Getting the gist of it: An investigation of gist processing and the learning of novel gist categories

Emma Marie Raat

PhD

University of York

Psychology

April 2023

Abstract

Gist extraction rapidly processes global structural regularities to provide access to the general meaning and global categorizations of our visual environment – the gist. Medical experts can also extract gist information from mammograms to categorize them as normal or abnormal. However, the visual properties influencing the gist of medical abnormality are largely unknown. It is also not known how medical experts, or any observer for that matter, learned to recognise the gist of new categories. This thesis investigated the processing and acquisition of the gist of abnormality. Chapter 2 observed no significant differences in performance between 500 ms and unlimited viewing time, suggesting that the gist of abnormality is fully accessible after 500 ms and remains available during further visual processing. Next, chapter 3 demonstrated that certain high-pass filters enhanced gist signals in mammograms at risk of future cancer, without affecting overall performance. These filters could be used to enhance mammograms for gist risk-factor scoring. Chapter 4's multi-session training showed that perceptual exposure with global feedback is sufficient to induce learning of a new gist categorisation. However, learning was affected by individual differences and was not significantly retained after 7-10 days, suggesting that prolonged perceptual exposure might be needed for consolidation. Chapter 5 observed evidence for the neural signature of gist extraction in medical experts across a network of regions, where neural activity patterns showed clear individual differences. Overall, the findings of this thesis confirm the gist extraction of medical abnormality as a rapid, global process that is sensitive to spatial structural regularities. Additionally, it was shown that a gist category can be learned via global feedback, but this learning is hard to retain and is affected by individual differences. Similarly, individual differences were observed in the neural signature of gist extraction by medical experts.

Table of Contents

Abstract 2
Table of Contents 2
List of Tables 5
List of Figures 7
Acknowledgements 11
Author's Declaration 12
Chapter 1: Literature Review 15
  1.1 Introduction 15
  1.2. Visual processes making up gist extraction 17

1.2.1. Spatial structural regularities 17
1.2.2. Summary statistics 22
1.2.3. Intermediate disjunctive features 24
1.2.4. Non-selective process 25
1.3. Neural mechanisms of gist extraction 29
1.3.1. Category processing 30
1.3.2. Spatial layout processing 32
1.3.3. Summary of neural mechanisms 35
1.4. Gist extraction in Medical Image Perception 35
1.5. Goals of this thesis 38
1.5.1. Perceptual properties 38
1.5.2. Learning processes 39
1.5.3. Neural signature 40
Chapter 2: Comparable prediction of breast cancer risk from a glimpse or a first impression of a mammogram 41
2.1. Declarations 41
2.2. Abstract 43
2.3. Significance statement 44
2.4. Introduction 45
2.5. Methods 47
2.6. Results 54
2.7. Discussion 60
2.8. List of Abbreviations 63
2.9. Open Practices Statement 63
2.10. Appendices 64
Appendix A: Detailed effect of time limit on performance in naïves 64
2.11 Commentary: AUC vs d’ 64
Chapter 3: Early signs of cancer present in the fine detail of mammograms. 65
3.1. Abstract 65
3.2. List of Abbreviations 66
3.3. Introduction 66
3.4. Methods 68
3.5. Results 73
3.6. Discussion 78
3.7. Conclusion 80
3.9. Declarations 81
3.10. Appendices 81

Appendix A: Factors influencing AUC performance measure ... 81

3.11: Commentary: Image processing ... 83

Chapter 4: Future studies using bandstop filters to selectively remove specific frequency bands likely will not need to use brightness boosting, as they would not remove 0 Hz. These bandstop filters could be helpful in further isolating the role of specific frequencies in carrying or masking the gist of medical abnormality. Using global feedback to induce learning of gist of abnormality in mammograms ... 84

4.1. Declarations ... 84

4.2. Abstract ... 86

4.3. Significance statement ... 87

4.4. Introduction ... 87

4.5. Methods ... 94

4.6. Results and Discussion ... 100

4.7. Conclusion ... 111

4.7. List of Abbreviations ... 113

4.8. Open Practices Statement ... 113

4.9. Appendices ... 113

Appendix A: Mammographic descriptors of obvious and subtle cases ... 113

Appendix B: Engagement and attention in training phases ... 114

Appendix C: Log-linear likelihood ratios ROC curves ... 115

Appendix D: Effect of training on rating time ... 116

Appendix E: Criterion for learners and non-learners ... 118

Appendix F: DNN probability of benign abnormality ... 118

Appendix G: DNN correlation with non-learners ... 120

Chapter 5: The neural signature of the gist of medical abnormality ... 121

5.1. Abstract ... 121

5.2. Introduction ... 121

5.3. Materials & Methods ... 123

5.4. Results ... 134

5.5. Discussion ... 151

5.6. Conclusion ... 156

5.7. Appendices ... 157

Appendix A ... 157

Appendix B ... 157

Appendix C ... 158

Chapter 6: Discussion ... 159

6.1. Effects of viewing time... 159
6.2. Spatial frequency bands.. 160
6.3. Statistical learning through perceptual exposure.. 163
6.4. Neural signature of gist extraction in experts 167
6.5. Implications for medical imaging... 168
6.6. Conclusion.. 171
Thesis Abbreviations 172

References 173

List of Tables

Table 2.1: Average reaction time in milliseconds for naïves (n=23) and radiologists (n=11) during no time limit conditions, per image type (± 95% CI)... 60
Table 3.1: Group average and mixed model estimated mean of D', criterion, and AUC for unfiltered mammograms and each high-pass filter frequency. 73
Table 4.1: AUCs for malignancy probability scores for the SBC and SBC+HM for obvious, subtle, and global mammograms versus the group of normal mammograms. This is calculated for the training set and the test set separately. Square brackets contain the lower and upper bands of 95% CIs. 110
Table 4.2: Spearman's rank correlations between the average human learner score pre- and post-training of human observers, and the SBC/SBC+HM malignancy probability scores. 111
Table 4.3: Percentage of mammograms with a high, intermediate, low, or unassessed DCIS grade for the obvious and subtle subsets of the image set. Where descriptors were not available in the OPTIMAM database, the mammogram was classified as unassessed. 114
Table 4.4: Percentage of mammograms that were positive, negative, borderline, or unassessed for different receptor groups, HER-2, Progesterone, and Estrogen receptors for the obvious and subtle subsets of the image set. Where descriptors were not available in the OPTIMAM database, the mammogram was classified as unassessed.. 114
Table 4.5: Number of participants performing at above chance levels (real AUC > 95th% simulated AUC) at each testing phase, split up for learners, non-learners, and total. 115
Table 4.6: AUCs for the probability of benign abnormality for the SBC and SBC+HM for obvious, subtle, and global mammograms versus the group of normal mammograms. This is calculated for the training set and the test set.. 119
Table 4.7: Spearman's rank correlations between the average human learner score pre- and post-training, and the SBC/SBC+HM probabilities of benign abnormality. 119

Table 4.8: Spearman's rank correlations between the average human non-learner score pre- and post-training, and the SBC/SBC+HM probabilities of malignant or benign abnormality. 120

Table 5.1: List of the ERPs that were tested for each of the clusters, with the start and end (ms) of the search window for the peak activity, the direction of the peak, and the target window length (ms) around the peak. The table combines some clusters where the same ERPs were investigated, to reduce the length of the table. 131

Table 5.2: Performance measures for each radiologist, showing their overall criterion, overall d' performance and d' separated for the obvious, subtle, and global mammograms. The years of experience and number of mammogram cases viewed in the previous year are also shown. 134

Table 5.3: Results of single subject bootstrapping for the Occipital cluster, showing the p-value and the average difference between abnormal and normal trials in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined. 138

Table 5.4: Results of single subject bootstrapping for the Occipital cluster, showing the p-value and the average difference between hits and true negatives in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined. 138

Table 5.5: Results of single subject bootstrapping for the Parietal cluster, showing the p-value and the average difference between abnormal and normal trials in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined. 139

Table 5.6: Results of single subject bootstrapping for the Parietal cluster, showing the p-value and the average difference between hits and true negatives in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined. 139

Table 5.7: Results of single subject bootstrapping for the Temporal cluster, showing the p-value and the average difference between abnormal and normal trials in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined. 140

Table 5.8: Results of single subject bootstrapping for the Temporal cluster, showing the p-value and the average difference between hits and true negatives in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined. 140

Table 5.9: Results of single subject bootstrapping for the Central cluster, showing the p-value and the average difference between abnormal and normal trials in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined. 144

Table 5.10: Results of single subject bootstrapping for the Central cluster, showing the p-value and the average difference between hits and true negatives in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined. 144

Table 5.11: Results of single subject bootstrapping for the Frontal cluster, showing the p-value and the average difference between abnormal and normal trials in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined. 147

Table 5.12: Results of single subject bootstrapping for the Frontal cluster, showing the p-value and the average difference between hits and true negatives in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined. 148
Table 5.13: Total sum and percentage (in brackets) of significant/trending effects per ERP per radiologist for hits vs true negatives. Percentage is calculated based on the number of clusters it was tested at (from Occipital, Parietal, Temporal, Central, and Frontal). 149
Table 5.14: Total sum and percentage (in brackets) of significant/trending effects per cluster per radiologist for hits vs true negatives. Percentage is calculated based on the number of ERPs that were tested for that cluster. 149
Table 5.15: Kendall rank correlation for correlations between the absolute difference between hits and true negatives and d’, cases viewed, and years of experience. The table shows the Kendall's τ coefficient and p-value for each correlation, with trending values in bold font. 151
Table 5.16: Total sum and percentage (in brackets) of significant/trending effects per ERP per radiologist for abnormal vs normal. Percentage is calculated based on the number of clusters it was tested at (from Occipital, Parietal, Temporal, Central, and Frontal). 157
Table 5.17: Total sum and percentage (in brackets) of significant/trending effects per cluster per radiologist for abnormal vs normal. Percentage is calculated based on the number of ERPs that were tested for that cluster. 157
Table 5.18: Kendall rank correlation for correlations between the absolute difference between abnormal and normal trials and d’, cases viewed, and years of experience. The table shows the Kendall's τ coefficient and p-value for each correlation. 158

List of Figures

Figure 2.1: Simplified overview of the experimental procedure (A) and example mammograms for each of the four types used in this experiment (B). 52
Figure 2.2: Average ratings for each observer group for each type of image. Statistical results are Dunnett’s multiple comparisons tests, comparing each type of abnormal image to the normal images. 54
Figure 2.3: ROC curves for the radiologist groups during no time limit and 500 ms time limit conditions per image type (subtle abnormal, contralateral, priors). Each plot contains individual ROCs (coloured dotted lines) and the group mean ROC (thick black line). The dashed grey diagonal line indicates the line of no discrimination. 56
Figure 2.4: Bar graphs representing the average d’, AUC, and criterion (±SEM) per image comparison category (subtle abnormal, contralateral, prior) and over the total image set for the radiologists and naïves under no time limit and 500 ms time limit conditions. 58

Figure 3.1: Procedure and stimuli used in the experiment. (A) Example visualization of the different screens in one trial, showing the fixation cross, mammogram case, mask, and rating screen. (B) Examples for the unaltered (0) and high-pass filtered versions (0.5, 1, 1.5, and 2 cpd) of a unilateral mammogram 70

Figure 3.2: ROC curves per image category. Average and individual (dotted) ROCs per frequency condition (0, 0.5, 1, 1.5, and 2 cpd) for each abnormal mammogram category (obvious, subtle, and prior). The black dashed line represents chance levels, with anything above it being above chance........ 74

Figure 3.3: Boxplots of D’ across conditions. Each boxplot shows the median as the line within the coloured box containing the 25th and 75th percentiles, with extending whiskers containing the 95% CI, with any outliers plotted as dots. Significance of pairwise comparisons is indicated in the figure with *=p<.05, **=p<.01,***=p<.001. (A) Boxplots showing D’ across frequency conditions (0, 0.5, 1, 1.5, and 2 cpd). Pairwise comparisons of frequency showed that D’ was significantly higher for F0 than F1 (estimated difference=0.482, t(271)=6.080, p=<.0001) and F2 (estimated difference=0.397, t(282)=4.502, p=<.0001), but did not differ significantly from F0.5 (estimated difference=-0.181, t(273)=-2.637, p=.067), and F1.5 (estimated difference=0.007, t(282)=0.083, p=1.000) –and even trended towards a higher D’ in F0.5. (B) Boxplots showing D’ for each mammogram category (obvious, subtle, and prior) and frequency, to illustrate mammogram category and interaction effects. Pairwise comparisons of mammogram category showed that D’ was significantly higher for obvious than subtle (estimated difference=0.462, t(258)=7.278, p=<.0001) and prior mammograms (estimated difference=0.683, t(258)=10.763, p=<.0001), and higher for subtle than prior mammograms (estimated difference=0.221, t(258)=3.485, p=<.001). 75

Figure 3.4: Boxplots of criterion across conditions. Each boxplot shows the median as the line within the coloured box containing the 25th and 75th percentiles, with extending whiskers containing the 95% CI, with any outliers plotted as dots. Significance of pairwise comparisons of main effects is indicated in the figure with *=p<.05, **=p<.01,***=p<.001. (A) Criterion across frequency conditions (0, 0.5, 1, 1.5, and 2 cpd). Pairwise comparisons of frequency showed that criterion was significantly higher for F0 than F1 (estimated difference=-0.3295, t(261)=-5.640, p=<.0001), F1.5 (estimated difference=-0.695, t(264)=-9.176, p<.0001), and F2 (estimated difference=-0.480, t(264)=-6.329, p=<.0001), but did not differ significantly from F0.5 (estimated difference=-0.086, t(261)=-1.477, p=.579). (B) Criterion for each mammogram category (obvious, subtle, and prior) and frequency, to illustrate mammogram category and interaction effects. Pairwise comparisons of mammogram category showed that criterion was significantly lower (less conservative) for obvious than subtle (estimated difference=-0.232, t(258)=-4.345, p=<.0001) and prior mammograms (estimated difference=-.347, t(258)=-6.517, p=<.0001), but did not differ significantly between subtle and prior mammograms (estimated difference=-.116, t(258)=-2.172, p=.078). 77

Figure 3.5. Boxplots of AUC across conditions. Each boxplot shows the median as the line within the coloured box containing the 25th and 75th percentiles, with extending whiskers containing the 95% CI, with any outliers plotted as dots. Significance of pairwise comparisons is indicated in the figure with *=p<.05, **=p<.01,***=p<.001. (A) Boxplots showing AUC across frequency conditions (0, 0.5, 1, 1.5, and 2 cpd). Pairwise comparisons of frequency showed that AUC was significantly higher for F0 than F1 (estimated difference=0.089, t(273)=6.533, p=<.0001) and F2 (estimated difference=0.098, t(284)=5.647, p=<.0001), but did not differ significantly from F0.5 (estimated difference=-0.025, t(273)=-1.827, p=.360) and F1.5 (estimated difference=0.023, t(284)=1.306, p=.688). (B) Boxplots showing AUC for each mammogram category (obvious, subtle, and prior) and frequency, to illustrate mammogram category and interaction effects. Pairwise comparisons of mammogram category showed that D' was significantly higher obvious than subtle (estimated difference=0.089, t(258)=7.058, p=<.0001) and prior mammograms (estimated difference=0.156, t(258)=12.368, p=<.0001), and higher for subtle than prior mammograms (estimated difference=0.068, t(258)=5.310, p=<.001). 82

Figure 4.1: Scene exemplars for beaches (A) and playgrounds (B) that illustrate the variation in viewing angle, lightning, configuration, and specific objects. Mammogram exemplars containing subtle abnormalities (C) or no abnormalities (D) illustrating the variation in shape, size, and textural patterns. 90

Figure 4.2: Overview of the experimental procedure and flow-chart schedule of the experiment. The screens show the presentation order within a training trial and the duration or button press to continue. Test trials always showed mammograms for 500 ms and omitted the feedback screen but were otherwise identical. The flow-chart schedule shows the order of experimental phases for each session, and the number of unilateral mammograms viewed per session. In the test phases, 200 mammograms were viewed, while the training phases had 4 blocks with 180 mammograms each. Session 1 to 9 were separated by 1 to 3 days each, while session 10 was delayed by 7 to 10 days after session 9. 97

Figure 4.3: Mean d', criterion, and AUC across test phases (± 95% confidence intervals) for all participants (n=15), plotted separately for each abnormal image type (○ Obvious, ● Subtle, □ Global). 101

Figure 4.4: Individual progression of d' across the 9 training phases, with the learners in green hues in the left plot and the non-learners in orange hues in the right plot. 103

Figure 4.5: Mean d', criterion, and AUC across test phases (± 95% confidence intervals) for the learners (n=9), plotted separately for each abnormal image type (○ Obvious, ● Subtle, □ Global). 104

Figure 4.6: Distribution of Single Breast Classifier (SBC) and SBC+Heatmap (SBC+HM) malignancy probability scores on the full image set of mammograms split into 25 bins for each of the image type categories, with a combined plot showing the overlap between normal (red), obvious (green), subtle (blue), and global (yellow) scores. 109

Figure 4.7: Maximum and average viewing time in milliseconds per participant at the end of each training phase. Maximum viewing time is calculated for the fourth block of the session. Individual lines are plotted, while the dashed black line represents the group average. 115

Figure 4.8: Individual average rating times are shown at pre-training (pre, green), post-training (post, orange), and retention (ret, purple) testing phases, as connected dot-clouds per participant-image type combination and boxplots to show both individual patterns and the population distributions. 117

Figure 4.9: Individual median rating times are shown at pre-training (pre, green), post-training (post, orange), and retention (ret, purple) testing phases, as connected dot-clouds per participant-image type combination and boxplots to show both individual patterns and the population distributions. 117

Figure 4.10: Distribution of Single Breast Classifier (SBC) and SBC+Heatmap (SBC+HM) benign abnormality probability scores on the full image set of mammograms split into 25 bins for each of the image type categories, with a combined plot showing the overlap between normal (red), obvious (green), subtle (blue), and global (yellow) scores. 118

Figure 5.1: The timeline of a trial, showing the fixation cross (1000 – 2000 ms total), auditory cue (100 ms), mammogram (500 ms), mask (500 ms), and 2-AFC answering window (2000 ms). 126

Figure 5.2: Timeline of topography maps of averaged activity across the five radiologists. 127

Figure 5.3: Channel layout and chosen clusters of interest: Frontal (blue), central (yellow), temporal (green), parietal (red), and occipital (pink). 128

Figure 5.4: Example histograms showing the distributions of 10.000 bootstrapped differences (µV) in Temporal P600 ERPs when viewing abnormal versus normal mammograms for each radiologist. The histograms show that the temporal P600 evoked by Abnormal mammograms was larger in amplitude for radiologists 1, 4, and 5, and smaller for radiologists 2 and 3. Each plot also shows the real observed difference (black line), zero difference line (red dashed line), and the proportion of samples below/above 0 (top left corner). 133

Figure 5.5: Single subject bootstrapping results showing the Temporal P600 target time windows for abnormal (green dashes) and normal (red dashes), averaged activity, and p-values per radiologist. 134

Figure 5.6: Time trace of the average neural activity in Frontal, Central, Temporal, Parietal, and Occipital clusters for Abnormal (solid) and Normal (dashed) mammograms for individual radiologists, and group averages (Abnormal, black line; Normal, grey line). 136

Figure 5.7: Time trace of the average activity in Frontal, Central, Temporal, Parietal, and Occipital clusters for Hits (solid) and True Negatives (dashed) mammograms for individual radiologists, and group averages (Hits, black line; True Negatives, grey line). 137
Figure 5.8: Single subject bootstrapping results showing the Temporal N1 target time windows for abnormal (green dashes) and normal (red dashes), averaged activity, and p-values per radiologist. 141
Figure 5.9: Single subject bootstrapping results showing the Temporal P600 target time windows for abnormal (green dashes) and normal (red dashes), averaged activity, and p-values per radiologist. 142
Figure 5.10: Single subject bootstrapping results showing the Temporal N1 target time windows for hits (green dashes) and true negatives (red dashes), averaged activity (hits: green, true negatives: red), and p-values per radiologist.. 143
Figure 5.11: Single subject bootstrapping results showing the Central P2 target time windows for abnormal (green dashes) and normal (red dashes), averaged activity, and p-values per radiologist. 145
Figure 5.12: Single subject bootstrapping results showing the Central P600 target time windows for abnormal (green dashes) and normal (red dashes), averaged activity, and p-values per radiologist. 146
Figure 5.13: Single subject bootstrapping results showing the Central P600 target time windows for hits (green dashes) and true negatives (red dashes), averaged activity (hits: green, true negatives: red), and p-values per radiologist.. 147

Acknowledgements

Firstly, my gratitude goes to my supervisor Dr Karla Evans. Thank you for giving me the freedom to work independently, but also for offering guidance when I needed it. Your support allowed me to develop many skills that I will use throughout my entire career, wherever it may take me.

