the difference between this participant and the average of all participants within each cluster.

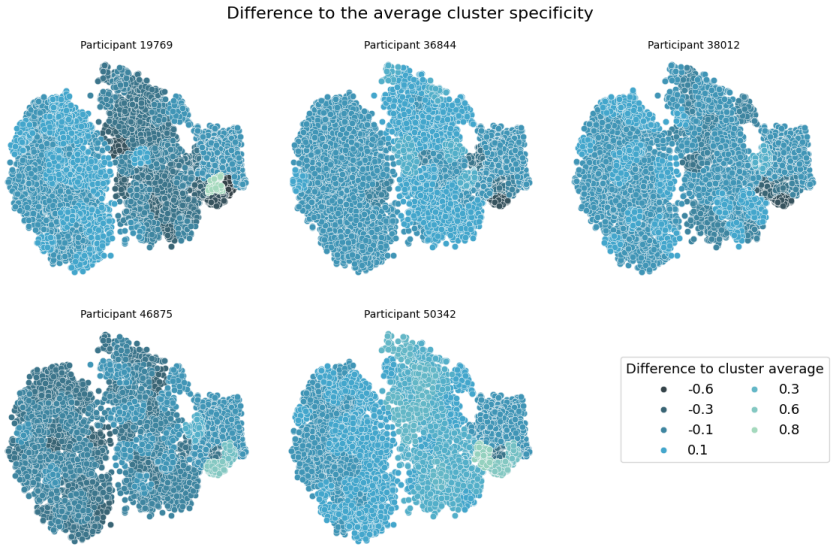

Figure 5.3: Relative diagnostic specificity per cluster for 5 participants

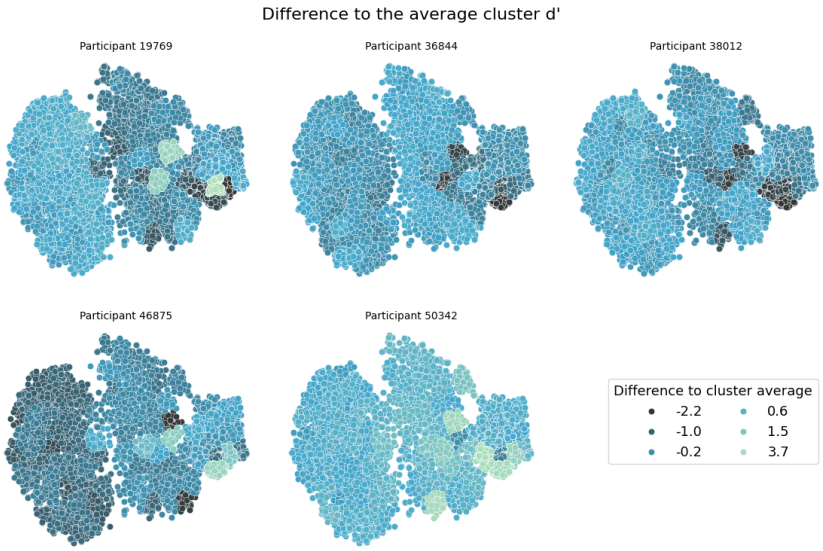

Figure 5.4: Relative diagnostic d' per cluster for 5 participants

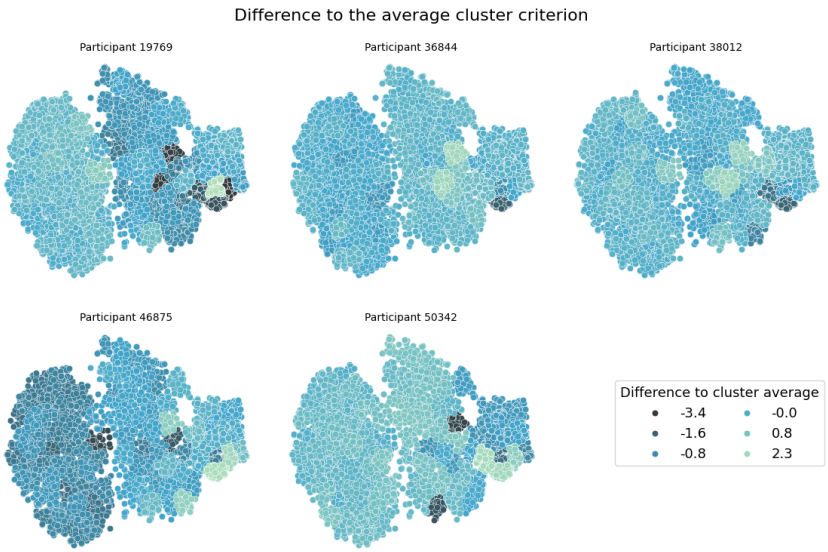

Figure 5.5: Relative diagnostic criterion per cluster for 5 participants

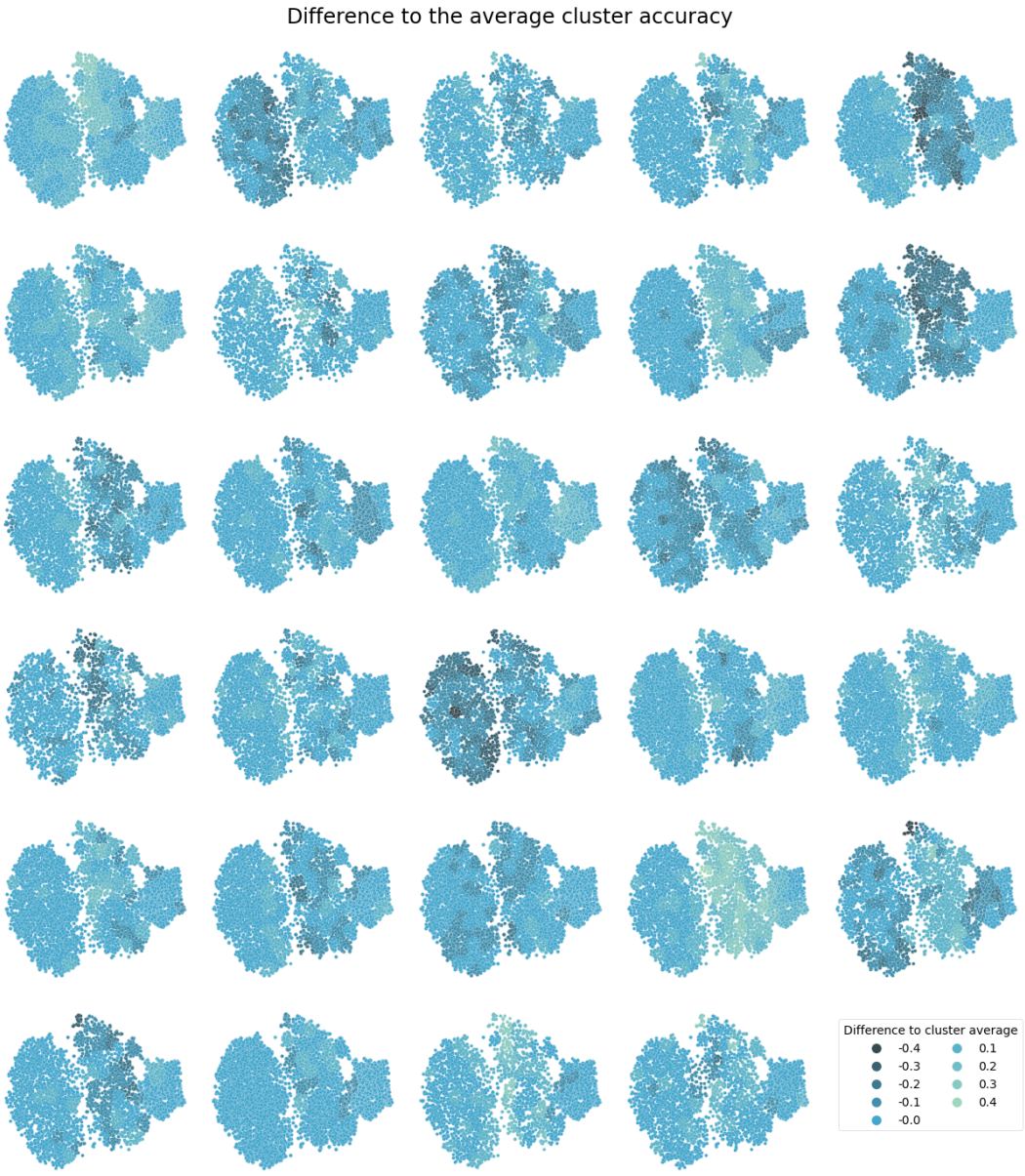

Figure 5.6: Relative diagnostic accuracy per cluster for 30 participants. Each dot represents a skin lesion image. Given that not all participants submitted a diagnosis for each image, some fingerprints contain fewer images (dots) than others. The accuracy (color) is computed at the cluster level.

Bibliography

[BCV13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. **Representation learning: A review and new perspectives**. *IEEE transactions on pattern analysis and machine intelligence* 35:8 (2013), 1798–1828 (see page 7).

[Ber+02] Wendie A Berg, Carl J D'Orsi, Valerie P Jackson, Lawrence W Bassett, Craig A Beam, Rebecca S Lewis, and Philip E Crewson. **Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography?** *Radiology* 224:3 (2002), 871–880 (see page 2).

[Ber05] Leonard Berlin. **Errors of omission**. *American Journal of Roentgenology* 185:6 (2005), 1416–1421 (see page 1).

[Ber07] Leonard Berlin. **Accuracy of diagnostic procedures: has it improved over the past five decades?** *American Journal of Roentgenology* 188:5 (2007), 1173–1178 (see page 1).

[Ber09] Leonard Berlin. **Malpractice issues in radiology: res ipsa loquitur**. *American Journal of Roentgenology* 193:6 (2009), 1475–1480 (see page 1).

[BHB16] Anna K Bobak, Peter JB Hancock, and Sarah Bate. **Super-recognisers in action: Evidence from face-matching and face memory tasks**. *Applied Cognitive Psychology* 30:1 (2016), 81–91 (see page 2).

[Bir15] D Birchall. **Spatial ability in radiologists: a necessary prerequisite?** *The British journal of radiology* 88:1049 (2015), 20140511 (see page 2).

[BLS96] Craig A Beam, Peter M Layde, and Daniel C Sullivan. **Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample**. *Archives of internal medicine* 156:2 (1996), 209–213 (see page 2).

[Cod+18] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. **Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)**. In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, 168–172 (see page 5).

[Cok+05] V Cokkinides, J Albano, A Samuels, M Ward, and J Thum. **American cancer society: Cancer facts and figures**. *Atlanta: American Cancer Society* 2017 (2005) (see page 3).

[Com+19] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. **Bcn20000: Dermoscopic lesions in the wild**. *arXiv preprint arXiv:1908.02288* (2019) (see page 5).

[Cor11] CA Corry. **The future of recruitment and selection in radiology. Is there a role for assessment of basic visuospatial skills?** *Clinical radiology* 66:5 (2011), 481–483 (see page 2).

[Cre+20] Aline F. Cretenoud, Lukasz Grzeczkowski, Marco Bertamini, and Michael H. Herzog. **Individual differences in the Müller-Lyer and Ponzo illusions are stable across different contexts**. *Journal of Vision* 20:6 (June 2020), 4–4. ISSN: 1534-7362. DOI: 10.1167/jov.20.6.4. eprint: https://arvojournals.org/arvo/content_public/journal/jov/938476/i0035-8711-453-1-07044.pdf. URL: https://doi.org/10.1167/jov.20.6.4 (see page 2).