I am also thankful to my TAP members, Mike Burton, Alex Wade, and Daniel Kaiser, for their time and thoughts, and for helping me fit my thesis within my timeline. I want to express special gratitude to Daniel Kaiser for his detailed feedback on my literature review, and to Alex Wade for introducing me to the EEG lab. Additionally, I'd like to thank both of my progression chairs, Steve Tipper, and later Sven Mattys, for their time.

Importantly, I want to thank all radiologists that took part in my studies. This work would not have been possible without their willingness to try something weird and look at mammograms for mere milliseconds. A special thanks goes to Dr Roisin Bradley for sharing her medical expertise throughout my PhD.

I also would like to thank Isabel Farr and Ryan Tan for their contributions to data collection during their volunteer research assistant positions.

A big thank you to my lovely lab members Emily, Lyndon, and Cameron. Thank you to Cameron for his help with computational modelling. Special shout-out to Lyndon for sticking with me on the EEG data collection and analysis journey. You guys made cycling all the way to the department worth it!

I am eternally grateful to my parents for always stimulating my curiosity and love for science. You always supported me to do anything I set my mind to – and I couldn't have done this without the unconditional support you always give me.

I would also like to thank my amazing friends from afar, Daphne and Justine, for always being there with advice and a listening ear. And to my sister Andrea, who knew all too well what I was going through, thank you for all the support.

Lastly, I could not have undertaken this journey without my partner Joe. You fully believed in me from the day I decided to apply for a studentship in York. It has been a whirlwind to complete this PhD, across Covid, buying a house, adopting a cat, and starting my career as a developer. Thank you for providing a cosy home base to read, analyse, and write, with cups of tea, hugs, and (so many) cat pictures. The past years have been stressful and difficult but knowing that you and Poppy were there to listen, support me, and celebrate each success meant the world.

Author's Declaration

I declare that this thesis is a presentation of original work that is my own unless specified otherwise below. None of this work has been submitted for examination at this or any other institution for another award or degree. The work was carried out under the supervision of Dr Karla K. Evans. This research was supported by a departmental studentship from the Department of Psychology, University of York, UK. All sources are acknowledged as references.

Chapter 2

Chapter 2 has been published as a journal article in Cognitive research: Principles and Implication. Raat, E. M., Farr, I., Wolfe, J. M., & Evans, K. K. (2021). Comparable prediction of breast cancer risk from a glimpse or a first impression of a mammogram. Cognitive Research: Principles and Implications, 6(1), 1-14. doi.org/10.1186/s41235-021-00339-5

Emma Raat made substantial contributions to the conception and design of the study, to the acquisition of the radiologist no time limit data and the naïve participant data, to the analysis and interpretation of data, and has drafted the work. Research assistant Isabel Farr contributed to the acquisition of the naïve participant data. Dr Jeremy M. Wolfe made substantial contributions to the design, analysis, and interpretation of the data and to the editing of the manuscript. Dr Karla K. Evans made substantial contributions to the conception and design of the study, to the acquisition of the radiologist time limit data, to the analysis and interpretation of the data, and revising the work. All authors read and approved the final manuscript.

Chapter 3

Chapter 3 has been published as a journal article in PLoS One. Raat E. M., Evans K. K. (2023) Early signs of cancer present in the fine detail of mammograms. PLoS ONE 18(4): e0282872. doi.org/10.1371/journal.pone.0282872

Emma Raat contributed significantly to conceptualization, methodology, software, investigation, formal analysis, visualization, project administration, and writing and editing of the manuscript. Dr Karla K. Evans made substantial contributions to the conception and design of the study, to the acquisition of the in-person radiologist data, to the interpretation of the data, and revising of the work. All authors read and approved the final manuscript.

Preliminary versions of this work were also presented at the following conferences:

- Raat, E.M., Farr, I., Evans, K.K. (May 2021) *The effect of spatial frequency on perceiving the gist of abnormality in mammograms.* Poster presentation at V-VSS 2021
- Raat, E.M., Evans, K.K (March 2022) The role of spatial frequencies in mammograms (2022), Talk at UoY and York NHS Networking meeting.
- Raat, E.M., Evans, K.K (July 2022). *The effect of spatial frequency on gist perception in medical imaging.* Medical Image Perception Society Conference XIX, York

Chapter 4

Chapter 4 was published as Raat, E. M., Kyle-Davidson, C., & Evans, K. K. (2023). *Using global feedback to induce learning of gist of abnormality in mammograms* in Cognitive Research: Principles and Implications, 8(1), 1-22. https://doi.org/10.1186/s41235-022-00457-8

Emma Raat made substantial contributions to the conception and design of the study, to the acquisition of the data, to the analysis and interpretation of data, and drafted the work, and took the lead in editing and revising of the work. Dr Cameron Kyle-Davidson made substantial contributions to the analysis and interpretation of the neural network data, and to the revising of the work. Dr Karla K. Evans made substantial contributions to the conception and design of the study, to the analysis and interpretation of the data, and to the editing and revising the work. All authors read and approved the final manuscript.

Additionally, research assistant Isabel Farr assisted with some administrative duties for the online data collection.

Preliminary versions of this work were also presented at the following conferences:

- Raat, E.M., Farr, I., Evans, K.K (August 2021) *Learning and retention of the gist of abnormality in mammograms after perceptual training of naïves.* Talk at Visual Cognition II session of the online ECVP conference.
- Raat, E.M., Farr, I., Evans, K.K (November 2021) *How to Teach Naïve Observers to Recognize the Gist of Cancer in Mammograms.* Poster presentation at the online Psychonomic conference.
- Raat, E.M., Farr, I., Kyle-Davidson, C., Evans, K.K. (May 2022). *Learning to perceive the gist of cancer through perceptual training.* Poster presentation at VSS conference, Florida, USA.
- Raat, E.M., Farr, I., Evans, K.K (June 2022) *Effects of perceptual training on the learning and retention of the gist of medical abnormality.* Research Seminar Talk at University of York Psychology Department.

Chapter 5

Emma Raat was responsible for the conception and design of the study, supported the data collection, performed the pre-processing and statistical analysis of the data, and wrote the chapter. Lab technician Lyndon Rakusen assisted in collecting the radiologist' EEG recordings and assisted with the code for the pre-processing of the EEG data. Research assistant Ryan Tan assisted in some of the data collection for this chapter.

Chapter 1: Literature Review

1.1 Introduction

In the medical field, categorization of complex images into normal and abnormal is an important first step for further diagnostics. Medical personnel performing this task are often described as experts, suggesting that they innately possess or have developed the specific capabilities necessary to diagnose these complex medical images. Diagnosing medical images is a difficult task, where an abnormality must be perceived and processed to lead to a decision. Indeed, medical experts such as radiologists or cytologists possess the extraordinary ability to view a medical image and form a diagnosis. While part of reaching a diagnosis comes from interpreting clinical history and symptoms, the main bulk must come from the perceptual information gathered from the medical images, suggesting a large role for perceptual expertise. Interestingly, medical experts frequently describe experiences where they know that something is wrong with a medical image, before they have performed a detailed search and decision-making process, seemingly based on global information they perceived in the blink of an eye. This suggests that a broad distinction between normal and abnormal cases can be achieved even after only a short exposure. Thus, these medical images might contain a so-called gist signal, a general meaning or categorization of our visual environment, which medical experts are able to extract through their expertise.

Indeed, previous research has shown that medical experts can distinguish abnormal from normal cases after rapid exposure to a medical image. For example, radiologists achieved 70% accuracy in categorising chest radiographs with just 200 ms of visual exposure (Kundel & Nodine, 1975), which is not enough for visual search to take place. This ability is not limited to chest radiographs, as medical experts were able to accurately rate the probability of an abnormality in mammograms and micrographs (cytology) with above-chance accuracy with 250 ms exposure time (Evans, Georgian-Smith, Tambouret, Birdwell, & Wolfe, 2013). Under these conditions, observers reached d' prime of approximately 1 with 250 ms exposure time, and ~1.14 with 500 ms for mammograms, compared with a d' prime of 1.86 for abnormal/normal categorisation during free viewing of a mammogram dataset under similar laboratory settings (Evans, Birdwell, & Wolfe, 2013). Thus, performance under gist conditions is lower than under normal diagnostic conditions with full scrutiny, but medical experts still show a surprisingly accurate perception of presence or absence of abnormalities when they view a medical image for only 100 to 500 milliseconds. In this thesis, I will refer to this ability as the extraction of the '*gist of abnormality*'.

While medical professionals are perceptual experts in their specific fields, any human can be said to possess a high level of expertise about their environment. In our daily lives, we are constantly exposed to complex visual environments, from which we need to rapidly extract relevant features to guide our actions. Indeed, within 20-30 ms, humans can distinguish superordinate categories, such as natural from man-made environments (Joubert, Rousselet, Fabre-Thorpe, & Fize, 2009), or even differentiate basic scene categories such as fields, forests, and rivers (Greene & Oliva, 2009). Similarly, broad categorisations such as the presence or absence of an animal in the scene are above-chance with 10 ms exposures, and reach a high accuracy with 40-60 ms (Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005). Thus, humans can extract surprisingly complex semantic and statistical information from rapidly presented scenes, analogous to the gist of abnormality extracted by medical experts.

Importantly, gist categorisation timeframes do not allow for detailed search of the image. Under normal circumstances, the average saccade latency in adults is 200-220 milliseconds (Darrien, Herd, Starling, Rosenberg, & Morrison, 2001; Gezeck, Fischer, & Timmer, 1997). This means that gist categorizations can be made without any eye movements, thus, the observer could not have foveated all the items in their environment. And while attention can be oriented without moving our eyes, either towards a saccade target preceding a saccade or while the eyes remain fixated, selective attention through covert attention still would not allow scrutinization of multiple image elements within the gist extraction timeline. For example, pre-saccadic covert attentional shifts occur approximately 50 ms beforehand and still rely on a saccade to occur (Deubel, 2008), while covert attention during fixation has estimated dwell times between 250 (Theeuwes, Godijn, & Pratt, 2004) and 500 ms (Duncan, Ward, & Shapiro, 1994) in simplistic visual arrays. Thus,, there must instead be some global, non-selective process in place which shapes our first, split-second impression without needing to direct selective attention overtly or covertly (Wolfe, Vo, Evans, & Greene, 2011). Indeed, this rapid recognition of categories of objects and scenes is thought to occur via a process called gist extraction, which allows us to rapidly extract general information about our visual environment without requiring focused attention or prolonged exposure.

In this thesis, gist extraction is defined as a collection of global, non-selective visual processes that extract structural and statistical regularities and spatial envelope properties over the whole image to extricate the general meaning, or gist, of any image. The gist of abnormality is defined as a specific type of gist extraction in which medical images are categorised into broad classes, usually normal and abnormal/suspicious. Combining the findings in medical image diagnosis and rapid scene categorisation raises the question of how medical experts developed their ability to extract the gist of abnormality, which can be generalized to the question of how human observers learn the gist of any new category. The main goals of this thesis are to further clarify what factors influence the extraction of the gist of medical abnormality, and to investigate the learning of the gist of a new category.

This literature review will firstly further substantiate the definition of gist extraction used here. Then, neuroimaging findings will be evaluated to form an understanding of potential mechanisms of gist extraction, through the timeline of processing and involved brain areas. Next, evidence for parameters influencing extraction of gist of abnormality by medical experts will be reviewed in more detail. Lastly, this chapter will outline the goals of this thesis.

1.2. Visual processes making up gist extraction

Gist extraction is likely not achieved through a singular process or mechanism. Instead, it is thought to consist of a collection of visual processes, occurring conjointly to extract gist, that share three key characteristics: These processes should all be **rapid**, occurring with short exposure times and reaction times; they should **not** require **focused** attention; and they should occur **globally** on the whole image, with consequently a lack of access to location information.

Gist extraction is a combination of rapid, non-selective, global processes that occur globally, across the visual field. Processes matching these three characteristics can be divided in three sub-groups with different properties. Firstly, the extraction of **spatial structural regularities,** which summarize the patterns of spatial frequencies present in an image; Secondly, the extraction of **summary statistics** of an ensemble of items, which are collapsed across an image; Thirdly, the extraction of **basic and intermediate features** across the whole visual field without focused attention**,** which are not bound to locations or other features. These three sub-groups of processes underlying gist will be discussed in more detail below, followed by a discussion of attentional theories that can explain how gist extraction occurs without focused attention.

1.2.1. Spatial structural regularities

Every image that we perceive is built up from textures in different orientations and spatial frequencies, overlapping and together forming the image. This also means that an image can be broken down into these spatial structural regularities. Detailed textural properties and their orientations can be extracted from a sample image using a set of linear bandpass filters with multiple orientations and scales according to a steerable pyramid structure, and applied to Gaussian noise to create an artificial image that carries the same spatial structural statistics or regularities as the sample image (Portilla & Simoncelli, 2000). Information extracted from scene photographs using spatial filters with scaling constants matching the receptive field sizes of the visual cortex area V2 based on foveal fixation at the centre of the scene can be appliedto Gaussian noise to create so-called 'metamer' images. These metamers which were not distinguished from their sample scene when they were presented sequentially followed by a target image (does this target match 1 or 2, 2-AFC) if this occurred prior to engagement of selective attention (200 ms exposure per image) (Freeman & Simoncelli, 2011), but upon further inspection were distinctly different and did not contain any recognizable objects. This shows that spatial structural regularities can capture the scene features that are extracted under rapid viewing conditions, resulting in similar perception to the original. However, observers were not asked to categorize the scenes or to detect which of the samples was artificial.

However, spatial structural regularities alone are not always sufficient for capturing the gist of an image. Materials such as paper, fabric, and glass can be accurately categorised after 40 ms exposure (Sharan, Rosenholtz, & Adelson, 2009), showing that these contain gist signals. However, accuracy on a match-to-sample task (250 ms exposure) was lower for artificially generated textures using Portilla & Simoncelli's model (with the same spatial structural properties as the texture) than real material patches (Balas & Conlin, 2015). Similarly, material categorization (water, metal, wood, stone) performance was significantly lower on presented artificial textures than natural images (Balas, Conlin, & Shipman, 2016).Even when the artificial texture was presented foveally and the natural image was presented peripherally, reducing the resolution of the natural image, natural images outperformed the artificial ones. Thus, the spatial structural regularities in our visual environment that can be extracted through series of multi-scale, multi-orientation filters analogous to the receptive fields in the V2 likely play a role in gist extraction but are unlikely to be sufficient to drive categorization on their own, at least as captured by the artificial texture models such as Portilla and Simoncelli (2000).

1.2.1.1. Low and high spatial frequency bands

Spatial frequency bands capture different aspects of visual information: lower spatial frequencies (LSF) provide coarse, 'blobby' information, and higher spatial frequencies (HSF) provide edges and contours of shapes. Both types of visual information can represent aspects of scene content, LSF captures larger surface areas of a scene, while HSF captures mainly areas of rapid change in the scene. Thus, both provide different aspects of scene information that could be used for the recognition of a gist category.

Early research on spatial frequency processing focussed primarily on the importance of LSF for rapid scene processing, which lead to the so-called coarse-to-fine hypothesis. This hypothesis defines temporal aspects of visual processing: coarse LSF information is accessed first after which finer, HSF information becomes available. The coarse-to-fine hypothesis states that this is caused by the difference in speed by the two different pathways carrying spatial frequency information: LSF information is conveyed to the inferior temporal cortex by the fast magnocellular pathways, while HSF is conveyed by the slower parvocellular pathways (Bar, 2004; Kauffmann, Ramanoël, & Peyrin, 2014). However, a large body of mainly macaque research shows that frequency tuning between magnocellular and parvocellular cells often overlaps considerably (Skottun, 2015), suggesting that the distinction in frequency sensitivity between the two pathways might not be as clear cut as sometimes believed.

In gist research, the coarse-to-fine hypothesis was investigated by experiments combining low and high frequency information from two scenes into hybrid images. It was found that the perceived scene category of a flashed hybrid image was predominantly the low frequency scene category at 30 ms, and the high frequency scene category at 150 ms (Schyns & Oliva, 1994). However, these hybrid images were constructed by merging the LSF information from one scene with the HSF of another. This breaks spatial space contiguity, disrupting boundaries and contours. Additionally, the scene category of both the low and high frequency scene in a hybrid image had a priming effect after 30 ms exposure time (Schyns & Oliva, 1994), indicating that both LSF and HSF were available to the visual system. Later research even showed that preceding exposure to images containing either meaningful LSF or HSF content influenced the perception of subsequently viewed hybrid images, with over 70% of perceived categories matching the previously seen meaningful spatial frequency band (Oliva & Schyns, 1997). Thus, gist extraction of a hybrid image might not be equivalent to that of a natural scene image, and both frequency bands are available after extremely brief exposure times, and the relative weight of LSF and HSF in our gist perception depends on both exposure time and task-relevance.

Anatomical properties of the visual system seem to underline the importance of LSF, as gist can be extracted from images in foveal regions but also far periphery. Peripheral vision has lower resolution and consequently cannot process high spatial frequencies that are processed foveally. The outer boundary of central vision differs depending on definition used, from only the macula (0 – 3.6°) to extending it up to the perifovea at 10°, while peripheral vision extends to ~62 degrees binocularly, and 105-110 degrees monocularly – the latter also called the far periphery (Loschky, Boucart, et al., 2015). Even when scenes were presented at large visual eccentricities up to 70°, gist extraction of superordinate scene category (naturalness, openness, expansiveness) was still reliable (Boucart, Moroni, Thibaut, Szaffarczyk, & Greene, 2013). However, reaction time increased and performance decreased with increased eccentricity, especially for more detailed categorisation (indoor/outdoor), which was at chance-level at 70°. A different study presented images for 28 ms, placed randomly at 1 of 9 locations, for an animal presence detection task, showed performances of 60.5% correct at 70° and performance increased almost linearly towards central vision (Thorpe, Gegenfurtner, Fabre-Thorpe, & BuÈlthoff, 2001). This suggests that gist extraction can occur even with the lower resolution of peripheral vision, especially for more broad categorisations (natural/man-made), although accuracy does go down, potentially because HSF information is less accurately extracted at the lower resolution of peripheral vision.

On the other hand, recent neuroimaging research has found that scene-selective areas respond preferentially to HSF rather than LSF information. The parahippocampal place area (PPA) showed increased activity with HSF checkerboards, scenes and even faces, compared to their LSF equivalents (Rajimehr, Devaney, Bilenko, Young, & Tootell, 2011). Similarly, when contrast was equalized, the PPA and the occipital place area (OPA) were activated more by HSF than LSF versions of indoor and outdoor scenes, while there was no difference in the retrosplenial cortex (RSC) (Kauffmann, Ramanoël, Guyader, Chauvin, & Peyrin, 2015). Thus, HSF seems to preferentially activate the two of the important scene processing areas, the PPA and OPA.

To get a more meaningful measurement of the role of spatial frequency bands in scene processing, further neuroimaging research has investigated the relationship between spatial frequency and scene categorization. Scene category could be decoded from BOLD signals in the PPA, RSC, and lateral occipital complex (LOC) from both photographs and their corresponding line drawings viewed for 2 seconds (Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011), suggesting the important role of contours and edges (HSF) in scene category encoding. Similarly, computational models could decode scene category from the brain activation patterns evoked when viewing HSF natural scene images for 800 ms from all tested scene-related regions (PPA, RSC, OPA and LOC), while for LSF images scene category could only be decoded in the posterior PPA (Berman, Golomb, & Walther, 2017). Overall, these correlational studies strongly suggest that HSF rather than LSF carries meaningful information of scene category in scene-selective areas during rapid viewing (800 ms), although this might not hold the same for ultra-rapid presentations (~30 ms).

One last point of note is that hybrid images or separate LSF and HSF presentations might not be directly analogous to a full natural image. Instead, LSF and HSF might be integrated in a super-additive manner, as viewing the LSF + HSF of a scene led to superior levels of gist extraction for a vehicle presence detection task (100 ms exposure), compared to the probability summation of separate LSF and HSF viewing (Kihara & Takeda, 2010). However, this super-additive performance was not found with exposure time up to 83 ms, suggesting that the super-additive integration of LSF and HSF information for scene categorization becomes available somewhere around 100 ms after image onset. Thus, integration of spatial frequencies occurs after approximately 100 ms viewing time and might be super-additive, leading to higher performance than based on probability summation of separate frequencies.

Thus, ultimately, and unsurprisingly, both LSF and HSF play important roles in rapid scene categorization, and their relative importance might differ depending on factors such as categorization type, location of the image on the retina, and exposure time. LSF information might be available earlier and could drive gist extraction for ultra-rapid presentations, but once available, evidence suggests that HSF might encode scene category information to scene-selective areas more strongly. LSF and HSF are integrated after only brief exposure time, which increases gist perception accuracy. Overall, spatial frequency information is extracted rapidly and globally through gist and the information carried by LSF and HSF both play an important role in informing scene categorization.