[Cre+21] Aline F Cretenoud, Lukasz Grzeczkowski, Marina Kunchulia, and Michael H Herzog. **Individual differences in the perception of visual illusions are stable across eyes, time, and measurement methods**. *Journal of Vision* 21:5 (2021), 26–26 (see page 2).

[CW20] Teresa Canas-Bajo and David Whitney. **Stimulus-specific individual differences in holistic perception of Mooney faces**. *Frontiers in Psychology* 11 (2020), 585921 (see page 2).

[Den+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. **Imagenet: A large-scale hierarchical image database**. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, 248–255 (see page 8).

[Din+96] Jacqueline Dinnes, Jonathan J Deeks, Matthew J Grainge, Naomi Chuchu, Lavinia Ferrante di Ruffano, Rubeta N Matin, David R Thomson, Kai Yuen Wong, Roger Benjamin Aldridge, Rachel Abbott, et al. **Visual inspection for diagnosing cutaneous melanoma in adults**. *Cochrane Database of Systematic Reviews* 2018:12 (1996) (see page 3).

[DL17] Barbara Dosher and Zhong-Lin Lu. **Visual perceptual learning and models**. *Annual review of vision science* 3 (2017), 343–363 (see pages 3, 29).

[DN06] Brad Duchaine and Ken Nakayama. **The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants**. *Neuropsychologia* 44:4 (2006), 576–585 (see page 2).

[Dod08] Yadolah Dodge. **The concise encyclopedia of statistics**. Springer Science & Business Media, 2008 (see page 12).

[Dwa57] Meyer Dwass. **Modified randomization tests for nonparametric hypotheses**. *The Annals of Mathematical Statistics* (1957), 181–187 (see page 14).

[Eis+14] Nora Eisemann, Annika Waldmann, Alan C Geller, Martin A Weinstock, Beate Volkmer, Ruediger Greinert, Eckhard W Breitbart, and Alexander Katalinic. **Non-melanoma skin cancer incidence and impact of skin cancer screening on incidence**. *Journal of Investigative Dermatology* 134:1 (2014), 43–50 (see page 1).

[Elm+02] Joann G Elmore, Diana L Miglioretti, Lisa M Reisch, Mary B Barton, William Kreuter, Cindy L Christiansen, and Suzanne W Fletcher. **Screening mammograms by community radiologists: variability in false-positive rates**. *Journal of the National Cancer Institute* 94:18 (2002), 1373–1380 (see page 2).

[Elm+09] Joann G Elmore, Sara L Jackson, Linn Abraham, Diana L Miglioretti, Patricia A Carney, Berta M Geller, Bonnie C Yankaskas, Karla Kerlikowske, Tracy Onega, Robert D Rosenberg, et al. **Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy**. *Radiology* 253:3 (2009), 641–651 (see page 2).

[Elm+94] Joann G Elmore, Carolyn K Wells, Carol H Lee, Debra H Howard, and Alvan R Feinstein. **Variability in radiologists' interpretations of mammograms**. *New England Journal of Medicine* 331:22 (1994), 1493–1499 (see page 2).

[Eme+19] Kara Emery, Vicki Volbrecht, David Peterzell, and Michael Webster. **Color vs. motion: decoding perceptual representations from individual differences**. *Journal of Vision* 19:8 (2019), 8–8 (see page 2).

[EO07] Eugene Edgington and Patrick Onghena. **Randomization tests**. CRC press, 2007 (see page 14).

[ET94] Bradley Efron and Robert J Tibshirani. **An introduction to the bootstrap**. CRC press, 1994 (see page 14).

[EWH98] Joann G Elmore, Carolyn K Wells, and Debra H Howard. **Does diagnostic accuracy in mammography depend on radiologists' experience?** *Journal of Women's Health* 7:4 (1998), 443–449 (see page 2).

[Fel+95] Judith Feldman, RobertA Smith, Ruthann Giusti, Barbara DeBuono, John P Fulton, and H Denman Scott. **Peer review of mammography interpretations in a breast cancer screening program.** *American journal of public health* 85:6 (1995), 837–839 (see page 2).

[FW14] Jason Fischer and David Whitney. **Serial dependence in visual perception**. *Nature neuroscience* 17:5 (2014), 738–743 (see pages 2, 4).

[Gam+23] Roberto Gammeri, Selene Schintu, Adriana Salatino, Francesca Vigna, Alessandro Mazza, Patrizia Gindri, Sonia Barba, and Raffaella Ricci. **Effects of prism adaptation and visual scanning training on perceptual and response bias in unilateral spatial neglect**. *Neuropsychological Rehabilitation* (2023), 1–26 (see pages 3, 29).

[Grz+17] Lukasz Grzeczkowski, Aaron M Clarke, Gregory Francis, Fred W Mast, and Michael H Herzog. **About individual differences in vision**. *Vision research* 141 (2017), 282–292 (see page 2).

[Gu+18] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. **Recent advances in convolutional neural networks**. *Pattern recognition* 77 (2018), 354–377 (see pages 8, 9).

[Hai+06] Qi Haijiang, Jeffrey A Saunders, Rebecca W Stone, and Benjamin T Backus. **Demonstration of cue recruitment: Change in visual appearance by means of Pavlovian conditioning**. *Proceedings of the National Academy of Sciences* 103:2 (2006), 483–488 (see pages 3, 29).

[HBJ11] Sarah J Harrison, Benjamin T Backus, and Anshul Jain. **Disambiguation of Necker cube rotation by monocular and binocular depth cues: Relative effectiveness for establishing long-term bias**. *Vision research* 51:9 (2011), 978–986 (see pages 3, 29).

[Her+06] Michael H Herzog, Knut RF Ewald, Frouke Hermens, and Manfred Fahle. **Reverse feedback induces position and orientation specific changes**. *Vision research* 46:22 (2006), 3761–3770 (see pages 3, 29).

[Hin+12] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. **Improving neural networks by preventing co-adaptation of feature detectors**. *arXiv preprint arXiv:1207.0580* (2012) (see page 8).

[HLL20] Qishen Ha, Bo Liu, and Fuxu Liu. **Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge**. *arXiv preprint arXiv:2010.05351* (2020) (see pages 7, 19).

[HT15] Roberta Heale and Alison Twycross. **Validity and reliability in quantitative studies**. *Evidence-based nursing* 18:3 (2015), 66–67 (see page 13).

[Kan+18] Sae Kaneko, Ikuya Murakami, Ichiro Kuriki, and David H Peterzell. **Individual variability in simultaneous contrast for color and brightness: Small sample factor analyses reveal separate induction processes for short and long flashes**. *i-Perception* 9:5 (2018), 2041669518800507 (see page 2).

[KLP89] Howard K Koh, Robert A Lew, and Marianne N Prout. **Screening for melanoma/skin cancer: theoretic and practical considerations**. *Journal of the American Academy of Dermatology* 20:2 (1989), 159–172 (see page 2).

[KR11] Ryota Kanai and Geraint Rees. **The structural basis of inter-individual differences in human behaviour and cognition**. *Nature Reviews Neuroscience* 12:4 (2011), 231–242 (see page 2).

[Kru10] Elizabeth A Krupinski. **Current perspectives in medical image perception**. *Attention, Perception, & Psychophysics* 72:5 (2010), 1205–1217 (see page 2).

[Kun06] Harold L Kundel. **History of research in medical image perception**. *Journal of the American college of radiology* 3:6 (2006), 402–408 (see page 2).