1.2.1.2. Global spatial envelope properties

Combining spatial frequencies and their orientation as a pattern across an image can be summarized as the global spatial envelope (Oliva & Torralba, 2001). This spatial envelope contains outlines of surfaces and their surface properties, such as textures. The global spatial envelope can be extracted using calculations of principal component analysis of the Fourier spectrum of the image. Similarly, the visual system can extract these spatial patterns through the Gabor-like receptive fields of orientation sensitive cells. Each pattern of orientations can then be summarized with their scores for descriptive labels such as naturalness, openness, roughness, expansion, ruggedness, transience, or mean depth. When the global property scores of scenes are projected on a multidimensional space, semantically related scenes are often grouped closely together. Differences in naturalness score can predict scene categorisations such as natural versus man-made scenes (Oliva & Torralba, 2001), as man-made scenes have more straight horizontal and vertical lines, and natural landscapes have a wider distribution of edge orientations, textured zones, and smooth contours, influencing their global spatial envelope properties. Taking this further to specific scene categories, openness score could differentiate between a street (closed) and a highway (open), which have similar naturalness and expansion scores. Thus, combinations of global spatial envelope properties can be used to categorize scenes and might be used by the visual system for gist extraction.

The importance of global spatial envelope properties for gist extraction was clearly demonstrated by Greene and Oliva (2009), who showed that the chance of a false alarm for a distractor scene during rapid viewing (30 ms exposure) could be accurately predicted and even computationally modelled based on its similarity to the target category in its global spatial envelope properties. When a set of trials contained distractors from a different category but similar global properties, there were more false alarms and if a hit occurred, it was associated with longer reaction times. This suggests that similarities in global properties result in gist signals that are harder to distinguish from each other, which increases false alarms and reaction times.

In conclusion, global spatial envelope properties and their descriptive labels directly relate to accuracy and reaction time, suggesting these properties play an important role in gist extraction. Gist categorization is more accurate and faster when the categories have dissimilar global spatial envelope properties from each other.

1.2.2. Summary statistics

Natural scenes contain a lot of regularities, such as the patterns from trees in a forest. And while these regularities can be informative in the form of spatial structural regularities, they can also be viewed as redundant. It is not necessary to process each individual tree to know you are in a forest. And where there is redundancy, information can be compressed with a more efficient coding scheme. This can produce summary statistics that can efficiently capture general information of a group of similar elements (e.g., trees).

Summary statistics capture the average and distribution of a feature without access to specific objects or their locations. For example, summary statistics can efficiently encode the mean and distribution of sizes of a group of circles. Indeed, observers accurately reported mean size of ensembles of 4 to 16 circles after brief exposure (500 ms), but they were unable to distinguish individual circles from random ones with similar sizes (Ariely, 2001). Mean size of an ensemble could even be extracted with just 50 ms of viewing time (Chong & Treisman, 2003). The same applies for other basic visual features, such as orientation (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), velocity and direction of motion (Williams & Sekuler, 1984), hue (Maule, Witzel, & Franklin, 2014), or the centre of mass (Alvarez & Oliva, 2008) of object ensembles. Interestingly, larger set sizes have been reported to slightly increase performance and reduce reaction times when reporting mean size or orientation (Parkes et al., 2001; Robitaille & Harris, 2011), suggesting that summary statistic processing occurs in parallel over the whole ensemble. The advantage of larger set sizes suggests that summary statistics are an efficient averaging procedure, where noise of individual items cancels out to increase accuracy.

While most summary statistics capture averages of basic visual features, observers can also extract summary statistics of more complicated features, such as the average emotion or gender of face ensembles (Haberman & Whitney, 2007), which could be extracted from sets of 4 or 16 faces shown for 2 seconds with similar discrimination thresholds for both set sizes. Again, observers did not have access to individual faces in the ensemble, as performance at a 2AFC task to identify individual faces was at-chance. Further research indicates that emotion extraction from 16 faces became more noisy with shorter exposure times of 500 or even 50 ms, but still occurred with above-chance accuracy (Haberman & Whitney, 2009). Thus, summary statistics can capture averages of complicated features, even at brief exposure times, occurring globally and without providing access to individual elements, and could thus play a role in gist extraction.

The use of summary statistics would allow gist extraction to occur in peripheral vision, as the averaging procedures would compensate for the increased size of the receptive fields in the periphery. The influence of summary statistics on peripheral vision can be demonstrated by a phenomenon called crowding, in which identification of a target is impaired by surrounding distractors, called flankers. Balas and colleagues (2009) suggested that crowding is caused by averaging of peripheral vision into summary statistics. This was supported by the similarity between performance on 4-AFC letter-detection in peripheral ensembles and foveal viewing of summary statistic representations (mongrels) of these ensembles. Thus, summary statistics might be extracted especially in the periphery of our vision to increase the accuracy of our perception of global features.

In addition to being available rapidly and globally, multiple summary statistics can also be extracted in parallel, suggesting the process does not require focused attention. Indeed, observers could monitor for both numerosity and mean size of a single ensemble, reporting one of the two with a post-cue without effects on accuracy (Utochkin & Vostrikov, 2017), and participants' numerosity and mean size estimates of the same ensemble were not correlated. This suggests that both statistics were calculated independently, in parallel. However, this might not be the case processing multiple ensembles simultaneously. When observers had to monitor mean size and/or numerosity of two differently coloured ensembles of circles simultaneously, dual task cost occurred, reducing accuracy. Similarly, largest mean size and mean orientation was more accurately perceived in 2 sets of sequentially presented duos than when 4 ensembles were presented in one view (Attarha & Moore, 2015), again showing a dual task cost of monitoring multiple ensembles. Other studies also identified reductions in accuracy with more than two subsets (Halberda, Sires, & Feigenson, 2006; Poltoratski & Xu, 2013). Lastly, in an ensemble of coloured letters, observers could accurately report both letter identity and letter colour proportions, but judgement of conjoined features (proportion green Ts) was much less accurate (Treisman, 2006). This indicates that processing of a summary statistic is limited to a low number of separate ensembles, but that the visual system can monitor for multiple distinct summary statistics in one ensemble, however these summary statistics are separately calculated and not easily bound to each other in specific combinations.

What's more, summary statistics do not require focused attention to be calculated. While observers were engaged in a foreground object tracking task, they were more likely to notice changes in the structured background when these changes altered the ensemble structure, showing that summary statistics allow automatic detection of changes in a naturalistic background, even under reduced attention to the ensemble due to the tracking task (Alvarez & Oliva, 2009).

In summary, summary statistics are calculated rapidly on a global ensemble, without requiring focused attention, and without providing access to individual objects within the ensemble. Summary statistics can be calculated for multiple visual properties simultaneously, and these properties can be simple (size, orientation) or more complicated (emotion). By averaging across items, summary statistics reduce redundancy and increase accuracy of the global properties of the image, which can contribute meaningfully to gist extraction.

1.2.3. Intermediate disjunctive features

Intermediate disjunctive features were first introduced under the Feature Integration Theory (FIT). FIT states that visual attention consists of two processes occurring serially: First, features are registered globally and in parallel under distributed attention. Then, focused attention is then needed to correctly bind the features of one object to each other based on a shared location (Treisman & Gelade, 1980). In the first stage, in absence of focused attention, features can be considered ‘free-floating’ (Treisman, 2006; Treisman & Schmidt, 1982). Distributed attention provides general parameters of statistical properties, while losing access to local object information. It spreads globally over the entire display or over a set of similar objects (Treisman, 2006), giving access to features of the attended group. These features are for example global shape, boundaries, and relations between elements. Distributed attention does not allow access to correctly conjoined features, as supported by participants’ inability to accurately estimate the proportion of green T’s in a coloured shape array with 500 ms exposure time, while their estimations of proportion of green shapes or T’s separately remained accurate (Treisman, 2006), similar to the summary statistics discussed above. Distributed attention is also thought to provide access to intermediate disjunctive features (Evans & Chong, 2012).

Intermediate features are formed from a combination of basic visual features, such as colour and shape, such as exemplars of for example a wing, beak, or arm. They are distinct features, rather than just a sum of these basic features. Neural evidence for intermediate features comes from animal research which has shown that higher cortical areas contain neurons that are preferentially tuned to specific intermediate features. For example, some of the neurons in the inferotemporal area of macaques selectively responded to elementary components of natural objects, such as a T-shaped element or a circle with a smaller circle protruding from it (Tanaka, 1996). Similarly, inferotemporal neurons showed preferential activation when the monkey was viewing monkey hands or faces (Gross, Rocha-Miranda, & Bender, 1972). Thus, there is evidence of neural representations of intermediate features. Computational evidence shows that an algorithm trained on intermediate features of faces and cars outperformed an algorithm based on either basic or more complex features, and was more generalizable, meaning it performed better stimuli outside of the training set than the basic or complex features model (Ullman, Vidal-Naquet, & Sali, 2002). This shows that intermediate features can increase the flexibility of categorisation.

Indeed, intermediate disjunctive features can influence rapid categorization of presence or absence of broad object categories within a scene. Firstly, absence of diagnostic features (eyes, muzzle/beak, limbs) in a rapidly viewed scene (32 ms) highly impaired animal detection in both accuracy and reaction times (Delorme, Richard, & Fabre-Thorpe, 2010). Similarly, when human distractors were added to an animal and vehicle detection task during rapid serial visual presentation (RSVP; 75 ms per frame), accuracy of animal detection went down, likely due to similarity in their intermediate disjunctive features, while there was no effect on vehicle detection (Evans & Treisman, 2005). In the same experiment, participants were asked to identify and localize the animal, which occurred with above-chance accuracy when the animal presence was correctly detected: localization occurred in 53% of the cases when the animal was correctly detected, and correct identification occurred in 44%. However, this is not as high as would be expected if the visual system had access to the animal object. Instead, it was theorized that an educated guess could be informed by the combination of intermediate features (furry, antlers → stag), leading to some correct identifications. Indeed, many of the incorrect identifications were of the correct general category (mammal, bird etc.), suggesting that observers were aware of likely animal sub-categories. Thus, intermediate features can provide rich information about the possible scene content. This is also the case for scene categorization, where the presence of a salient incoherent object (e.g. tree or animal in man-made scene, boat or human in natural scene) reduced accuracy and slowed reaction times (Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007), showing that intermediate features can also influence broader scene categorization. Thus, extracting intermediate features play a role in gist extraction by allowing rapid, generalizable detection of presence of for example animals or humans within a scene, which in turn can also guide the perceived scene category.

1.2.4. Non-selective process

Part of the definition of gist extraction given at the start of this thesis stated that processes involved in gist extraction should not rely on focused, selective attention. Instead, they occur over whole visual field**,** and extract features are not bound to locations or to other features co-occurring in the same location. However, gist information does require attention to be entered into visual processing, which is often described as distributed or non-selective, which is needed to access the unbound features. A non-selective process can process multiple items in parallel without selection, which leads to a higher capacity limit. Thus, as gist extraction should not require focused attention on the stimulus, it should not suffer from serial processing limitations.

Indeed, gist extraction for an animal detection task in the periphery occurs without diminished performance during a demanding foveal letter discrimination task requiring focused attention (F. F. Li, VanRullen, Koch, & Perona, 2003). When two scenes are presented in parallel, performance on a rapid animal present/absent task drops only slightly. This small drop in performance matches independent parallel processing, as for example the likelihood of a false alarm increases when processing two distractor images instead of one. Similarly, reaction times remained the same during parallel processing of two scenes for animal presence detection, and no changes in occipital neuronal activity could be detected using electroencephalogram (EEG) (Rousselet, Fabre-Thorpe, & Thorpe, 2002), providing evidence for the parallel nature of gist extraction of two scenes. This was further supported by a follow-up study using 4 quadrants for inter- and intra-hemifield presentations of 1, 2, or 4 scenes (Rousselet, Thorpe, & Fabre-Thorpe, 2004). Again, the small drop in performance for the 2 and 4 scene conditions fit with the model of parallel processing. Thus, gist processing indeed shows properties of a non-selective process, as it occurs simultaneously with selective tasks or on multiple scenes in in parallel without large effects on performance.

What's more, gist extraction does not require pre-existing knowledge of the exact targets to monitor for, further demonstrating the non-selective nature of gist extraction. When a task is known before image onset, goal-directed top-down processes can tune lower level neural responses to be more sensitive to the stimulus, for example, increasing sensitivity to task-relevant features in monkeys performing a bisection or Vernier task (W. Li, Piëch, & Gilbert, 2004). In many gist experiments, a limited set of pre-cued properties is used, which does not account for the possibility that perhaps a limited amount of non-selective filters can be active at the same time. A study comparing pre- and post-cues with 9 possible categories showed that pre-cue performance was significantly higher, but post-cue detection still outperformed expected levels if observers could only monitor for one cue (Evans, Horowitz, & Wolfe, 2011), suggesting that multiple filters can be active simultaneously. Even more strikingly, the RSVP study by Potter, Wyble, Hagmann, and McCourt (2014) also showed above-chance performance even if the target category was revealed 200 ms after the RSVP (Potter et al., 2014). Similarly, gist extraction can monitor for multiple cues simultaneously in one image. Observers could effectively monitor for the presence of one or both of the categories (e.g. animal and/or beach), with increased performance when detecting the presence of at least one cued category (Evans, Horowitz, et al., 2011). Thus, gist extraction can occur for on multiple cues simultaneously, or even when the target is unknown, removing top-down task-context influences.

However, there is some evidence that attentional limits do exist in gist extraction, as studies show inattentional blindness and dual-task cost effects, indicating that gist extraction is not a completely 'cost-free' parallel process. When observers are unaware that a specific feature will appear and should be monitored, they often do not consciously perceive it – this is called inattentional blindness. Dual-task cost occurs when performance on the primary task goes down when simultaneously performing the secondary task. Both inattentional blindness and dual-task costs occur for summary statistics. A substantial proportion of participants was unable to report the ensemble property colour diversity of a letter grid when they were tasked with monitoring letter identity in a cued row, showing inattentional blindness. Similarly, letter identity accuracy was reduced when they subsequently monitored for both the cued row and colour diversity of the ensemble (Jackson-Nielsen, Cohen, & Pitts, 2017), indicating that monitoring the summary statistic carried some attentional demand. Likewise, for scene gist, during a primary task of multiple object tracking task or an RSVP of letters/digits, an unexpected scene background as the second to last mask was often not perceived (inattentional blindness), and subsequent monitoring for both the primary task and scene gist decreased performance (dual-task cost) (Cohen, Alvarez, & Nakayama, 2011). So, while gist extraction is generally a highly efficient process that can occur in parallel to itself or other demanding tasks, a level of attention and alertness is required to process perception of the gist, which can lead to inattentional blindness when unaware of the need to process the gist for reporting, or to dual-task costs when monitoring gist alongside certain demanding primary tasks.

So, as gist extraction requires a certain amount of attention, this thesis will briefly discuss some of the attentional theories for visual processing that best fit with the observed qualities of gist extraction. Most visual attention theories define a split between two types of attention, one accessing the image features globally, and the other deploying attention to locations for precise processing and refined object recognition. These theories should account for the rapid processing of gist, as the time course of gist extraction clearly shows that it does not rely on fixation of individual elements and is unlikely to involve deployment of attention to specific locations. Thus, gist extraction belongs to the global axis of attention. One attentional theory which illustrates how gist could be processed rapidly and with minimal attentional demands is the reverse hierarchy theory, which is supported by anatomical, neuroimaging, computational, and behavioral evidence, which will be discussed below.

1.2.4.1. Reverse hierarchy theory

The reverse hierarchy theory (RHT) was informed by the cortical structure of the visual system. Neurons in early visual cortex areas, like the V1, often have small receptive fields, they respond to for example a precise orientation of a specifically sized bar of light in one part of the visual field. On the other hand, higher cortical areas contain neurons with larger receptive fields, which are often tuned to more general, higher-order stimuli, such as a specific face. Indeed, neuronal responses in the inferior temporal cortex of macaques were robust to changes in viewpoint and size of complex shape stimuli (Vogels & Orban, 1996). According to the RHT, as visual information enters the visual system, a feed-forward sweep of visual information is propagated through the network and rapidly reaches these higher visual areas. There, the arrival of this first sweep of activity enables ‘vision at glance’: a high-level, general interpretation of the world, associated with sparse attention (Hochstein & Ahissar, 2002). This way, explicit, but global perception is formed in high cortical areas. Next, re-entrant feedback from the higher cortical areas returns towards lower cortical areas, which fine-tunes and adapts the binding of features to their respective objects and results in ‘vision with scrutiny’: A detailed perception of subordinate categories. Re-entrant feedback imposes a top-down influence on further computations in the lower cortical areas (Gilbert & Li, 2013). A constant feedback loop between lower and higher cortical areas further refines perception. In the RHT model, gist extraction should occur in the feed-forward sweep, the first information that reaches higher visual areas before re-entrant processing occurs, meaning gist perception is part of ‘vision at glance’.

EEG research identified three stages of visual processing: a pre-110 ms stage with feedforward flow of information towards mainly extrastriate visual areas, a post-110 ms stage, where re-entrant processing to early visual areas such as V1 occurred, and a 200-300 ms stage which relies on the earlier stages, with activity in extrastriate areas and beyond. It also showed that the pre-110 ms stage was uninterrupted by masking, while the post-110 ms stage re-entrant processing was disrupted (Fahrenfort, Scholte, & Lamme, 2007). Thus, demonstrating that re-entrant processing takes place after an initial feed-forward only stage, which is not disrupted by subsequent masking.

If gist extraction is indeed achieved through a feed-forward sweep, gist extraction should occur even when a briefly viewed image is masked immediately afterwards to disrupt re-entrant processing. Indeed, in an RSVP where the next image functionally masks the previous one, observers could perform an absence/presence 2-AFC of scene category with above-chance accuracy even if each image was only shown for 13 ms (Potter et al., 2014). However, natural scenes might not sufficiently mask all areas of the previous images, which might allow some persistent visual processing. Using more structured geometric or coloured line masks, RSVP detection of scene category occurred reliably at 53 ms per image, but not 13 or 27 ms (J. F. Maguire & Howe, 2016). This indicates that the length of exposure for processing likely needs to exceed 27 ms, but 53 ms exposure is still within the 110 ms of the feedforward stage described by Fahrenfort et al. (2007).

Computational models of visual processing based on human physiology also offer support for the feed-forward nature of gist extraction, matching with the RHT. It is well-established that neurons in higher processing areas of our visual system have increasingly stable responses to objects or scenes, regardless of differences in for example position or scale (Gross et al., 1972; Hubel & Wiesel, 1968; Logothetis, Pauls, & Poggio, 1995; Perrett & Oram, 1993; Quiroga, Reddy, Kreiman, Koch, & Fried, 2005), and that their receptive fields simultaneously increase in size (Perrett & Oram, 1993; Smith, Singh, Williams, & Greenlee, 2001; Tanaka, 1996), which together allow humans to flexibly recognise variations of objects or scenes from different viewpoints (Evans & Chong, 2012). Feed-forward models of processing show that these higher-order neurons can rapidly respond to their preferred stimuli, without needing top-down feedback. For example, a feedforward model of the primate ventral stream trained on an animal absent/present categorization task produced performance patterns that were highly comparable with human rapid visual processing (Serre, Oliva, & Poggio, 2007). Similarly, a simulation of a single wave of spikes (neuronal firing) showed that this first sweep of information sufficed for broad recognition of faces in natural images by a computational model (VanRullen, 2007), clearly demonstrating that the feedforward sweep is sufficient to extract rich information about our visual environment.

In summary, gist extraction occurs through a separate mode of processing distinctly different from the processing underlying detailed object recognition, in line with the RHT. Gist extraction is non-selective and does not require focused attention, allowing it to occur during a different focal task, on parallel displays, and with post-trial cues. Similarly, the feed-forward nature of gist extraction is supported by psychophysics experiments, in which gist extraction occurs with rapid presentations followed by masking, even if the task-context is not known beforehand to remove top-down influences. This behaviour can also accurately be described with feed-forward computational models. This non-selective, feed-forward sweep of visual processing gives access to spatial structures, summary statistics, and intermediate features that together inform gist extraction rapidly and globally.

1.3. Neural mechanisms of gist extraction

As defined above, gist extraction is made up of a collection of processes, combining low-, mid-, and high-level visual features, ranging from orientation, spatial frequency, and colour (low), to shapes, depth, and textures (mid), to faces, bodies, and objects (high) (Groen, Silson, & Baker, 2017). It therefore seems unlikely that there is a single “gist extraction” brain area. Instead, gist extraction likely takes place across a network of cortical regions, which extract spatial structural regularities, summary statistics, and intermediate disjunctive features and integrate these into gist categorizations of the global visual environment.

So, while it is unlikely that one singular area performs gist extraction, there are some brain areas that might be involved. A likely group of candidate areas is the scene processing network, which consists of areas that are known to respond preferentially to scenes over objects and/or faces: The parahippocampal place area (PPA) (Downing, Chan, Peelen, Dodds, & Kanwisher, 2006;

Epstein, Harris, Stanley, & Kanwisher, 1999; Epstein & Kanwisher, 1998), the occipital place area (OPA) (Downing et al., 2006; Grill-Spector, 2003), and the retrosplenial cortex (RSC) (Henderson, Larson, & Zhu, 2008). Additionally, while the lateral occipital complex (LOC), is mainly known as an object-selective area (Grill-Spector, Kourtzi, & Kanwisher, 2001), there is also evidence of some evidence for LOC activity to scenes.

As a rapid process, another important aspect of gist extraction is its timing. Only neural activity patterns that are observable after rapid exposure can influence gist extraction. Scene-specific activity occurs early in visual processing. Comparing scenes to objects and faces, no difference was found in P1 (80-130 ms), but the P2 (200-320 ms) peak amplitude was significantly higher for scenes with 500 ms exposure (Harel, Groen, Kravitz, Deouell, & Baker, 2016). The N1 or N170, commonly described as face-processing specific, did not differ between objects and scenes, which both had a lower peak amplitude than faces. Magnetoencephalography (MEG) measurements of neural activity to faces and outdoor scenes (buildings, landscapes) showed even earlier onset, with scene-specific activity significantly higher in a left and right hemisphere medio-occipital region during the M100 (100-130 ms after stimulus onset) with 1000 ms exposure (Rivolta, Palermo, Schmalzl, & Williams, 2012).

Thus, the scene network might play a role in gist extraction, and some neural correlates related to scene processing are available early in processing after rapidly viewing a scene. However, differential activity between scene categories is more relevant for gist extraction, where differences in activity patterns and amplitudes would be expected between for example natural and man-made scenes. If indeed spatial structural regularities play a role in informing gist extraction, one would also expect differential neural activity based on these visual properties. In the next sections, evidence for gist category representations in neural activity is explored, as well as evidence for neural representations of different spatial regularities.

1.3.1. Category processing

Early neural correlates of scene categories have been observed when comparing natural and man-made scenes. These evoked differential activity in the N170 and P2, which for the P2 was modulated by spatial expanse (open/closed), but not distance (close/far) of natural scenes (Harel et al., 2016). A hierarchical linear regression model containing contrast energy, spatial coherence, and naturalness ratings factors explained 22.7% of the P2 amplitude, showing the influence of image properties on P2 amplitude as well asperceived naturalness of the scene. An even earlier effect of naturalness was found when observers viewed scenes (exposure time 1.5 s) during a memory task, where significant decoding between natural and man-made scenes first occurred at 125 ms at an occipital electrode (Oz), and differential activity remained significant during the

P2 (150-275 ms) (Lowe, Rajsic, Ferber, & Walther, 2018). In both experiments, stimulus type and category information were not task relevant. These findings suggest that early activity, such as the P2 ERP, automatically captures diagnostic scene information which could inform global scene categorization, and additionally showed the influence of image properties such as contrast energy and spatial coherence.

Further research at gist extraction timing within environmental scenes again found early differential activity depending on task-relevant category, differentiating between targets and distractors depending on the specific task. In natural scenes, EEG signals diverged between GO (animal) and NO-GO (no animal) trials 150 ms after stimulus onset (20 ms exposure) (Thorpe, Fize, & Marlot, 1996). Specifically, distractor scenes without animals evoked a strong frontal negativity starting at 150 ms after stimulus onset. This difference occurred with a widely varied set of images, and it was unrelated to the reaction time of the trial (faster reaction times did not evoke earlier differences). The same was observed by Bacon-Macé et al. (2005) in both frontal and occipital electrodes, even when the scene was masked 12 ms after the scene was flashed, although longer mask latency increased the amplitude of differential activity. This differential activity might be related to decision making in the no-go trials, for example inhibiting a GO response. Later studies similarly reported frontal negativity from 160-170 ms, but additionally showed differential occipito-temporal positivity amplitudes at around 150 ms (Delorme, Rousselet, Macé, & Fabre-Thorpe, 2004; Rousselet et al., 2002). The same frontal and occipito-temporal activity, with additional parietal activity was also found at 150 ms in an animal/no animal 2 alternative forced choice (2-AFC, 30 ms exposure), showing that the differential activity cannot be solely attributed to differences in motor commands between a go and no-go trial (Antal et al., 2001).

Later processing might only occur when the scene content is task-relevant, as prolonged scene category sensitive ERPs (>250 ms) observed during rapid natural/man-made categorization task are not present when the scenes are not task-irrelevant (Groen, Ghebreab, Lamme, & Scholte, 2016). Additionally, when scenes were task-relevant, neural activity beyond 250 ms remained correlated to the scene's spatial coherence. This suggests that later differential activity occurs only when scenes are attended for decision making but is still influenced by spatial layout properties. It also underlines the importance of task-relevance during research into gist extraction, as much research instead uses passive viewing or attentional tasks such as monitoring for change in fixation cross colour/length, in which the scene is not task-relevant, which means later stages of gist extraction are under-represented or absent.

Lastly, a brief view at the areas in the scene selective network shows that some of these areas might be functionally involved in gist extraction as their activity is related to scene categorization.

Intracerebral electrodes in the PPA of epilepsy patients demonstrated scene-selective activity 80 ms after stimulus onset and scene-category differential activity at 170 ms (Bastin et al., 2013). Similarly, disrupting OPA processing with transcranial magnetic stimulation impaired rapid (~100 ms viewing time) scene, but not object categorization (Dilks, Julian, Paunov, & Kanwisher, 2013). Thus, the PPA and OPA likely play a functional role in scene categorization. On the other hand, patients with RSC lesions retain unimpaired scene recognition (E. A. Maguire, 2001). Instead, the RSC has been shown to mainly be involved in spatial navigation, as well as wider cognitive tasks, such as recalling real or imaginary events (for a review of RSC research refer to Vann, Aggleton, and Maguire (2009)). Thus, the RSC might support various cognitive functions, rather than specifically being involved in scene category processing. Lastly, LOC's activation patterns to scenes seem to be mainly driven by the objects in the scene. Activity in the LOC evoked by a scene containing a task-relevant object correlated with isolated representations of the object (Peelen, Fei-Fei, & Kastner, 2009). Similarly, scene-evoked activity in the LOC could be predicted by the average activity patterns of signature objects (e.g. bathtub, toilet → bathroom) (MacEvoy & Epstein, 2011), while there was no such relationship for PPA activity. Thus, the PPA and OPA show functional relevance for global scene categorization, while this is less likely for the RSC and LOC.

Distributed pattern analysis of fMRI activity showed that scene categories (1600 ms viewing time) could be decoded from activation patterns with above-chance accuracy in the LOC (24% correct), PPA (31%), RSC (27%), and V1 (26%) (Walther, Caddigan, Fei-Fei, & Beck, 2009). Additionally, error patterns of the computational models based on LOC and PPA activity correlated with human errors in a rapid scene categorization task (14-45 ms exposure, 6AFC), showing that human gist categorization made similar mistakes as decoders based on PPA and LOC activity, suggesting gist extraction might utilize information that is encoded in the PPA and LOC. However, top-down re-entrant feedback of scene content from other areas might have provided scene-context information to the RSC and LOC, as scenes were viewed for 1600 ms. Indeed, a study with scenes differing in spatial layout and object content (300 ms exposure) showed that only spatial layout could be decoded from RSC activity, while both layout and object content could be decoded from the PPA, and only object content could be decoded in the LOC (Harel, Kravitz, & Baker, 2013). Thus, the RSC seems to be sensitive to spatial layout for navigation, while the LOC is sensitive to objects in a scene context. But the PPA encoded both scene content (objects) and spatial layout, again emphasizing its likely involvement in scene gist extraction.

1.3.2. Spatial layout processing

Various spatial layout properties influence neural activity during rapid processing of scenes, supporting how these properties could be available to guide gist extraction. For example, spatial coherence reflects the level of scene fragmentation, with higher scores indicating more variety/fragments, while contrast energy represents the amount of contrast throughout the scene. Spatial coherence and contrast energy correlated with both neural activity and behavioural performance on a natural/man-made categorization task with 100 ms exposure time (Groen, Ghebreab, Prins, Lamme, & Steven Scholte, 2013). Single trial linear regression analysis showed that a combination of spatial coherence and contrast energy correlated with EEG amplitude between 109 and 137 ms across occipital and parietal electrodes. Additionally, spatial coherence influenced peri-occipital activity up to 250 ms after stimulus onset. Behaviourally, higher spatial coherence increased the likelihood of subjects rating a scene as natural. Interestingly, participants' perception (natural/man-made) could be predicted from EEG signals as early as 80 ms, showing a clear link between neural activity and perception. At the early phase, predictions of perceptual ratings mainly relied on occipital and peri-occipital activity, but after 260 ms activity was distributed across the scalp, suggesting initial localized processing followed by a distributed cortical representation of the gist category of the scene.

Size (e.g. small: kitchen, large: factory hall) and clutter level (low/high) of indoor scenes could also be decoded in MEG signals during attentional viewing without a categorization task (Cichy, Khosla, Pantazis, & Oliva, 2017). Clutter decoding accuracy peaked first, at 107 ms, and size decoding peaked approximately 250 ms after stimulus onset (500 ms exposure). Size and clutter decoding activity were independent of each other, scene category, or low-level visual properties contrast or luminance, showing they were distinct representations of scene properties. Clutter level might be related to the level of openness in a scene, although the current study did not review this possibility. What's more, both the PPA and the RSC showed differential activation to level of clutter (amount and organization of objects) in scenes (Park, Konkle, & Oliva, 2015).

In summary, early neural activity correlates with various spatial layout properties and human perception, suggesting that these properties could be extracted rapidly and become available for gist extraction. Spatial layout processing could give rise to summary statistics or global spatial envelope characteristics that could be diagnostic of scene category. Localization of spatial coherence neural correlates suggest that initial processing of structural layout properties occurred in occipital and parietal regions, after which distributed cortical processing occurred.

From the scene network areas, the PPA seemed to be the most promising for scene gist categorisations. Looking at visual properties of scenes, the PPA has also been shown to be sensitive to changes in surface properties such as texture or geometrical shape of object ensembles (Cant & Xu, 2017), scenes, and even objects in scenes (Lowe, Rajsic, Gallivan, Ferber, & Cant, 2017), while the PPA responded markedly less to these changes when they occurred in isolated objects. Additionally, activation patterns in the PPA were most related to the expanse in the scene (open/closed), with less influence of distance (near/far), or, surprisingly, content (man-made/natural) or scene category (Kravitz, Peng, & Baker, 2011). Thus, the PPA shows early activation, which seems to represent spatial structure and texture of scenes, which could inform gist categorisations.

Texture and layout could be decoded from neural activity in artificially constructed rooms shown for 2 seconds (Henriksson, Mur, & Kriegeskorte, 2019). The OPA was mostly sensitive to layout, irrespective of surface textures. In contrast, both the V1 and PPA were more sensitive to differences in surface texture than layout but showed highest scene discriminability when both layout and texture varied. Lastly, the RSC showed no consistent differential activity to the room stimuli. A separate MEG experiment showed that the texture-invariant layout representations could first be identified at 60 ms and peaked at 100 ms after stimulus onset. Additionally, the effect of layout on MEG activity patterns correlated with OPA activity in the fMRI experiment, suggesting layout information is available early in processing in the OPA. Both the PPA and OPA responded to changes in relative length and angle of a simple scene layout (300 ms exposure), but not to similar changes in length and angle of objects, showing scene-specific sensitivity to layout, while the RSC did not respond differentially to any of these changes (Dillon, Persichetti, Spelke, & Dilks, 2018).

Lastly, spatial layout also has a functional role in scene categorization. Scene category could be decoded from both color photographs and line drawings in V1, V2+VP, V4, PPA, and RSC (Walther et al., 2011). LOC activity did not allow decoding of scene category, potentially because objects were less diagnostic of scene category in this experiment. Interestingly, scene category of line drawings could be decoded by a model trained on photographs, and vice versa, in the primary visual cortex and PPA areas. This shows that contours and edges play an important role in scene categorization for computational models, as photographs and line drawings could interchangeably be decoded from activity patterns in primary visual cortex and PPA.

Overall, it seems that PPA and OPA might have complementary roles regarding spatial structure and textural content, with the OPA responding specifically to layout, while the PPA is tuned both texture and layout, which both could inform gist extraction, while the role of the RSC is less clear and might be driven by top-down re-entrant processing. Neural representations of spatial coherence, clutter, spatial layout and other spatial properties were shown to be available rapidly after briefly viewing a scene, indicating that these signals would be available for use in gist categorization.

1.3.3. Summary of neural mechanisms

In summary, gist extraction is indeed likely facilitated through a network of scene-selective cortical areas, which provide information on aspects such as scene layout (OPA, PPA), global spatial envelope, textural information, and shape (PPA), and scene content (PPA, LOC), potentially driven by intermediate features of diagnostic objects. These areas and the associated encoded information can be used to decode the scene category, showing their functional relevance. Scene-selective areas likely give rise to a distributed representation of gist.

Perceptually driven effects of scene or object-category-in-scene activity are evident rapidly after stimulus onset, around 80 ms. After 150 ms, meaningful and behaviourally relevant differential activity is consistently found, especially in frontal and occipital regions. This occurs even when re-entrant processing is disrupted by masking, which indicates that sufficient processing has occurred to differentiate task-relevant categorical features to inform gist extraction. This also fits with the feed-forward nature of gist extraction under the RHT. At 200 ms (P2), clear scene-category specific activity can be differentiated, and differential activity persists after 250 ms in a distributed fashion, especially when the scene category is task-relevant. Spatial structural properties influence this neural activity.

1.4. Gist extraction in Medical Image Perception

As shown in the introduction, medical experts can detect a gist of abnormality in chest radiographs, mammograms, and cytology within 200-250 ms, (Evans, Georgian-Smith, et al., 2013; Kundel & Nodine, 1975). Additionally, medical gist extraction is not constrained to just 2D images: videos from stacks of breast tomography images, each frame displayed for 20 ms, allowed for above-chance performance, even after exclusion of trials where the abnormality was localized (Wu, D'Ardenne, Nishikawa, & Wolfe, 2019). Similarly, expert readers of prostrate images could extract the gist of abnormality from a serial display of slices from a 3D imaging device (Treviño et al., 2020). Thus, gist of abnormality is a global signal that can be extracted by experts in many different medical disciplines and in 2D and 3D displays, demonstrating its flexible nature.

Certain global measures, such as mean breast density (Boyd et al., 2010; Vachon et al., 2007), as well as bilateral symmetry of breast density (Zheng et al., 2012) and breast volume (Scutt, Lancaster, & Manning, 2006) have been shown to be significant predictors of breast cancer risk, which might lead to the suggestion that gist extraction of medical abnormality in mammograms could be explained by these factors. However, unilateral mammograms of the breast containing a cancerous abnormality still contain a strong gist signal without any way to extract (a)symmetry information (Evans, Haygood, Cooper, Culpan, & Wolfe, 2016), showing that symmetry measures may assist distinguishing abnormal from normal cases, but is not required. Additionally, while mean breast density ratings also categorize abnormalities with above-chance accuracy, the d' prime is lower than that of gist ratings, and density and abnormality ratings were not correlated across images, thus, breast density is not the driving factor for the gist of medical abnormality in mammograms.

In line with the non-selective, global nature of gist signals, medical gist extraction does not provide accurate localisation information. Radiologists recognised abnormalities with above-chance accuracy after 250 ms exposures but were unable to localize these abnormalities on a subsequent masked outline. One later study suggested that partial localization of a region of interest might sometimes occur within a gist extraction timeframe (250 ms) when an abnormality is correctly rated as such (Carrigan, Wardle, & Rich, 2018), although even then, localisation accuracy was around 35% for low density mammograms, in which abnormalities are generally easier to localize, and only 10% for high density mammograms**.** In conclusion, while partial localisation information might be available in some instances, extracting the gist of medical abnormality is largely reliant on global features and does not consistently allow access to locations of abnormalities.

The fact that detecting the gist of abnormality is a global process that does not require localized features is emphasized by the fact that it still occurs when viewing sections without the localized abnormality. Firstly, gist of abnormality could be extracted from unilateral images of the contralateral breast with above chance accuracy (Evans et al., 2016), although this signal is notably weaker than in the breast containing the abnormality. Strikingly, gist can even be extracted from small patches taken from either the breast containing the abnormality or the contralateral breast with above-chance accuracy, whether this patch contained the abnormality or not (Evans et al., 2016). This further strengthens the argument that gist of abnormality is present globally, even in the parts of the breast parenchyma that do not contain the cancerous abnormality. Extending on this, gist of abnormality is even detectable in so-called priors: mammograms of women diagnosed with cancer taken 3 years prior to their eventual diagnosis. In these priors, no actionable signs of cancer could be found even after additional inspection, but gist ratings were above-chance accurate (Patrick C. Brennan et al., 2018; Evans, Culpan, & Wolfe, 2019). Thus, gist of abnormality is a global structural signal, that can even be detected before local abnormalities are apparent, suggesting it could be used as a risk-marker, to for example increase screening for at-risk women. Overall, these findings all support the assertion that the signal of abnormality must originate from robust, global differences between normal and abnormal breast tissue. For example, an abnormality might distort the statistical and global regularities of a normal breast parenchyma, or conversely there could be specific statistical and global regularities associated with an abnormality.

As discussed, the role of spatial frequency bands has been extensively investigated in scene gist extraction, with mixed findings on the relative importance of low and high frequency information for rapid categorization of scenes (see section 1.2.1.1. low and high spatial frequency bands). Expanding this research to medical gist showed that low-pass filtering strongly reduced accuracy of rating normal vs abnormal mammograms from a d’ of 1.06 with full spectrum mammograms to only 0.26, while high-pass filtered mammograms retained most gist information, with a d’ of 0.96 (Evans et al., 2016), while localization remained at-chance across all conditions. Thus, gist of abnormality seems to be preferentially contained in higher spatial frequencies, although more detailed assessments are needed to further narrow down the specific roles of frequency bands.

Interestingly, a study where mammograms were presented for 1 second, an inversion effect occurred, where accuracy was higher for upright than inverted mammograms – which is indicative of holistic processing(Chin, Evans, Wolfe, Bowen, & Tanaka, 2018), which was stronger for experts than novice readers (residents). This suggests that compound processing plays a role in recognizing the gist of medical abnormality. However, the inversion effect was not present with shorter exposure times of 500 ms (K.K. Evans, personal communications, 2020), suggesting that holistic processes might occur to improve detection of abnormalities in medical images but likely does not occur on the same rapid time scale as gist extraction.

Lastly, gist extraction performance is (unsurprisingly) strongly correlated with expertise in radiologists reading mammograms (Evans et al., 2019). More specifically, performance correlated the number of cases read in the previous year, but not with years in practice or percentage of time spend reading mammograms, reflecting the importance of recent perceptual rather than medical expertise. What’s more, experience in breast tomography correlated with performance on gist perception in digital breast tomosynthesis images (C. C. Wu et al., 2019), while experience in digital mammography did not show a significant correlation, suggesting that this correlation is modality-specific and driven by recent perceptual experience, rather than pure medical knowledge.

One important consideration is how scene and medical gist can best be compared to each other to support how we can use medical gist research to further our general knowledge of gist extraction. Categorization of normal versus abnormal medical images can be viewed as a superordinate level categorization task, as complex medical images are categorized on a broad level (normal or abnormal), rather than specific types of abnormalities (for a review see Makki (2015)). This influences comparisons with scene gist studies, as these have often used animal/vehicle presence/absence detection tasks, which are more akin object-in-scene categorization, which might rely more heavily on for example intermediate disjunctive features rather than global spatial envelope properties. A better comparison for medical gist is studies investigating superordinate scene categorization, such as natural/man-made distinctions, or to a lesser extent basic level categorization (beach, forest etc.). Thus, it is important to be aware of distinctions between different stimuli and categorization tasks in gist extraction research, and how they influence our comparisons between scene and medical gist literature.

1.5. Goals of this thesis

This thesis has three aims that each further our knowledge about gist extraction in medical images. The first overarching aim is to investigate which **perceptual properties** influence the extraction of the gist of medical abnormality and how. The second aim is to study if **gist of a new category can be learned** through perceptual training, and how this ability develops. In concert with both of these aims, the third aim of this thesis is to explore the neural signature of gist extraction from mammograms in medical experts.