[KW17] Anna Kosovicheva and David Whitney. **Stable individual signatures in object localization**. *Current Biology* 27:14 (2017), R700–R701 (see pages 2, 4).

[Lan+15] Jean Langlois, George A Wells, Marc Lecourtois, Germain Bergeron, Elizabeth Yetisir, and Marcel Martin. **Spatial abilities of medical graduates and choice of residency programs**. *Anatomical Sciences Education* 8:2 (2015), 111–119 (see page 2).

[Laz+06] Elizabeth Lazarus, Martha B Mainiero, Barbara Schepps, Susan L Koelliker, and Linda S Livingston. **BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value**. *Radiology* 239:2 (2006), 385–391 (see page 2).

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. **Deep learning**. *nature* 521:7553 (2015), 436–444 (see page 7).

[Lin+92] MN Linver, SB Paster, RD Rosenberg, ChR Key, ChA Stidley, and WV King. **Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases.** *Radiology* 184:1 (1992), 39–43 (see page 2).

[LLH20] Chang-Cheng Lo, Ching-Hung Lee, and Wen-Cheng Huang. **Prognosis of bearing and gear wears using convolutional neural network with hybrid loss function**. *Sensors* 20:12 (2020), 3539 (see page 9).

[Llo82] Stuart Lloyd. **Least squares quantization in PCM**. *IEEE transactions on information theory* 28:2 (1982), 129–137 (see page 10).

[Man+06] David Manning, Susan Ethell, Tim Donovan, and Trevor Crawford. **How do radiologists do it? The influence of experience and training on searching for chest nodules**. *Radiography* 12:2 (2006), 134–142 (see page 2).

[Man+21] Mauro Manassi, Cristina Ghirardo, Teresa Canas-Bajo, Zhihang Ren, William Prinzmetal, and David Whitney. **Serial dependence in the perceptual judgments of radiologists**. *Cognitive research: principles and implications* 6 (2021), 1–13 (see page 2).

[MC04] Neil A Macmillan and C Douglas Creelman. **Detection theory: A user's guide**. Psychology press, 2004 (see page 11).

[Mol+08] Eduard Molins, Francesc Macià, Francesc Ferrer, Maria-Teresa Maristany, and Xavier Castells. **Association between radiologists' experience and accuracy in interpreting screening mammograms**. *BMC health services research* 8:1 (2008), 1–10 (see page 2).

[NGL16] Andrey R Nikolaev, Sergei Gepshtein, and Cees van Leeuwen. **Intermittent regime of brain activity at the early, bias-guided stage of perceptual learning**. *Journal of Vision* 16:14 (2016), 11–11 (see pages 3, 29).

[OPG18] Jacob L Orquin, Sonja Perkovic, and Klaus G Grunert. **Visual biases in decision making**. *Applied Economic Perspectives and Policy* 40:4 (2018), 523–537 (see page 2).

[RCN12] Richard Russell, Garga Chatterjee, and Ken Nakayama. **Developmental prosopagnosia and super-recognition: No special role for surface reflectance processing**. *Neuropsychologia* 50:2 (2012), 334–340 (see page 2).

[RDN09] Richard Russell, Brad Duchaine, and Ken Nakayama. **Super-recognizers: People with extraordinary face recognition ability**. *Psychonomic bulletin & review* 16:2 (2009), 252–257 (see page 2).

[Ren+23] Zhihang Ren, Xinyu Li, Dana Pietralla, Mauro Manassi, and David Whitney. **Serial Dependence in Dermatological Judgments**. *Diagnostics* 13:10 (2023), 1775 (see page 2).

[Ric+19] Jennifer J Richler, Andrew J Tomarken, Mackenzie A Sunday, Timothy J Vickery, Kaitlin F Ryan, R Jackie Floyd, David Sheinberg, Alan C-N Wong, and Isabel Gauthier. **Individual differences in object recognition.** *Psychological Review* 126:2 (2019), 226 (see page 2).