1.5.1. Perceptual properties

As set out in this introduction, gist extraction is characterized by its rapid speed of processing, but less is known about the effects of longer exposures on our ability to process and access gist categorizations. Many attentional theories make a clear distinction between global and local, rapid and slow, or distributed and focused processing, but it is not known how longer exposure time influences our ability to extract gist information and access gist categorizations. A switch might occur when local, selective information becomes available, that might make gist information less relied upon, but still accessible as a secondary source of information, or even inaccessible. However, it is also possible that gist extraction keeps accumulating information as more time is available to process the gist signals, which might actually strengthen our gist perceptions. Thus, the question is: Does the gist signal only influence perception during initial rapid exposure, or does it remain available during later stages of processing – and if so, is it perhaps an additive process, where longer exposure can strengthen and improve accuracy of gist signal, leading to a more accurate ‘first impression’? **Chapter two** investigated the differences in performance of medical experts on global assessments of medical images between rapid flashes (500 ms) or unlimited exposure times.

Next, spatial frequency bands underlying spatial structural regularities clearly play an important role in scene perception, especially in gist extraction. Spatial frequency gives information about global shapes, spatial layout, borders/contrasts. It also underlies the global spatial envelope properties that have been shown to influence the perceived gist of scenes. However, the roles of low and high spatial frequency information in gist extraction remain unclear, with conflicting evidence for the relative importance of one or both. One previous study suggested the importance of HSF for medical gist extraction. **Chapter three** investigated the effect of different levels of high-pass spatial filters on perception of the gist of abnormality in medical experts, in order to further investigate the role of low and high spatial frequency in medical gist extraction.

1.5.2. Learning processes

The gist of medical abnormality forms a key opportunity to investigate the learning processes underlying a human's ability to learn to recognise the gist of novel categories. It is largely unknown how humans gain the knowledge needed to accurately distinguish a beach from a forest, in a robust fashion, from a variety of viewpoints and variations, all within the blink of an eye. The development of this expertise is difficult to investigate, as scene gist abilities are present in all healthy adults, and it is unknown what the childhood development of gist extraction is. Furthermore, developmental studies using children face confounding factors from other developmental factors, such as communication constraints. On the other hand, medical experts have gained the ability to extract a new category of gist in their field of expertise, a skill which the general population does not possess. We therefore know this ability is not innate, and there is evidence showing its strong relation to recent perceptual exposure (number of images seen) rather than medical knowledge (years of experience). What's more, gist extraction should be generalizable and rely on global patterns rather than specific elements, to enable observers to recognise a beach or abnormal mammogram across the wide variability that exists within that category (e.g., viewpoint or breast density). This fits with statistical learning, the process through which humans can extract naturally occurring statistical patterns in space and/or time (Turk-Browne, Jungé, & Scholl, 2005), without feedback on the exact features to extract. Statistical learning of spatial regularities might allow observers to learn to recognise the invariant global properties of a forest, beach, or even an abnormal mammogram through perceptual exposure to many exemplars of the category, without explicit feedback on the visual features that represent the category. Training naïve observers on medical images will allow us to investigate whether humans can learn to recognise the gist of a novel categorization through perceptual exposure alone.

This thesis aims to investigate whether people can learn to recognise the gist of a new category through perceptual training with global feedback alone. **Chapter four** contains a multi-session online perceptual training experiment to see if people can learn to recognize the gist of a new category, as well as exploring individual differences in this ability. Additionally, human learning patterns were compared to the performance of computational models to gain insight into the overlap in information captured by the neural networks and human perception.

1.5.3. Neural signature

Lastly, while some studies have investigated the neural correlates of scene gist extraction, no such study has been performed for the gist of medical abnormality. Investigating the neural signature of extracting the gist of medical abnormality will provide further insight into similarities and potential differences between scene and medical gist. **Chapter five** used EEG measurements to investigate the brain activity patterns evoked by viewing and rating normal and abnormal mammograms in a group of expert radiologists. Calculating the differential activity between these two categories gives insight into areas that carry information on the gist of medical abnormality. Single subject bootstrapping allowed an in-depth exploration of the neural activity patterns in each individual radiologist.

The research in this thesis will increase our understanding of medical gist extraction, by scrutinizing image parameters influencing the gist of abnormality, and will make a first effort to explore the learning of a new gist categorization. Assessing neural activity patterns in medical experts will allow understanding of the neural mechanisms involved in recognition of the medical abnormality gist and subsequent decision making. This thesis also has important practical applications for the medical imaging field. Firstly, gist of abnormality could provide a novel medical diagnosis tool for mammography screening. For example, a rapid triage of cases for their gist scores, in which cases with high gist scores would be prioritized for further assessment and maybe even receive earlier invitations for their next screening appointment, as high gist scores are a known risk factor for development of breast cancer. Further understanding of the timeline of medical gist extraction (Chapter 2) could inform us how this triage system could best be implemented, while knowing how high-pass filters influence the accuracy of gist extraction could allow us to finetune the mammogram presentations to boost performance (Chapter 3). Secondly, knowledge on the learning of the gist of abnormality (Chapter 4) could inform new, effective medical training approaches to help residents reach higher levels of perceptual expertise earlier in their career or to keep perceptual performance high in experienced medical experts that might have less exposure to cases in their day-to-day roles.

Chapter 2: Comparable prediction of breast cancer risk from a glimpse or a first impression of a mammogram.

Chapter 2 was published as Raat, E.M., Farr, I., Wolfe, J.M. *et al.* Comparable prediction of breast cancer risk from a glimpse or a first impression of a mammogram. *Cogn. Research* **6**, 72 (2021). This version of the article has been accepted for publication after peer review but is not the Version of Record and does not reflect post-acceptance improvements. It also has edited headers and figure numbers to fit with the thesis format. The Version of Record is available online at doi.org/10.1186/s41235-021-00339-5

2.1. Declarations

Ethics approval and consent to participate

All participants provided informed consent before participating in this study. The study received ethical approval from the Departmental Ethics Committee of the Department of Psychology, University of York, United Kingdom (ID: 805).

Consent for publication

The mammograms used in the methodology figure were sourced from the OPTIMUM database and have been fully anonymised.

Availability of data and material

The datasets generated and analysed during the current study are available on our OSF repository, dx.doi.org/10.17605/OSF.IO/5NWP8

This data is available under Creative Commons Attribution-NonCommercial-ShareAlike 2.0 UK: England & Wales (CC BY-NC-SA 2.0 UK).

Competing interests

JMW holds a role as Editor-In-Chief for the CPRI journal but was not involved in the peer review of this paper.

Funding

EMR is funded through a departmental scholarship from the department of Psychology at the University of York, United Kingdom for the 3-year duration of her PhD project.

JMW was supported by NIH-NCI grant **CA207490.**

KKE was supported by Cancer Research UK & EPSRC grant EDDCPJT\100027

These funding bodies had no influence on the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

EMR made substantial contributions to the conception and design of the study, to the acquisition of the radiologist no time limit data and the naïve participant data, to the analysis and interpretation of data, and has drafted the work.

IF made substantial contributions to the acquisition of the naïve participant data.

JMW made substantial contributions to the design, analysis, and interpretation of the data and to the editing of the manuscript.

KKE made substantial contributions to the conception and design of the study, to the acquisition of the radiologist time limit data, to the analysis and interpretation of the data, and revising the work.

Acknowledgements

Not applicable

2.2. Abstract

Expert radiologists can discern normal from abnormal mammograms with above-chance accuracy after brief (e.g., 500 ms) exposure. They can even predict cancer risk viewing currently normal images (priors) from women who will later develop cancer. This involves a rapid, global, non-selective process called “gist extraction”. It is not yet known whether prolonged exposure can strengthen the gist signal, or if it is available solely in the early exposure. This is of particular interest for the priors, that do not contain any localizable signal of abnormality. The current study compared performance with brief (500 ms) or unlimited exposure for four types of mammograms (normal, abnormal, contralateral, priors). Groups of expert radiologists and untrained observers were tested. As expected, radiologists outperformed naïve participants. Replicating prior work, they exceeded chance performance though the gist signal was weak. However, we found no consistent performance differences in radiologists or naïves between timing conditions. Exposure time neither increased nor decreased ability to identify the gist of abnormality or predict cancer risk. If gist signals are to have a place in cancer risk assessments, more efforts should be made to strengthen the signal.

Key words: gist, radiology, mammography, holistic impression, gestalt

2.3. Significance statement

Breast cancer is the most common cancer in women and causes the highest number of cancer-related deaths in women globally. Because early detection is highly beneficial to treatment outcomes, mammographic screening for breast cancer is widely implemented. In earlier work, we have found that expert radiologists can detect a ‘gist of abnormality’ at above chance levels after a brief exposure to a mammogram from a woman with cancer, even if the cancer is not visible in the image (e.g., the image is from the contralateral breast). The gist signal can be detected in mammograms acquired several years before the woman is diagnosed with actionable cancer. If the gist signal is to be of clinical use, it would help if it were more robust. Previous studies used brief exposures for research purposes (e.g., to thwart eye movements). Here we test if a stronger signal is available when no time limit is imposed. We did not find any effect on accuracy measures, so the effort to strengthen the signal will need to pursue other paths.

2.4. Introduction

The visual system has the remarkable capability to extract information about our environment in the proverbial blink of an eye. Within a 100 ms, humans can identify the general meaning (or "gist") of what they are seeing (Potter, 1975). They can extract information about the scene category (Greene & Oliva, 2009) or detect the presence of certain object categories (Bacon-Macé et al., 2005). Gist extraction is a global, non-selective process, by which our visual system rapidly extracts structural and statistical regularities over the whole image to make broad categorizations of the stimulus perceived (Wolfe et al., 2011). The global, non-selective nature of the process means that the observer might be quite sure something like an animal is present but not be sure of its precise identity or location (Evans & Treisman, 2005).

This rapid gist extraction also occurs with specialized scenes like radiological images. To a non-expert, the gist of a mammogram may be nothing more than 'this is a mammogram'. However, expert radiologists can extract a "gist of abnormality" (Evans, Georgian-Smith, et al., 2013) from a brief glimpse of, at least, some medical images. Medical experts can distinguish abnormal from normal images with above-chance accuracy after rapid exposures. Experimental studies typically use exposures of 250 to 500 ms. Reliable detection of this gist of abnormality has been found for different types of medical images, for example chest radiographs (Kundel & Nodine, 1975), prostate images (Treviño et al., 2020), cervical micrographs in cytology as well as 2D mammograms (Evans, Georgian-Smith, et al., 2013) and 3D breast tomosynthesis (C. C. Wu et al., 2019).

While the exact perceptual features driving the extraction of the gist of abnormality are not yet known, previous research has investigated several potential factors. Breast density, which is known to be a predicting factor for breast cancer (Boyd et al., 2010; Vachon et al., 2007), cannot explain the gist signal, as it is less predictive of abnormality than gist, and shares only a small and negative correlation (r -0.10-0.26), with gist ratings on the same cases (Evans et al., 2019). Similarly, global symmetry between the two breasts might facilitate gist ratings of abnormality, but is certainly not essential, as gist ratings of unilateral abnormal cases reached d' of 1.16 (Evans et al., 2016), showing that, while symmetry may assist distinguishing abnormal from normal cases, it is not required. On the other hand, there seems to be an important role of high spatial frequencies, as performance dropped considerably when high frequency information was removed (low-pass filtered d' = 0.26). High-pass filtered images supported performance (d' = 0.96) that was not markedly worse than full spectrum images (d' = 1.06) (Evans et al., 2016).

One of the leading lines of evidence that the gist of abnormality is global in nature is that the gist can be detected even when no lesions are present in the presented image. Radiologists detected the gist of abnormality in patches of breast parenchyma that did not include the lesion as well as in mammograms of the breast contralateral to the one with the cancerous abnormality (Evans et al., 2016). Under these conditions, performance is reduced, but still above-chance (d' = ~0.4 for patches, ~0.6 for contralateral breast). There is evidence that the global gist of abnormality is present even before any visibly actionable cancerous abnormalities are present. Radiologists distinguished between 'abnormal' mammograms, taken 3 years before a woman developed any actionable abnormalities and 'normal' mammograms from women who did not develop cancer. Accuracy was above chance with 500 ms exposure (Patrick C. Brennan et al., 2018; Evans et al., 2019) to these 'prior' images. Thus, gist of abnormality is a relatively small, but robust, global signal present in medical images, although the exact perceptual features contributing to the gist of abnormality remain a gap in the literature that requires further research.

The existence of this gist of abnormality may initially sound implausible. However, think about your first glimpse of a store. You might ask yourself if you are likely to find something that you want here. You could not do this perfectly in half a second, but neither would you be at chance. Your expertise as a consumer would allow you to register the gist of the store, even if the item you wanted was not in that first view. An expert radiologist can do something similar with a mammogram.

Unsurprisingly, gist extraction performance does not reach the performance levels obtained by experts when the stimulus remains visible during regular clinical reading. For example, a d' of 1.0 was found for gist extraction of chest radiographs in 200 ms, compared with a d' of 2.5 achieved during free-viewing (Kundel & Nodine, 1975). Similarly, free-viewing of a set of mammograms in a laboratory setting produced a d' of 1.9 for distinguishing abnormal from normal images (Evans, Birdwell, et al., 2013), while 250 ms exposure produced gist performance of d' ≈ 1 with 250 ms exposure (Evans, Georgian-Smith, et al., 2013) and 1.14 after 500 ms exposure.

The increase in performance between rapid exposure and free viewing seemingly fits with two-stage detection models in medical image perception that propose to divide visual processing into an early and later stage. The first stage occurs rapidly and extracts global information about the image, not unlike gist extraction (Sheridan & Reingold, 2017). Swensson's *Two-Stage Detection Model* asserted that a first stage filters the image and identifies features that require further examination and that a second stage carries out a search over the identified locations (Swensson, 1980). Swensson argued that medical experts have acquired perceptual mechanisms that allow them to extract and use this global information more effectively than novices. Similarly, Nodine and Kundel's *Global-Focal Search Model* postulated that, when viewing a medical image, experts obtain a global impression of the image, which constrains their subsequent search (C. Nodine & Kundel, 1987). The global information is extracted from an image and compared to a schema built from prior knowledge. Schemas of normal and abnormal medical images help identify potential perturbations, and focal attention is guided to these locations for further examination. In an updated version renamed the *Holistic Model,* an expert rapidly assesses an initial holistic impression in order to constrain a subsequent search-to-find process. During the search-to-find stage, holistically identified perturbations are attended foveally, while the expert also scans the image for any less salient abnormalities that were missed in the holistic stage (Kundel, Nodine, Conant, & Weinstein, 2007). Kundel has argued for a model of radiologist performance that has a prominent role for an “initial holistic, gestalt-like” stage of processing that is conceptually quite similar to global gist processing as we have described here (Kundel, Nodine, Krupinski, & Mello-Thoms, 2008). However, there is an important difference between the holistic analysis of the image as Kundel et al. understand it and global gist processing as we are using it here. The holistic representation contains information used to guide attention to locations where targets are likely to be, while the gist representation is a non-localized sense that this patient might or might not have disease.

Another important difference between the Kundel account and global gist processing concerns the time frame. The holistic phase of the Kundel et al. model encompass roughly the first full second of the reading of an image. More modern work in visual attention would envision that first second to be a mix of fast global gist processing and selective attention to a substantial number of specific objects or locations in the field (Evans et al., 2016). In the global gist experiments, stimuli were flashed briefly (typically for 500 ms or less), for the purpose of limiting volitional eye movements and attentional scrutiny of the images. This raises an interesting question; would the global gist signal continue to grow if observers had more time to look at the image? Alternatively, might the signal *only* be available if the images are briefly presented? There are phenomena that behave in this way, vanishing if the observer sees the stimuli for too long (e.g. abnormal fusion in binocular vision (Wolfe, 1983)). Accordingly, in the present experiment we compare performance of novice and expert viewers who view mammograms either for 500 ms or for as long as they like. The most interesting conditions in this experiment are those where there is no localized pathology in the image. Is the gist signal bigger, smaller, or unchanged by the ability to look longer to establish a ‘first impression’.

2.5. Methods

We compared two experiments involving rapid assessment of the same set of image stimuli using two different groups of participants: novice and expert. The first experiment presented the images very briefly for 500 milliseconds while the second allowed unlimited viewing time but asked the observers to make a decision on the basis of their “first impression”. The main experimental observers were two groups of medical experts in radiology and the control group was a group of observers without medical experience (“naïves”). Prior research has shown that naïve participants, without medical training are unable to assess if a mammogram is abnormal or not in 500 ms (Evans, Georgian-Smith, et al., 2013). The control group allowed us to determine if naïve observers would have access to the “gist of abnormality” if they just had a bit more time. Radiologists were tested as part of the Medical Image Perception “pop-up” lab supported by the US NIH: National Cancer Institute at the annual meeting of the Radiological Society of North America (RSNA) in 2018 and 2019. The RSNA meeting presents a unique opportunity to test expert radiologists in numbers that are otherwise difficult to access. That opportunity comes with methodological constraints. A between-subjects design was needed as the RSNA setting did not allow for a sufficient time for ‘wash-out’ of memory for specific images between a first and second assessment of that image. Additionally, there is an inherent level of unpredictability of testing in such settings. This is reflected, for example, in the unequal numbers of observers in the two radiologist groups, one group tested in 2018, the other in 2019.

Participants

A total of 50 participants took part in this study. A group of 11 radiologists with experience in mammography (7 female, age 32 to 65 years, 11 right-handed) participated in the no time limit condition, while 16 radiologists took part in a 500 ms time limit condition version of the experiment (9 female, age 38 to 63 years, 12 right-handed), which was part of a previously collected dataset in which spatially filtered mammograms were compared to unaltered mammograms, of which the ratings for unaltered cases formed the dataset used in the current experiment. A single group of 23 naïve observers (21 female, age 18 to 33 years old, 21 right-handed) participated both in the no time limit and the 500 ms time limit conditions.

Radiologists in this experiment were all at least at the resident level, who were currently practicing reading mammograms. They were all experienced at reading mammograms in a clinical setting, which was defined as having read at least 2000 scans in the last year. The radiologists in the no time limit group read on average 5195 scans (std 2757, range 3000 to 10000) a year. They averaged 16 years in practice (std 9.6 years, range 4 to 30), and on average spent 63% of their time diagnosing mammograms (std 33%, range 15 to 100%) in their work. The radiologists in the rapid display time limit group read on average 5056 scans (std 3828, range 2000 to 12000) a year, averaged 22 years in practice (std 11.9 years, range 2 to 38), and on average spent 59% of their time diagnosing mammograms (std 35%, range 15 to 100%) in their work.

The lowest value of years in practice was slightly less than used as a cut-off for expertise in some previous studies, which used a cut-off of 5 years (Chin et al., 2018; Evans, Georgian-Smith, et al., 2013), but matches the minimum years in practice used by Carrigan et al. (2018). Additionally, number of annual cases is a key determinant for good reading performance (Rawashdeh et al., 2013). A study found that readers with 2000 to 4999 annual cases outperformed those who read 1000 cases or less on malignancy detection, but were not outperformed by those with more than 5000 annual cases (Reed, Lee, Cawson, & Brennan, 2010). Thus, the radiologists in this study could all be considered experienced observers of mammograms.

For the no time limit condition, radiologists were recruited during RSNA 2019. For the 500 ms time limit condition, radiologists were recruited during RSNA 2018. Naïve observers were undergraduates at the Psychology Department of the University of York (UK), participating for course credit. All participants had normal or corrected-to-normal vision. This study was approved by the Psychology Departmental Ethics Committee of the University of York, and all participants gave informed consent.

Two separate groups of radiologists were tested because a within-subject design would have required a sufficient time window between measurements to avoid memorization effects. This would not have been practical in the RSNA setting.

Stimuli and apparatus

The 500 ms group of radiologists saw a total of 120 stimuli. The 120 stimuli were mammograms of either mediolateral oblique (MLO) or craniocaudal (CC) view of two breasts (bilateral). Of these, 60 were abnormal, composed of 20 with obvious lesions, 20 with subtle lesions and 20 mammograms acquired 2 to 3 years prior to cancer showing no visibly actionable lesions at that time. The categories obvious and subtle abnormal were based on how easily detectable the abnormality was judged to be by an experienced collaborating radiologist. The other 60 were normal mammograms that did not contain cancerous abnormalities. The 60 normal mammograms were preassigned to the three categories of abnormal, so that each performance measure was calculated between 20 abnormal and 20 normal cases. Only the trials with subtle abnormal and prior stimuli, and their pre-assigned normal stimuli were analysed in this study, since these categories were also used in the other conditions, resulting in a total of 80 trials used for analysis.

The number of normal mammograms was reduced to a singular set of 20 normal cases in the no time limit condition (and both conditions for naïves) in an effort to reduce the duration of the experiment and increase ease of data collection given that in the no time limit experiment image viewing was self-paced. Thus, for the no time limit group of radiologists, and both conditions for naïves, results are based on 80 trials. The 80 stimuli were images of either MLO view or CC view of a single breast (see figure 1B for an example). These images were subdivided into four categories: normal mammograms of healthy women (normal), mammograms with relatively subtle cancerous abnormalities (subtle abnormal), mammograms of the breast contralateral to a breast containing a cancerous abnormality (contralateral), mammograms from women who later developed cancerous abnormalities but showed no visibly actionable lesions in these mammograms that were acquired on earlier screening (priors). Given that unilateral mammograms were presented in the no time limit experiment, we were able to add the category of contralateral images – images of a breast that did not contain a lesion but was contralateral to a breast that did contain a lesion. Thus, the no time-limit version of the experiment used a sub selection of the cases from the time limit version, 20 of the 60 normal cases from the time limit version, the 20 subtle cases which were split to create the unilateral subtle and contralateral categories, and all 20 prior cases. Neither priors nor contralaterals contained visible cancerous abnormalities, as determined by a study radiologist. Thus, they would have been labelled as 'normal' in regular practice. No mask was used in the no time limit condition, since the goal was to have unlimited visual processing until the participant chose to continue to the rating screen. Due to experimental limitations, the 500 ms condition of the naïves also did not include a mask, but since this would only increase the chance of naïves detecting the gist of abnormality, this is not considered a limitation.

For the radiologists, the images were presented on a 24' inch colour medical imaging display (1920 x 1200 pixels). For the naïve observers, the images were presented on 19.7' inch colour monitor (1280 x 1024 px). The stimuli, themselves, were presented in the centre of the screen at a size of 800 × 1000 pixels. The experiment was run using Matlab, utilizing the Psychophysics Toolbox 3 extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007). All mammograms were selected from the Complex Cognitive Processing lab database of stimuli, which can be shared with other researchers upon request to the last author (K.K. Evans).

Procedure

The procedures for both the no time-limit and time-limit version of the experiment were largely the same. The experiment consisted of 3 practice trials and 80 test trials (for no time-limit radiologists and for naïve observers) or 6 practice and 120 test trials (time-limited radiologists). In the practice trials, participants were familiarized with the display and rating screen, and feedback on the stimulus (normal or abnormal) was given after they confirmed their rating. On the test trials, no feedback was given. There were 20 trials for each of the abnormal types, but the time limit version for radiologists contained 60 rather than 20 normal cases (see stimuli and apparatus). Presentation order was randomized for each participant.

Each trial began with a white fixation cross presented at the centre of the screen (500 ms), followed by the mammogram being visible for either 500 ms (time-limited condition) or until the spacebar was pressed (no time-limit condition). For the time-limited experiment, the mammogram presentation was followed by presentation of a mask composed of the same breast outline, but with tissue replaced by a solid white field for 500 ms, before the rating screen was shown. No mask was used in the no time limit condition since the goal was to have unlimited visual processing until the radiologist chose to continue to the rating screen (see stimuli and apparatus). On the rating scale, participants used the mouse to move a slider to register their rating on the scale from 0 to a 100 (see figure 1A). Participants had to confirm their rating by pressing the spacebar, after which the next trial would start automatically. There was no masking display following the rating-scale screen.

Participants were asked to rate how certain they were that the image came from a woman with breast cancer or that the woman would develop cancer in the near future. The specific instructions given in the no time limit condition were: “You will be presented with 80 mammograms. View them for a time of your own choosing, but do not perform a detailed search of the image. Rather, focus on your first impression, your gut feeling, of the mammogram, without trying to scrutinize and search the image to localize abnormalities. Remember that 50 percent of the mammograms in the study contains or will develop cancer in the near future. You will then rate the mammograms on the likelihood of it containing cancer or developing it in the near future, based on your general impression, on a scale from 0, certainly no cancer, to a 100, certainly cancer present or will develop.” Instructions for the time limit condition were similar, except that it did not warn them to avoid detailed search, but instead emphasized that the image would only be visible for 500 ms.
Participants were asked to adopt a liberal rating criterion with regards to their decisions on whether a case contained or would develop cancer, while being as accurate as possible. There was no time constraint for choosing a rating in either condition, but participants were asked to report their first impression.

Different groups of radiologists participated in each of the two versions of the experiment (time limit of 500 ms and no time limit first impression). The versions were conducted a year apart. A single group of naïve participants participated in both the no time limit and the 500 ms time limit version in two different sessions, in a counterbalanced order. For naïve participants there was no masking used after the mammograms were presented in either experiment, due to the way the experiment was programmed. For naïves, each condition was tested in a separate session with at least one day and at most 1 week between sessions. Before each session, naïve participants were shown a short PowerPoint presentation to familiarize them with the concept of mammogram rating. This presentation explained how mammograms are made, how the brightness of the mammogram relates to tissue density, and common signs of abnormalities, as selected by a radiologist.

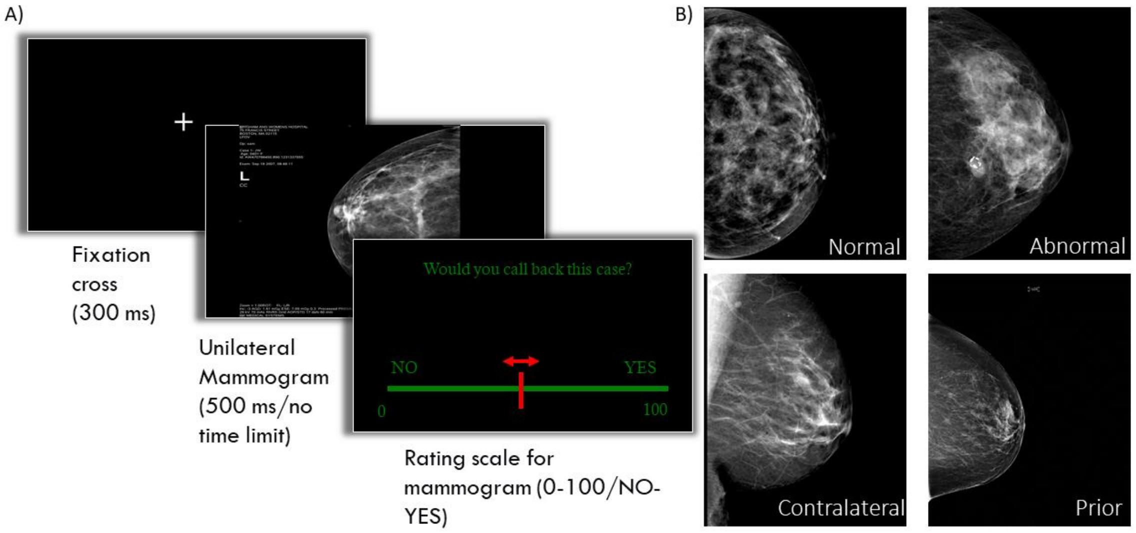

A)
Fixation cross (300 ms)
Unilateral Mammogram (500 ms/no time limit)
Would you call back this case?
NO
YES
0
100
Rating scale for mammogram (0-100/NO-YES)
B)
Normal
Abnormal
Contralateral
Prior

Figure 2.1: *Simplified overview of the experimental procedure (A) and example mammograms for each of the four types used in this experiment (B).*

Data analysis

The data were analysed using the framework of signal detection theory for binary classification. Given a rating, a mammogram was considered to be classified as either “abnormal” or “normal”, depending on whether the rating is higher or lower than some threshold. That classification was then compared to the ground truth. Signal detection measures were used to separately assess performance and response biases of the observer. Performance was represented by the d’ measure (d’ = z(true positive rate) – z(false positive rate)), where z denotes the inverse normal or z-transformation of the rates). In the cognitive literature, d’ is referred to as “sensitivity”. Unfortunately, “sensitivity” refers to the “true positive” or “hit” rate in the medical literature. We will refrain from using the term in order to avoid confusion. Response bias was measured by the criterion value, C (C = (z(true positive rate) + z(false positive rate))/-2). A negative criterion means that the observer was more likely to label the item as abnormal while a positive criterion means that observer was more likely to label the item as normal.

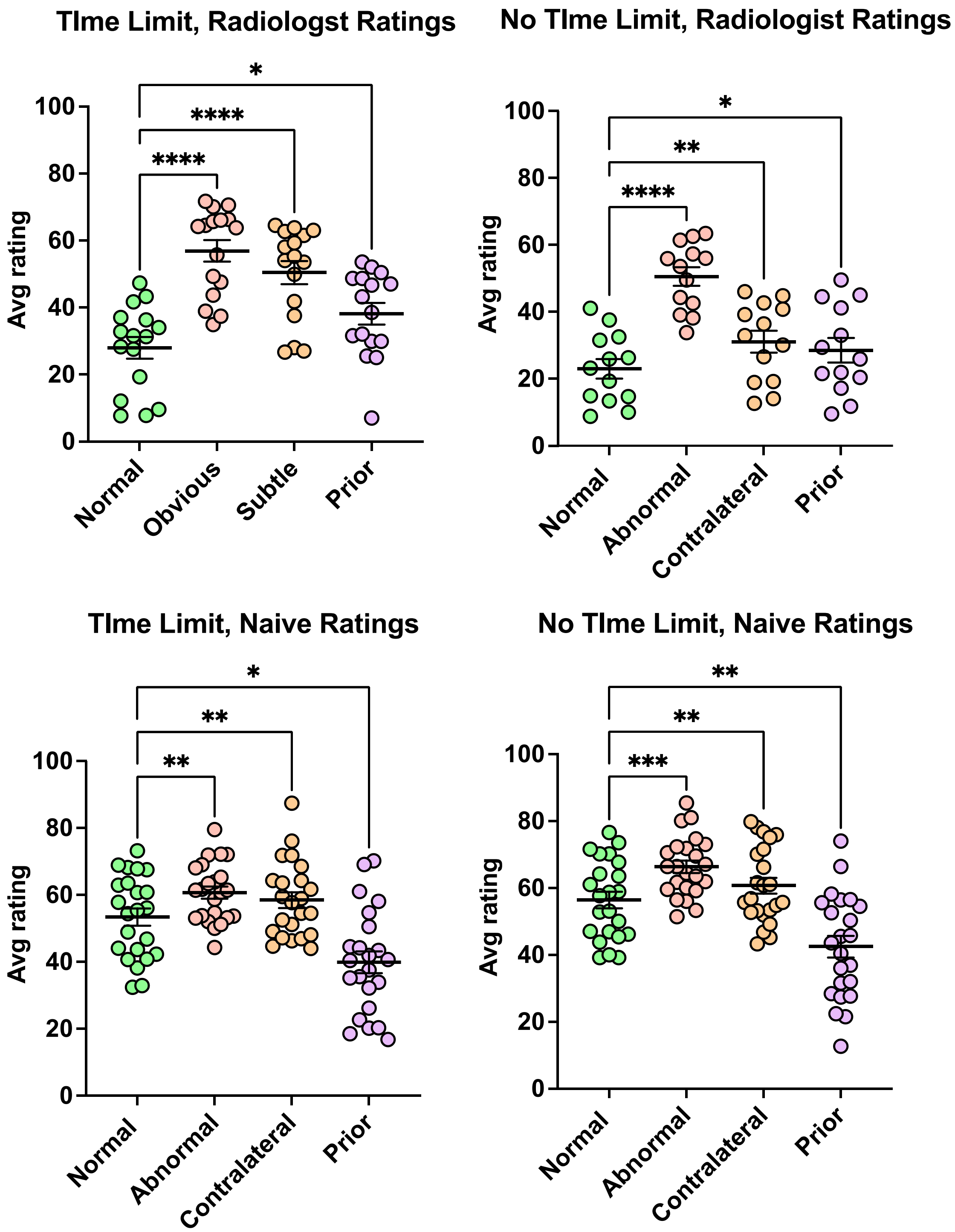
Receiver operating characteristic curves (ROC) were constructed by repeating this division of trials into proportions of true positive (hits) and false positive (false alarms) using different normal/abnormal rating cut-offs (here, 10, 20, 30, 40, 50, 60, 70, 80, and 90). The area under the curve (AUC) of an ROC, ranging from 0.0 to 1.0, represents the probability that a randomly chosen abnormal case will be rated higher than a randomly chosen normal case (Hanley & McNeil, 1982). Chance performance yields an AUC of 0.5. Higher AUCs indicate better performance in detecting the signal of cancerous abnormalities. AUCs were calculated using the trapezoid function in Matlab.

D’, criterion and AUC performance measures were calculated for each of the groups and conditions. For statistical analysis, we used the d’ and c values derived using a rating cut-off of 50, the middle of the ROC. In all cases, false positives were derived from ratings of 20 normal images that functioned as the negative cases, using the pre-allocated subset of 20 normal cases per image type in the radiologist time limit version, or the single set of 20 in the other experiments. The true positive rates were derived separately from responses to abnormal, contralateral, and prior images. Statistical analysis was used to compare these performance measures between image types, conditions, and group. The main statistical test used was mixed ANOVA, as there were the within-group measures of image type, and the between-group factors of either group (naïve/radiologist) and/or condition (500 ms/no limit). For comparing condition effects in naïves, a repeated measures ANOVA was used as this was measured with a within-subject design. Paired t-tests, corrected for multiple comparison, were used to compare specific conditions. One-sample t-tests were used to compare performance measures to chance.

In addition, reaction time (RT) data was collected in the no time limit condition. RT was defined as the time between the appearance of the mammogram and the time when the observer confirmed their rating. Average reaction time of radiologists and naïves was compared using an independent samples t-test. Repeated measures ANOVAs were used to compare reaction times within each group between image types.

Where possible, a combination of frequentist and Bayesian statistics are reported. Bayes factors can indicate the relative strength of evidence for two theories, where BF_{10} indicates the probability of the alternative compared to the null hypothesis under the observed data. Thus, Bayesian statistics can indicate whether a non-significant p-value from a frequentist test provides evidence towards the null hypothesis or if the evidence is insensitive (Dienes, 2014). The latter is generally considered the case with Bayes factors between 0.33 and 3. Values outside of this range provide evidence towards the null or alternative hypothesis, according to the heuristic classification scheme that was proposed by Jeffreys (1998) and is widely used to interpret Bayes factors. Bayesian statistics were calculated using the computer software JASP, version 0.14.1 (JASP-Team, 2020).

2.6. Results



S

Figure 2.2: *Average ratings for each observer group for each type of image. Statistical results are Dunnett's multiple comparisons tests, comparing each type of abnormal image to the normal images.*

Figure 2 shows the average ratings for each observer group (Radiologist and Naïve) for each type of image. For the radiologists, Dunnett's multiple comparisons tests show that all types of abnormal images are rated as significantly more abnormal than the normal images when viewing time was limited or unlimited (all $p < 0.05$). Interestingly, the data for the naïves also show significant differences between normal images and the other images, though the pattern of ratings is different than that seen with the radiologists. It is notable that the naïve observers rated the prior images as more normal than the normal images. This can be seen as type of artifact of stimulus selection. On returning to our image set, it appears that naïves might have used some rough assessment of density/complexity as a basis for their ratings, as the priors in this study are inadvertently systematically less dense than the normal images. The radiologists appear to be sensitive to some signal beyond density/complexity since they rate the priors as more abnormal. Since density and complexity are correlated with cancer risk, we can imagine that the radiologists took those factors into account as well. Had the images been more carefully balanced for density and complexity, it seems likely that the difference between radiologist ratings of normal and prior images would have been greater.

Turning to signal detection measures, Figure 3 shows that the ROCs for individual radiologists mostly lie above the diagonal chance performance line. As noted, the effects for the priors are weaker than what has been seen in other studies (Patrick C. Brennan et al., 2018; Evans et al., 2019), but this should be seen in light of the inadvertently lower density and complexity of the prior images.

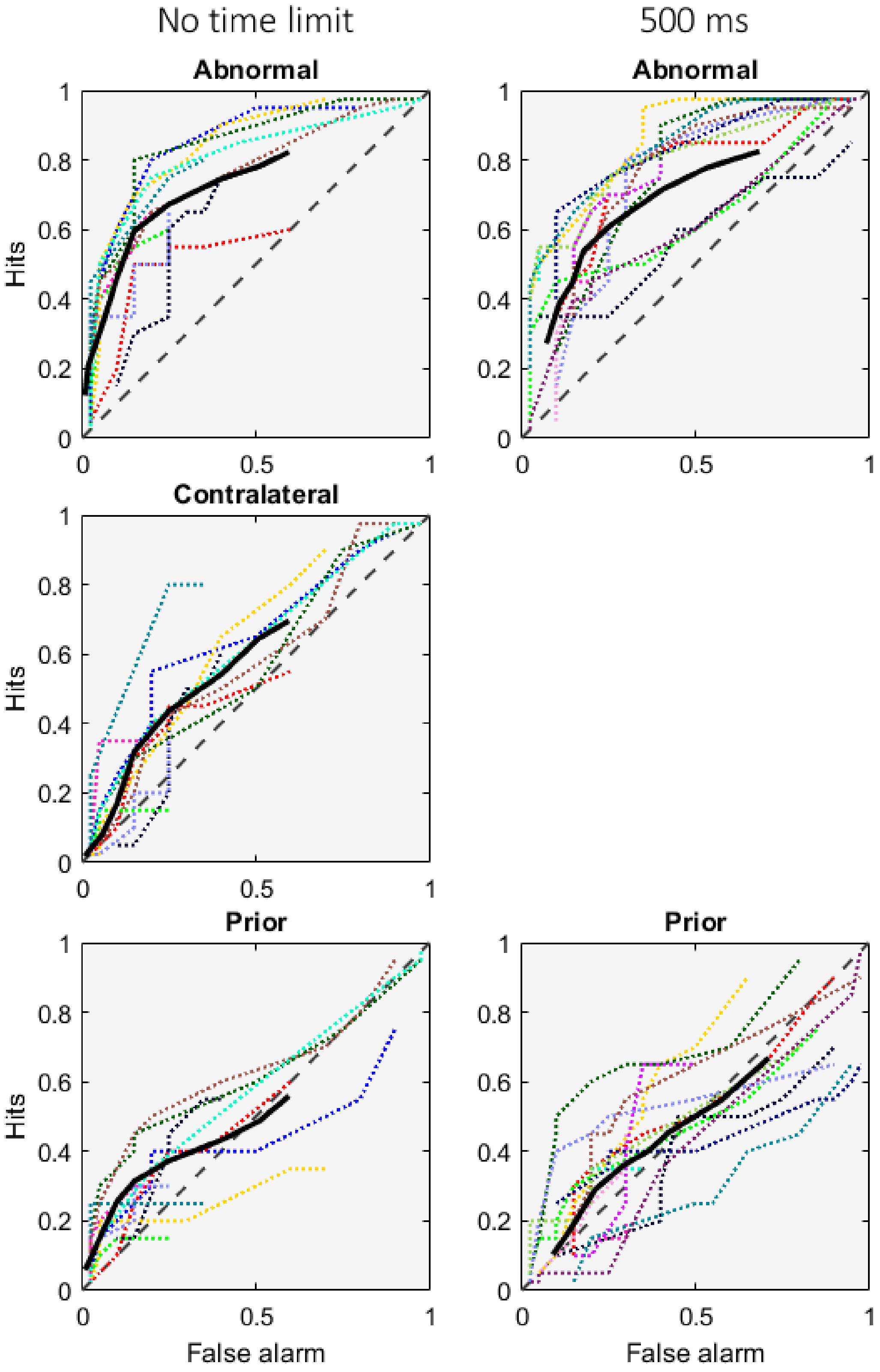
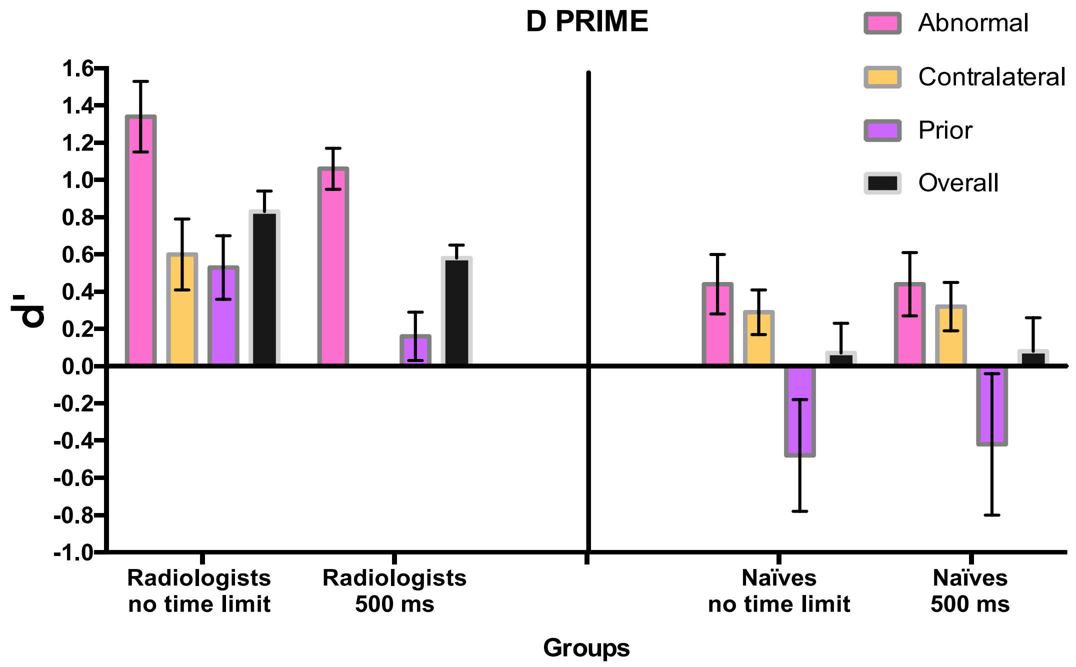
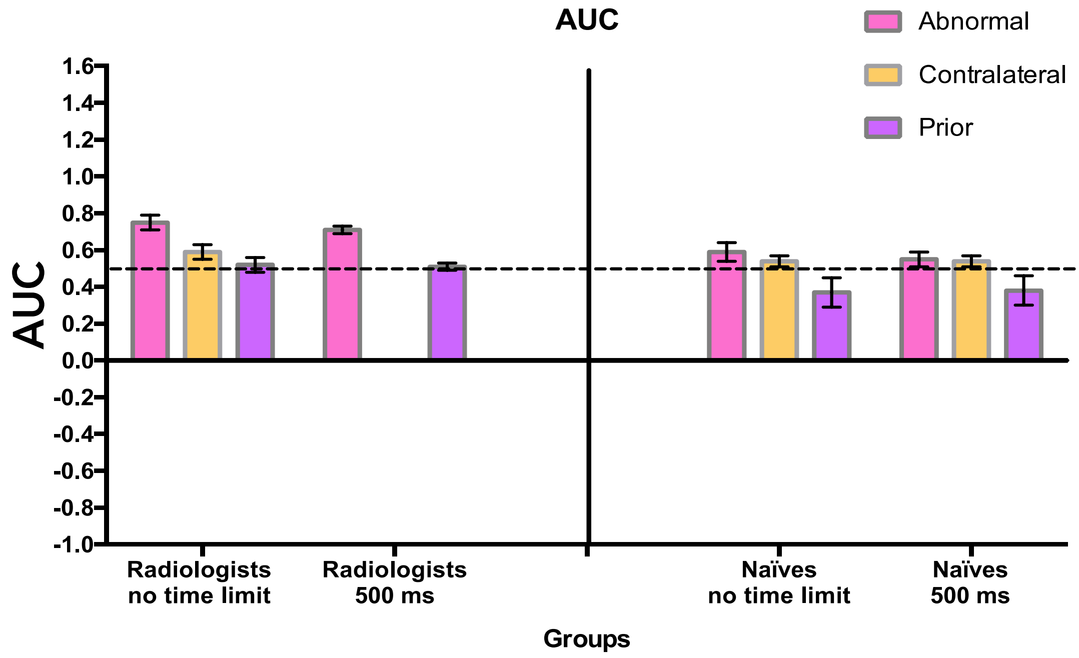