[Ros+16] Tony Rosen, Elizabeth M Bloemen, Jasmin Harpe, Allen M Sanchez, Kevin W Mennitt, Thomas J McCarthy, Refky Nicola, Kieran Murphy, Veronica M LoFaso, Neal Flomenbaum, et al. **Radiologists' training, experience, and attitudes about elder abuse detection**. *AJR. American journal of roentgenology* 207:6 (2016), 1210 (see page 2).

[Rot+21] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. **A patient-centric dataset of images and metadata for identifying melanomas using clinical context**. *Scientific data* 8:1 (2021), 34 (see page 7).

[Sal+08] L. J. Salomon, N. Winer, J. P. Bernard, and Y. Ville. **A score-based method for quality control of fetal images at routine second-trimester ultrasound examination**. *Prenatal Diagnosis* 28:9 (Sept. 2008), 822–827. ISSN: 01973851. DOI: 10.1002/PD.2016 (see page 2).

[Sch14] Alexander C Schütz. **Interindividual differences in preferred directions of perceptual and motor decisions**. *Journal of vision* 14:12 (2014), 16–16 (see page 2).

[SD87] N Clayton Silver and William P Dunlap. **Averaging correlation coefficients: Should Fisher's z transformation be used?** *Journal of applied psychology* 72:1 (1987), 146 (see page 14).

[SDG17] Mackenzie A Sunday, Edwin Donnelly, and Isabel Gauthier. **Individual differences in perceptual abilities in medical imaging: the Vanderbilt Chest Radiograph Test**. *Cognitive Research: Principles and Implications* 2:1 (2017), 1–10 (see page 2).

[SDG18] Mackenzie A Sunday, Edwin Donnelly, and Isabel Gauthier. **Both fluid intelligence and visual object recognition ability relate to nodule detection in chest radiographs**. *Applied Cognitive Psychology* 32:6 (2018), 755–762 (see page 2).

[SG17] Erich Schubert and Michael Gertz. **Intrinsic t-Stochastic Neighbor Embedding for Visualization and Outlier Detection: A Remedy Against the Curse of Dimensionality?** In: *Similarity Search and Applications: 10th International Conference, SISAP 2017, Munich, Germany, October 4-6, 2017, Proceedings 10*. Springer. 2017, 188–203 (see page 9).

[SK18] Ehsan Samei and Elizabeth A Krupinski. **The handbook of medical image perception and techniques**. Cambridge University Press, 2018 (see page 2).

[Smo+84] WR Smoker, KS Berbaum, NH Luebke, and CG Jacoby. **Spatial perception testing in diagnostic radiology**. *American journal of roentgenology* 143:5 (1984), 1105–1109 (see page 2).

[Son+15] Yang Song, Weidong Cai, Heng Huang, Yun Zhou, David Dagan Feng, Yue Wang, Michael J Fulham, and Mei Chen. **Large margin local estimate with applications to medical image classification**. *IEEE transactions on medical imaging* 34:6 (2015), 1362–1377 (see page 4).

[Str03] David L Streiner. **Starting at the beginning: an introduction to coefficient alpha and internal consistency**. *Journal of personality assessment* 80:1 (2003), 99–103 (see page 13).

[SZ14] Karen Simonyan and Andrew Zisserman. **Very deep convolutional networks for large-scale image recognition**. *arXiv preprint arXiv:1409.1556* (2014) (see page 8).

[Tan+06] Alai Tan, Daniel H Freeman, James S Goodwin, and Jean L Freeman. **Variation in false-positive rates of mammography reading among 1067 radiologists: a population-based assessment**. *Breast cancer research and treatment* 100 (2006), 309–318 (see page 2).

[Tip85] Steven P Tipper. **The negative priming effect: Inhibitory priming by ignored objects**. *The quarterly journal of experimental psychology* 37:4 (1985), 571–590 (see page 2).

[TL19] Mingxing Tan and Quoc Le. **Efficientnet: Rethinking model scaling for convolutional neural networks**. In: *International conference on machine learning*. PMLR. 2019, 6105–6114 (see page 8).

[TRK18] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. **The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions**. *Scientific data* 5:1 (2018), 1–9 (see page 5).

[VO85] Rufin Vogels and Guy A Orban. **The effect of practice on the oblique effect in line orientation judgments**. *Vision research* 25:11 (1985), 1679–1687 (see pages 3, 29).

[VPV+09] Laurens Van Der Maaten, Eric Postma, Jaap Van den Herik, et al. **Dimensionality reduction: a comparative**. *J Mach Learn Res* 10:66-71 (2009), 13 (see page 9).

[Wan+12a] Ruosi Wang, Jingguang Li, Huizhen Fang, Moqian Tian, and Jia Liu. **Individual differences in holistic processing predict face recognition ability**. *Psychological science* 23:2 (2012), 169–177 (see page 2).

[Wan+12b] Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. **End-to-end text recognition with convolutional neural networks**. In: *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*. IEEE. 2012, 3304–3308 (see page 8).

[Wan+14] Xiang Wan, Wenqian Wang, Jiming Liu, and Tiejun Tong. **Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range**. *BMC medical research methodology* 14 (2014), 1–13 (see page 13).

[Wan+18] Ying Wang, Li Wang, Qian Xu, Dong Liu, Lihong Chen, Nikolaus F Troje, Sheng He, and Yi Jiang. **Heritable aspects of biological motion perception and its covariation with autistic traits**. *Proceedings of the National Academy of Sciences* 115:8 (2018), 1937–1942 (see page 2).

[Wan+22] Zixuan Wang, Mauro Manassi, Zhihang Ren, Cristina Ghirardo, Teresa Canas-Bajo, Yuki Murai, Min Zhou, and David Whitney. **Idiosyncratic biases in the perception of medical images**. *Frontiers in Psychology* (2022) (see pages 2, 3, 22).

[WDM15] Mark Wexler, Marianne Duyck, and Pascal Mamassian. **Persistent states in vision break universality and time invariance**. *Proceedings of the National Academy of Sciences* 112:48 (2015), 14990–14995 (see page 2).

[Wil+10] Jeremy B Wilmer, Laura Germine, Christopher F Chabris, Garga Chatterjee, Mark Williams, Eric Loken, Ken Nakayama, and Bradley Duchaine. **Human face recognition ability is specific and highly heritable**. *Proceedings of the National Academy of sciences* 107:11 (2010), 5238–5241 (see page 2).

[Wil17] Jeremy B Wilmer. **Individual differences in face recognition: A decade of discovery**. *Current Directions in Psychological Science* 26:3 (2017), 225–230 (see page 2).

[WMW20] Zixuan Wang, Yuki Murai, and David Whitney. **Idiosyncratic perception: a link between acuity, perceived position and apparent size**. *Proceedings of the Royal Society B* 287:1930 (2020), 20200825 (see page 2).

[ZF14] Matthew D Zeiler and Rob Fergus. **Visualizing and understanding convolutional networks**. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, 818–833 (see page 8).

[Zhu+10] Qi Zhu, Yiying Song, Siyuan Hu, Xiaobai Li, Moqian Tian, Zonglei Zhen, Qi Dong, Nancy Kanwisher, and Jia Liu. **Heritability of the specific cognitive ability of face perception**. *Current Biology* 20:2 (2010), 137–142 (see page 2).

[Zhu+21] Zijian Zhu, Biqing Chen, Ren Na, Wan Fang, Wenxia Zhang, Qin Zhou, Shanbi Zhou, Han Lei, Ailong Huang, Tingmei Chen, et al. **A genome-wide association study reveals a substantial genetic basis underlying the Ebbinghaus illusion**. *Journal of Human Genetics* 66:3 (2021), 261–271 (see page 2).