Figure 2.3: *ROC curves for the radiologist groups during no time limit and 500 ms time limit conditions per image type (subtle abnormal, contralateral, priors). Each plot contains individual ROCs (coloured dotted lines) and the group mean ROC (thick black line). The dashed grey diagonal line indicates the line of no discrimination.*

Z-transformed versions of the ROCs (zROCs) produced curved functions. zROCs are straight lines if the underlying signal and noise distributions are normal. The curved zROCs could be taken as evidence that the underlying distributions are not normal; an interesting possibility beyond the scope of the current project.





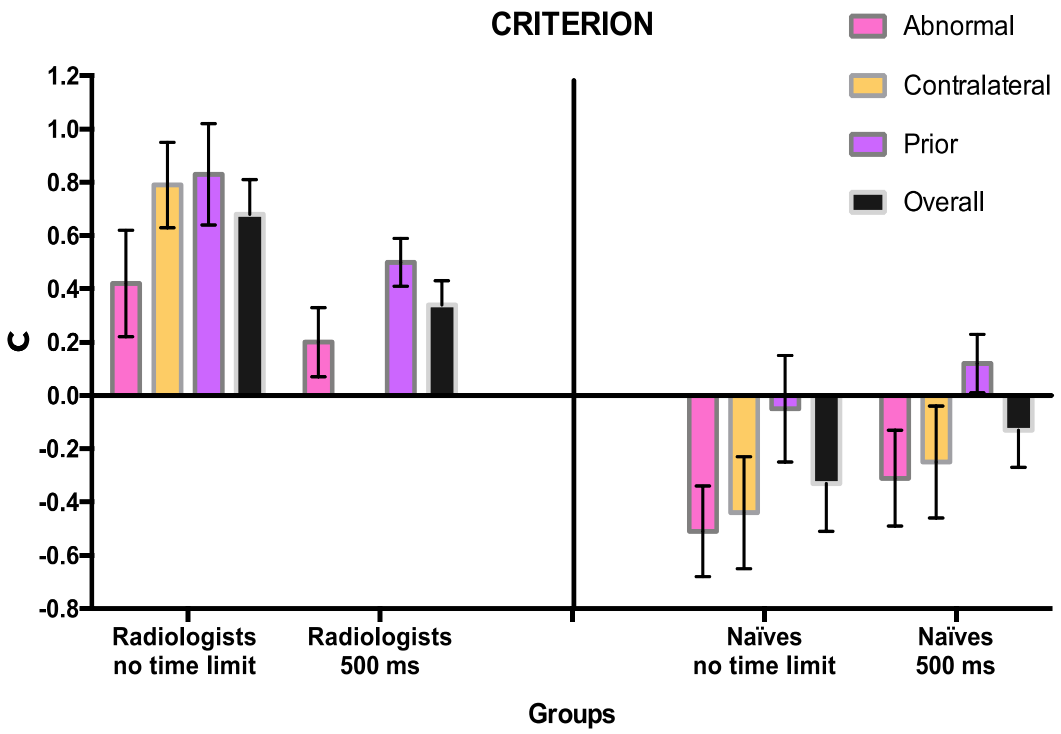


Figure 2.4: *Bar graphs representing the average d', AUC, and criterion (±SEM) per image comparison category (subtle abnormal, contralateral, prior) and over the total image set for the radiologists and naïves under no time limit and 500 ms time limit conditions.*

Effect of time limit on performance in radiologists

To see how time limitations affect performance of mammography experts, 2x2 mixed ANOVAs were conducted on d' and AUC with timing condition (no time limit, 500ms time limit) as a between-group factor and image type (subtle abnormal, priors) as a within-group factor. As stated in the methods, no contralaterals were shown in the time limit condition for the radiologists, so these were not included in this part of the analysis. For d', there was strong evidence for a main effect of image type (F(1,25)=59.409, p=<.001, ηp2=.704, $BF_{inclusion}$=5.87e7 and moderate evidence for a main effect of timing condition (F(1,25)=7.819, p=.010, ηp2=.238, $BF_{inclusion}$=3.828). There was no significant interaction effect (F(1,25)=.312, p=.576, ηp^2=.013, $BF_{inclusion}$=0.727). In the AUC data, there was, again, a large main effect of image type (F(1,25)=110.85, p=<.001, ηp^2=.816, $BF_{inclusion}$=1.241e10), but no statistically significant evidence of a main effect of timing condition (F(1, 25)=1.757, p=.197, ηp^2=.014, $BF_{inclusion}$=.613). There was no evidence for an interaction effect (F(1, 25)=0.440, p=.513, ηp^2=.017, $BF_{inclusion}$=0.392). The $BF_{inclusion}$ for both condition and interaction effect can be classified as anecdotal evidence for H_0.

Our particular interest was in whether more time allowed experts to extract more meaning from the prior images. Post-hoc comparisons showed that unlimited time produced a larger d' (t(25)=2.796, p=.010, BF_{10}=1.942) but not a larger AUC (t(25)=1.325, p=.197, BF_{10}=0.378) on average, and the Bayes Factor for the d’ difference shows only anecdotal evidence. The combination of non-significant effect on AUC and anecdotal Bayes Factor for d’ suggest that this might not be a true difference. Looking at Figure 3, it is clear that performance is above chance in both conditions but that the variability between observers makes it hard to determine if unlimited time improves performance. Certainly, unlimited time does not produce a massive improvement.

Turning to the criterion, there was a main effect of image type ($F(1,25)=52.290$, $p=<.001$, $\eta p^2=.677$, $BF_{inclusion}=322.440$). There was no evidence of main effect of timing condition ($F(1,25)=3.247$, $p=.084$, $\eta p^2=.115$, $BF_{inclusion}=1.331$) or an interaction effect ($F(1, 25)=.405$, $p=.530$, $\eta p^2=.016$, $BF_{inclusion}=0.423$). Criterion was significantly higher for priors than subtle abnormal cases (mean difference=.345, $p=<.001$, $BF_{10,U}=416.754$).

These findings showed some indication that additional time might improve performance of radiologists on detecting future abnormality in the priors, but this effect was inconsistent, as it was observed for d’ but not AUC. Additionally, for d’, the Bayesian statistics suggested only anecdotal evidence, further weakening the evidence. Overall, our results show no clear evidence of an advantage of either time condition.

Effect of time limit on performance in naïves

Overall performance as measured by d’ of the naïve participants was not significantly different from zero, as measured by a one sample t-test for the 500 ms ($t(22)=1.330$, $p=0.196$, $BF_{10}=0.308$) and the no time limit ($t(22)=1.309$, $p=0.204$, $BF_{10}=0.301$) condition. This is in line with previous findings, and suggests that overall, the naïve participants could not detect the gist of abnormality in abnormal, contralateral, and prior images with above-chance accuracy, even without a time limit, emphasizing the necessity for perceptual expertise. More detailed analysis of the performance of naïves is available in appendix A.

Effect of image type and expertise on reaction times

To investigate whether observers spend longer judging certain cases we examined reaction times under no time limit conditions. Radiologists had an average reaction time of 5526ms ± 1884 while naïves had an average reaction time of 4213 ± 942. Radiologists' RTs were higher for each image type (table 1). The difference between groups was significant (independent samples t-test, mean difference=1298ms, $t(34)=2.6$, $p=.014$, $d=0.9$) probably indicating that experts had more to think about when looking at an image.

For naïves, a one-way RM-ANOVA on image type (normal, subtle, contralateral, prior) showed no significant effect of image type ($F(3,66)=1.49$, $p=.226$) on reaction time, which was also supported by the Bayesian RM-ANOVA with a BF_{10} of 0.285 indicating moderate evidence towards this null effect. On the other hand, for radiologists, a one-way RM-ANOVA on image type (normal, subtle, contralateral, prior) showed a significant main effect of image type (F(3,36)=8.80, p<.001), which was also strongly supported by the Bayesian RM-ANOVA with a BF_{10} of 139.55 indicating extreme evidence towards this main effect. Frequentist post-hoc tests with Holm correction for multiple comparisons showed that responses were significantly slower for normal (p=.048) and subtle (p<.001) than prior cases, which was supported by the Bayesian post-hoc tests with moderate evidence for normal and prior ($BF_{10,\,u}$=6.83) and very strong evidence for subtle and prior ($BF_{10,\,u}$=38.33). The frequentist post-hoc tests trended towards faster responses to normal than subtle cases (p=.052), faster responses to contralateral than subtle cases (p=.052), and faster responses to prior than contralateral cases (p=.052). Among these trends, Bayesian post-hoc tests showed strong evidence for a difference between normal and subtle ($BF_{10,\,u}$=17.27), but only anecdotal evidence for subtle and contralateral ($BF_{10,\,u}$=1.74) and contralateral and prior cases ($BF_{10,\,u}$=1.77). The strong Bayes factor for normal and subtle cases suggests that this is a true effect, while there is only anecdotal evidence for the other two trends. Overall, reaction times differed significantly between image types, with faster responses to prior than both subtly abnormal and normal cases, and faster responses to normal than subtly abnormal cases.

Table 2.1: *Average reaction time in milliseconds for naïves (n=23) and radiologists (n=11) during no time limit conditions, per image type (± 95% CI).*

	Normal	Subtle abnormal	Contralateral	Priors	Overall
Naïves	4263 ± 978	4132 ± 929	4085 ± 911	4377 ± 942	4213 ± 942
Radiologists	5537 ± 1661	6162 ± 2261	5501 ± 1715	4846 ± 1735	5526 ± 1884

2.7. Discussion

In previous work, we and our colleagues have found that, with 500 ms of viewing time, expert radiologists can use a global gist of abnormality signal to classify normal from unilateral abnormal mammograms. More strikingly, we found that that this gist of abnormality can be detected in contralateral and prior-abnormal mammograms (Patrick C. Brennan et al., 2018; Evans et al., 2019; Evans et al., 2016). In the present study, we asked if that gist signal would be markedly stronger if experts could scrutinize the image or, alternatively, if the brief exposure was required, with any gist signal being hidden by sustained exposure. In fact, the data did not show either of these effects. The existence of a gist signal was replicated but there were no dramatic effects of exposure duration.

The data from naïve participants continues to show that detection of the gist of abnormality requires expertise. As expected, performance of naïve participants was not significantly different from chance in either the no time limit or the 500 ms condition. The prior images were judged to be more normal than the actual normal images; a result that seems to reflect lower density particular in the prior images we used. This finding fits with the previous reports of at-chance performance of naïves with rapid exposure (Evans, Georgian-Smith, et al., 2013), and also shows that more time does not enable naïves to access an accurate first impression to perform above chance. Thus, radiologists possess an ability that allows them to accurately perceive the gist of abnormality in mammograms, that does not seem to be present in naïve participants, regardless of time constraints.

A central question for this study was whether the gist of abnormality would still be available to expert observers when the stimulus was not flashed but was available until response. It could have been that, with longer exposures, a transient gist signal becomes diluted or cancelled by more sustained processes. Alternatively, it could be that experts could exploit the gist signal more effectively given more time. The data show that experts continue to perform at above chance levels with unlimited time, with some evidence that d’ was higher in the no time limit condition, but since this was not replicated in the AUC data there was no consistent evidence for improvement in performance without time-limited exposure. In thinking about a possible clinical role for gist, this is something of a disappointment. The gist signal for prior images is reliable but weak. The possible use of such a signal as imaging biomarker would be strengthened if conditions could be found that produced a more robust signal.

For the abnormal images, the images that contained visible lesions, our experts seem to have followed our instructions not to scrutinize the images. While this is a difficult instruction to verify, it is certainly the case that our average total reaction time of 5.53 ± 1.88 seconds is markedly lower than any normal interpretation times in the clinic (e.g. 128 to 138 seconds for routine screening examinations of digital mammography (Berns et al., 2006; Kuzmiak et al., 2010)) or in the lab (e.g. average reading time per 2D mammography case was 33 seconds in a screening-like condition (10% prevalence) of an archival set by 3 radiologists (Bernardi et al., 2012). Those cases included multiple images but even so, 5.5 second for one image would be hasty under normal instructions. In a two-decision stage study on bilateral cases, the initial normal/abnormal distinction took 23 seconds on average, followed by an additional 39 seconds to localize any abnormalities in the final decision phase (C. F. Nodine, Mello-Thoms, Kundel, & Weinstein, 2002). Thus, in the current no time-limit condition, radiologists were relatively fast in making their decision, supporting the notion that they were indeed using a first impression rather than a detailed examination to inform their rating.

Response times of radiologists were significantly affected by image type, with faster responses to priors than normal (+704 ms) or subtle abnormal cases (+1323 ms). Additionally, responses to normal cases were faster than subtle abnormal cases (+619 ms). These differences suggest that the presence of a local abnormality increased reaction times. One could speculate that once there was no time limit the experts started looking for a visibly localizable signal of abnormality rather than a global perturbation of the parenchyma. Basing one's decision on detection of a visible local lesion is in line with clinical practice to reduce false alarms, cognisant of low prevalence of breast cancer in screening population. In contrast, the possibility to search for local lesions is not present when the image is flashed for 500 ms, meaning the radiologist must heavily rely on their global gist impression. This might make it easier to focus on information conveyed by global, non-localizable signals of abnormality during the first impression and thus maybe a more optimal approach when aiming to develop a method for early-stage triage to identify at-risk women for more frequent screening. On the other hand, this could also result in missing possibly critical information present in the global parenchymal perturbation absent of a visible lesion. However, as our data showed no consistent changes in either performance or criterion, any changes in rating strategy between the conditions did not significantly affect radiologist ratings in our paradigm. This might be due to the mix of mammograms containing visibly actionable lesions and mammograms without visible abnormalities (contralateral, priors), which could prevent the radiologists from shifting to a strategy aimed at detecting the gist of abnormality in these more ambiguous cases. It might be interesting to repeat the no time limit condition in a new experiment using a test set composed exclusively of normal images and abnormal prior images. Such a set would lack any localizable abnormalities. With such a set, one could, give readers the information that, in this stack of 100 images, 50 came from women who would develop cancer within 3 years. Readers could be asked to sort the images into normal and abnormal categories, taking as much time as they cared to. Readers could be given case-by-case feedback after each response. Perhaps these conditions would produce stronger evidence of sensitivity to the gist of abnormality.

One additional consideration is that rating cases based on either a glimpse or a first impression is not a typical behaviour for radiologists. It is possible that further training with the task for possible triage of cases could improve their performance in gist and/or first impression ratings. For example, they might become more accustomed to suppressing their inclination to perform a detailed examination without a time limit or become more attuned to their first impression in both conditions. Or, if feedback is given, they might be able to further finetune their gist categorisation, although this might require intensive training to affect perceptual processing. These options could be explored in future experiments using training paradigms.

Conclusion

In the present study there was no clear evidence of additional additive benefit to the overall global impression of an image with no time limit exposure without search. Medical experts show the same overall performance detecting abnormalities in mammograms whether they use the global gist signal based on rapid viewing or using their first impression assessment with no time constrained viewing. Medical experts are not more sensitive to the signal of cancer with more time following first impression rather than gist but maintain a conservative criterion for images with no locally visible lesions.

In conclusion, it remains interesting that experts are sensitive to a global signal of abnormality that can be detected in images acquired years before the cancer produces a localized sign in the images. However, this signal remains small and was not meaningfully enhanced by removing the viewing time limit when rating a mixed set of cases in a laboratory setting. Thus, if this signal is to have some clinical utility, it is worth continuing efforts to enhance that signal by for example image enhancement.

2.8. List of Abbreviations

AUC: Area under the curve

BF: Bayes Factor

2.9. Open Practices Statement

The Matlab code used to run the experiment, and the datasets generated and analysed during the current study are available on our OSF repository. dx.doi.org/10.17605/OSF.IO/5NWP8

This data is available under Creative Commons Attribution-NonCommercial-ShareAlike 2.0 UK: England & Wales (CC BY-NC-SA 2.0 UK). All mammograms were selected from the Complex Cognitive Processing lab database of stimuli, which can be shared with other researchers upon request to the last author (K.K. Evans).

Experiments were not preregistered.

2.10. Appendices

Appendix A: Detailed effect of time limit on performance in naïves

The performance of naïve observers was characterized by lower d’ values, and AUC values close to chance (0.5) in both conditions. Following the very low ratings for priors, shown in Figure 2, we performed one-sample t-tests to further investigate this, which showed that naïve observers’ d’ actually falls below 0 and the AUCs is less than 0.5. This is the case for the 500 ms presentation (AUC: t(22)= -2.774, p=0.011, BF_{10}=4.51; d: t(22)=-2.139, p=.044, BF_{10}=1.47) and no time limit conditions (AUC: t(22)= -3.233, p=.004, BF_{10}=11.09; d: t(22)=-3.06, p=.006, BF_{10}=7.85). However, in the 500 ms condition, the Bayes factor for d’ provides only anecdotal evidence towards a significantly negative d’ prime in that condition.

As with the radiologists, the naïve observer data for d', AUC, and criterion were analysed in separate 2x3 repeated measures ANOVA with condition (no time limit, 500 ms time limit) and image type (normal-abnormal, normal-contralateral, normal-priors) as factors. There was a main effect of image type for both d’ (F(1.26, 27.78)=26.18, p=<.001, ηp^2=.543; $BF_{inclusion}$ across matched models=3.75e12) and AUC (F(1.23,43.01)=27.808, p<.001, ηp^2=.558, $BF_{inclusion}$ across matched models=3.75e12), but no evidence of a main effect of timing condition. Nor were there significant interaction effects. In fact, the $BF_{inclusion}$ across matched models for condition was 0.187 (d’) and 0.195 (AUC), both providing moderate evidence for the null hypothesis of no main effect of timing condition.

For criterion, there was evidence of a main effect of image type (F(1.26, 27.78)=26.18, p=<.001, ηp^2=.543, $BF_{inclusion}$ across matched models=2.97e7), and a main effect of timing condition (F(1.16, 22.00)=4.67 p=.042, ηp^2=.175, $BF_{inclusion}$ across matched models=30.55). A pairwise comparison showed that criterion was higher (more conservative) in the 500 ms time limit conditions (mean difference = .184, p=.042). Pairwise comparisons of image types showed that criterion was significantly higher when rating priors than abnormal (mean difference=.45, p=<.001) and contralaterals (mean difference=.38, p=<.001). This analysis suggests that removal of time limit had no effect on performance in naïves aside from making their ratings more conservative.

2.11 Commentary: AUC vs d’

While both AUC and d’ are measures of the ability of an observer to distinguish between signal and noise, they can yield slightly different patterns of results depending on the signal and noise distributions. The calculation for d’ assumes that signal and noise are normally distributed with equal variance. When the AUC is calculated using the trapezoid method, as it was done in this study, it is based on the ranks of the hits and false alarms, and does not rely on an assumption of normal distributions. Thus, AUC is a more reliable measure of the overall discriminability of normal/abnormal. In contrast, d' represents the discriminability at a specific threshold, which is more informative for clinical applications where a threshold must be established to identify at-risk cases, but the d’ can be influenced by the distribution of signal and noise. Therefore, it is useful to analyse both measures to obtain a complete picture of performance.

In this study, the lack of timing condition effect on the AUC measure suggests that viewing time did not significantly influence overall performance, especially since the Bayesian statistics showed the evidence for improvement in d’ with no time limit was only anecdotal (BF_{10}=1.942).

Chapter 3: Early signs of cancer present in the fine detail of mammograms.

Chapter 3 was published as Raat E. M., Evans K. K. (2023) Early signs of cancer present in the fine detail of mammograms. PLoS ONE 18(4): e0282872. doi.org/10.1371/journal.pone.0282872. This version of the article has been accepted for publication after peer review but is not the Version of Record and does not reflect post-acceptance improvements. It also has edited headers and figure numbers to fit with the thesis format. The Version of Record is available online at doi.org/10.1371/journal.pone.0282872

3.1. Abstract

The gist of abnormality can be rapidly extracted by medical experts from global information in medical images, such as mammograms, to identify abnormal mammograms with above-chance accuracy - even before any abnormalities are localizable. The current study evaluated the effect of different high-pass filters on expert radiologists’ performance in detecting the gist of abnormality in mammograms, especially those acquired prior to any visibly actionable lesions. Thirty-four expert radiologists viewed unaltered and high-pass filtered versions of normal and abnormal mammograms. Abnormal mammograms consisted of obvious abnormalities, subtle abnormalities, and currently normal mammograms from women who would go to develop cancer in 2-3 years. Four levels of high-pass filtering were tested (0.5, 1, 1.5, and 2 cycles per degree (cpd) after brightening and contrast normalizing to the unfiltered mammograms. Overall performance for 0.5 and 1.5 did not change compared to unfiltered but was reduced for 1 and 2 cpd. Critically, filtering that eliminated frequencies below 0.5 and 1.5 cpd significantly boosted performance on mammograms acquired years prior appearance of localizable abnormalities. Filtering at 0.5 did not change the radiologist’s decision criteria compared to unfiltered

mammograms whereas other filters resulted in more conservative ratings. The findings bring us closer to identifying the characteristics of the gist of the abnormal that affords radiologists detection of the earliest signs of cancer. A 0.5 cpd high-pass filter significantly boosts subtle, global signals of future cancerous abnormalities, potentially providing an image enhancement strategy for rapid assessment of impending cancer risk.

3.2. List of Abbreviations

Cpd: cycles per degree

ROC: receiver operator curve

AUC: area under the curve

AIC: Akaike Information Criterion

LSF: Low spatial frequencies

HSF: high spatial frequencies

3.3. Introduction

Breast cancer is (one of) the most prevalent and deadly cancers in women world-wide, according to global data from 1990 to 2015 (Fitzmaurice et al., 2017) and 2020 GLOBOCAN cancer statistics (Ferlay et al., 2021). As with most cancers, early detection is vital, as it allows for treatment before the disease progresses and improves clinical outcomes (Coleman, 2017). Currently, the most commonly used methods of screening and early detection are clinical breast exams and digital mammography, as they are effective and cost-efficient (Coleman, 2017) and have been estimated to reduce mortality by 30% to 50% (Tabár et al., 2014). Digital mammography is especially for early detection, as it allows detection of small, pre-clinical tumours of <15mm that are not detectable with a clinical breast exam (Tabár et al., 2014),. However, 20-30% of cancers are still estimated to be missed during screening in North America (Bird, Wallace, & Yankaskas, 1992; Majid, de Paredes, Doherty, Sharma, & Salvador, 2003).

Further reducing breast cancer mortality through screening could be achieved by increasing screening frequency. However, increasing screening frequency across the entire population is not cost-effective, and risks increasing false positives (Mandelblatt et al., 2009) or even over-diagnosis of benign breast conditions, which has been associated with unnecessary cost (Chubak, Boudreau, Fishman, & Elmore, 2010) and negative mental health effects (Jatoi, Zhu, Shah, & Lawrence, 2006; Sandin, Chorot, Valiente, Lostao, & Santed, 2002).

Instead, women at an increased risk should be offered more frequent screening. Currently, at-risk women are often identified through familial history of breast cancer, or genetic markers, such as

BRCA1 or BRCA2 mutations, which cause approximately 60% of hereditary breast cancer (Pruthi, Gostout, & Lindor, 2010). However, gene screening is costly and BRCA1 or BRCA2 mutations cause only 5% of breast cancer, limiting applicability to the general population. An alternative, more universal approach would be to identify at-risk women based on perceptual features in their existing mammograms. This method relies on the robust observation that experienced radiologists can capture both current and future cancer risk in the blink of an eye through extraction of the gist of abnormality.

This gist of abnormality is extracted through a process that rapidly and non-selectively extracts global structure and statistical regularities from our visual environment (Oliva, 2005; Oliva & Torralba, 2006). In normal observers, this allows them to categorize a flashed scene (30 ms) as a beach or a forest with high accuracy (Greene & Oliva, 2009; Joubert et al., 2009). In addition, medical experts are extract the gist of medical images, allowing them to distinguish normal from abnormal cases after 100 to 500 milliseconds of viewing time for chest radiographs (Kundel & Nodine, 1975), cytological images from PAP smears (Evans, Georgian-Smith, et al., 2013), and mammograms (Evans, Georgian-Smith, et al., 2013; Evans et al., 2016). Importantly, mammograms of women taken 3 years prior to their eventual diagnosis (priors), that did not contain detectable cancer even when viewed retrospectively, are scored as significantly more abnormal than mammograms of women that did not go on to develop cancer in the near future (Patrick C. Brennan et al., 2018; Evans et al., 2019). Thus, the gist of abnormality is a robust signal that can rapidly be extracted from mammograms.

Thus, a high gist of abnormality score could be a promising risk factor to flag mammograms for a secondary opinion (current risk) or to recommend women for more frequent scanning (future risk). Advantages of the gist signal are that it can be extracted from already existing mammograms, and it is already visible in cases up to 3 years prior to cancer onset, without visibly actionable lesions. Unfortunately, the signal strength in priors is relatively weak with an observed d’ of 0.22 and an Area Under the Curve (AUC) of 0.54-0.6 for priors without visible abnormalities (Patrick C. Brennan et al., 2018; Evans et al., 2019). Thus, methods to strengthen the gist of abnormality signal, especially in priors, are needed to make it more clinically viable.

Spatial frequency filtering might allow a way to isolate and enhance the perceptual features that comprise the gist signals in mammograms. Visual information can be summarized as spatial frequencies in various orientations. Low spatial frequencies (LSF) provide coarse information spread across a large area, whereas high spatial frequencies (HSF) provide finer details of for example edges and contours. Together, LSF and HSF provide important information about the texture and structural regularities in our visual environment. But it is possible that the gist of abnormality is stronger in specific frequency bands, or that it is masked by other frequency bands that make it harder to perceive. Interestingly, filtering out HSF strongly reduced accuracy of rating normal vs abnormal mammograms from a d’ of 1.06 with full spectrum mammograms to only 0.26, while filtering out LSF resulted in a relatively high d’ of 0.96 (Evans et al., 2016). Thus, gist of abnormality seems to be preferentially contained in HSF, although there was still a small reduction in performance.

Conflicting findings have been reported on the effects of spatial frequency filtering on general gist extraction. Merged spatial frequencies from two scenes were most frequently perceived as the LSF scene with 30 ms view time, but with 150 ms HSF dominated (Schyns & Oliva, 1994), suggesting an early importance for LSF. However, recent evidence points to the importance of HSF for scene gist when taking contrast normalization into account. Natural images contain more LSF than HSF contrast energy, following an inverse power law (Perfetto, Wilder, & Walther, 2020). This means that HSF-only images have lower overall visibility. After contrast normalization human observers showed equal performance on gist categorization of LSF and HSF scene images (Perfetto et al., 2020).

Since Evans et al. (2016) did not contrast normalize the mammograms, the reduction in performance for HSF compared to full spectrum mammograms might have be caused by a reduction in contrast energy. Additionally, HSF-retaining filters might differentially affect gist signals in different conspicuities. The current study aimed to investigate the effects of five levels of high-pass spatial frequency filtering on the gist of abnormality in mammograms with three different conspicuities when applying contrast normalization. Contrast normalization was combined with a brightness increase to ensure that the higher spatial frequencies were bright enough to be perceived. Our results show that some high-pass filters preserved overall performance, and more importantly, enhanced performance in mammograms taken prior to development of visible, actionable abnormalities. These findings provide a promising avenue of using high-pass filtering image enhancements to improve gist of abnormality risk factors.

3.4. Methods

Participants

A total of 34 participants took part in this experiment, which was conducted in two versions, an in-person experiment and an online experiment. The online version was set up to avoid in-person contact during the COVID-19 pandemic. All participants were radiologists with experience reading mammograms in a clinical setting, which was defined as having read at least 1000 scans in the last year.

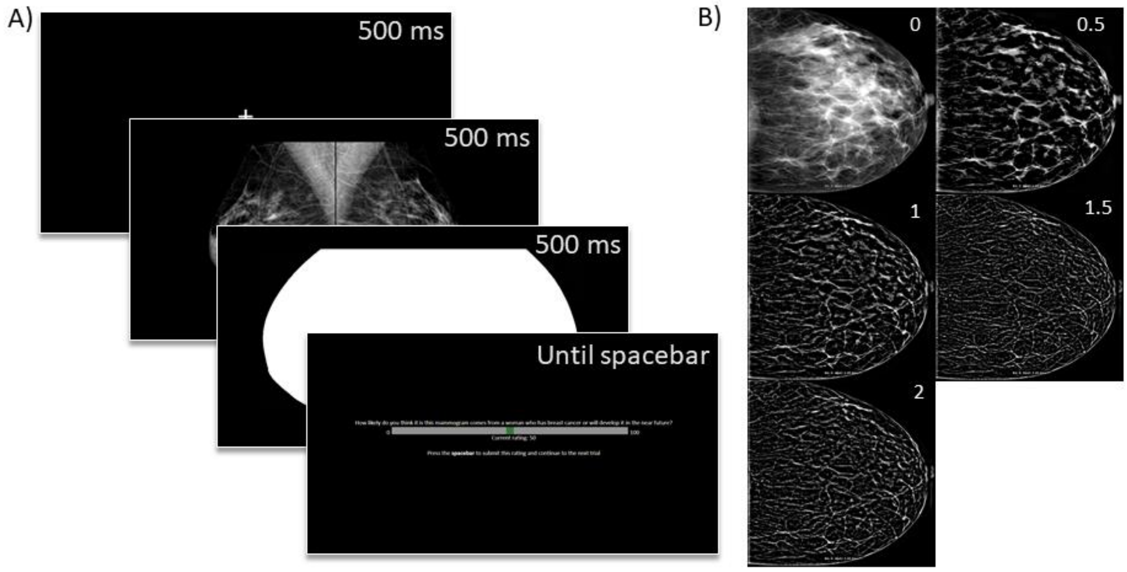
Sixteen participants took part in the in-person version of the experiment (9 female, 32 to 64 years old, mean 50.7+-10.8). They read on average 5056 scans (std 3707, range 1000 to 12000) over the last year, average 22 years in practice (std 11.6years, range 2 to 37), and on average spend 59% of their time diagnosing mammograms (std 34%, range 10 to 100%) in their work. Eighteen participants took part in the online version of the experiment (13 female, 33 to 67 years old, mean 46.9 +- 10.1). They read on average 5694 scans (std 2996, range 1000 to 10000) over the last year, average 14 years in practice (std 10.6 years, range 2 to 37), and on average spend 70% of their time diagnosing mammograms (std 27.1%, range 25 to 100%) in their work. The 5 radiologists at the lower end of cases read in the last year (<2000) had been practicing for 7, 18, 19, 30, and 37 years respectively, indicating extensive experience.

Participants were recruited in-person during the Radiological Society of North America (RSNA) 2018 conference, and online over a period from 2020 to 2022, with recruitment emails sent to individual contacts, collaborating hospitals in the United Kingdom, and newsletters of various radiology profession groups in the UK and the Netherlands. The sample size of the radiologist groups was dictated by the availability of participants. This study was approved by the Psychology Departmental Ethics Committee of the University of York (ID 307), and all participants gave informed consent either written on paper (in-person) or digitally by clicking a button “I understand and agree” after reading the consent form (online).

Stimuli and apparatus

The stimuli used in this experiment were de-identified bilateral mammograms sourced from the Complex Cognitive Processing Lab database of stimuli, in mediolateral oblique (MLO) or craniocaudal (CC) view. Four mammogram categories were used: normal mammograms of healthy women (normal), mammograms with obvious cancerous abnormalities (obvious), mammograms with subtle cancerous abnormalities (subtle), and mammograms without visibly actionable lesions taken three years prior to sign of abnormality (priors). Normal mammograms were defined as cases without abnormalities, of which the woman did not develop cancer in the next three years. Obvious and subtle mammograms were selected from a set of mammograms containing an abnormality, which were conspicuity-rated by an experienced mammogram-reading radiologist based on the visibility of the abnormality (obvious, subtle). Priors were defined as mammograms without any visible cancerous abnormalities of women who were then found to have developed cancer within the next three years retrospectively.

MATLAB was used to create the spatially filtered stimuli. Stimuli were filtered using a high-pass 2nd order Butterworth filter with four different cut-off points. Filtered stimuli were brightened using a custom setting multiplying any pixel values above 10 (out of a 0 to 255 scale) by 3.5. Next, the filtered images were contrast normalized with the SHINE Toolbox for each group of filtered images together with the unfiltered images. Contrast normalization removes effects from overall differences in brightness between the filter groups. Four groups of spatially filtered images were created, namely 0.5, 1, 1.5, and 2 cycles per degree (cpd), examples of which can be seen in Fig 1B.



How likely do you think it is this mammogram comes from a woman who has breast cancer or will develop it in the near future?

Current rating: 50

Press the **spacebar** to submit this rating and continue to the next trial

Figure 3.1: *Procedure and stimuli used in the experiment. (A) Example visualization of the different screens in one trial, showing the fixation cross, mammogram case, mask, and rating screen. (B) Examples for the unaltered (0) and high-pass filtered versions (0.5, 1, 1.5, and 2 cpd) of a unilateral mammogram*

The in-person experiment was run using MATLAB, utilizing the Psychophysics Toolbox 3 extensions (Brainard, 1997; Kleiner et al., 2007). The online experiment was run on a custom web page. Participants were instructed to sit at a comfortable viewing distance of approximately 57 cm. In-person, stimuli were presented on a 17’ inch Dell colour display (1920 x 1200 pixels) with an 85 Hz refresh rate. For the online experiment, participants performed the experiment on their own laptop or PC. For the online experiment, a screen calibration method based on the work by Q. Li, Joo, Yeatman, and Reinecke (2020) was used to ensure the stimuli were presented at 10 degrees of visual angle in height.

Procedure

The experiment consisted of 3 practice trials followed by 3 blocks of test trials. In the practice trials, participants were familiarized with the display and rating screen, and were given feedback on the stimulus (normal or abnormal) after they confirmed their rating. In the test trials, no feedback was given. Each trial started with a fixation cross in the centre of the screen (500 ms), followed by the bilateral mammogram being shown for 500 ms. Then, a mask consisting of the solid white shape of the breast tissue was shown for 500 ms. Next, the rating screen appeared, on which moving the mouse changed the rating on a scale from 0 to a 100. Pressing the spacebar would confirm the current rating, after which the next trial automatically started (see Fig 1A).

Participants were asked to rate how certain they were that the image came from a woman with breast cancer, or who will develop it in the near future. Participants were asked to adopt a liberal call back criterion, while being as accurate as possible. There was no time constraint for the rating in either condition, but participants were asked to report their first impression. During the in-person experiment, ratings were made on a scale from 0 (abnormal) to 100 (normal), while the online experiment used a scale from 0 (normal) to 100 (abnormal), due to a difference in coding. This is not expected to be any hindrance in comparing the two experiments, as the rating scale was clearly labelled in the instructions and on each rating screen, and 3 practice trials were available.

As previously stated, each participant completed three blocks of test trials. The same mammograms were used in each test block, to allow for direct comparison of performance. Each test block consisted of 120 trials: 60 normal, 20 obvious abnormal, 20 subtle abnormal, and 20 prior abnormal, in randomized order. One of the blocks always showed unaltered mammograms (F0) to ensure a baseline of performance, and the two other blocks showed two out of the four possible filter groups. Selected blocks and their order were randomized, although the switch from in-person to online measurements caused a lower number of participants for the F1 filter and the F1.5 filters than the F0.5 and the F2 filter. In total, all 34 participants rated F0, 21 participants rated F0.5, 15 participants rated F1, 13 participants rated F1.5, and 19 participants rated F2.

Data analysis

To analyse our data, a signal detection theory framework was used to calculate performance measures, as previously described in an earlier publication (Raat, Farr, Wolfe, & Evans, 2021): "Given a rating, a mammogram was considered to be classified as either "abnormal" or "normal", depending on whether the rating is higher or lower than some threshold. That classification was then compared to the ground truth. Signal detection measures were used to separately assess performance and response biases of the observer. Performance was represented by the D' measure (D' = z(true positive rate) – z(false positive rate)), where z denotes the inverse normal or z-transformation of the rates). In cognitive literature, d' is referred to as "sensitivity". However, "sensitivity" refers to the "true positive" or "hit" rate in the medical literature. We will refrain from using the term in order to avoid confusion. Response bias was measured by the criterion value, C (C = (z(true positive rate) + z(false positive rate))/-2). A negative criterion means that the observer was more likely to label the item as abnormal while a positive criterion means that observer was more likely to label the item as normal.

Receiver operating characteristic curves (ROC) were constructed by repeating this division of trials into proportions of true positive (hits) and false positive (false alarms) using different normal/abnormal rating cut-offs (here, 1 to 99). The area under the curve (AUC) of an ROC, ranging from 0.0 to 1.0, represents the probability that a randomly chosen abnormal case will be rated higher than a randomly chosen normal case (Hanley & McNeil, 1982). Chance performance yields an AUC of 0.5. Higher AUCs indicate better performance in detecting the signal of cancerous abnormalities."

Additionally, a technique for averaging ROCs from multi-reader, multi-case datasets was used to calculate an average ROC for visualization purposes (W. Chen & Samuelson, 2014). D' and criterion were derived using a rating cut-off of 50, as this is the middle point of the rating scale. AUCs were calculated across the entire rating scale and were calculated using the *sklearn.metrics auc* function in Python. These performance measures were calculated per participant for each of the filter conditions and mammogram category (obvious, subtle, and prior) combinations. Pre-processing into signal detection measures was performed in Python 3 using the following packages: json, scipy.stats, numpy, glob, sklearn.metrics auc, and csv. Next, statistical analysis was performed using SPSS 28.0.0.0 (190) for the univariate analysis of variance. For the primary analysis using linear mixed models, we used R version 4.1.3 in RStudio, and the following packages: tidyverse, lme4, sjPlot, rstatix, ggpubr, and emmeans. Additionally, boxplot figures were created using ggplot's geom_boxplot function. These boxplots follow the standard arrangement, except for the whiskers, which contain 1.58 times the inter-quartile range, which is approximately equivalent to the 95% confidence interval of the data (McGill, Tukey, & Larsen, 1978).

Firstly, univariate analysis of variance was performed to determine if there was any between-subjects difference in performance between the in-person and online groups of participants, using group as fixed factor, adding number of cases read as a covariate as previous research has shown a clear positive correlation between cases read and gist performance (Evans et al., 2019). As no main effect of group was found, the two groups could be merged into one dataset.

The primary goal of this study was to investigate the effects of each high-pass filter on performance per image type relative to the unfiltered condition, for which a linear mixed model was used. The model was run separately for D', criterion, and AUC, each with the factors Category (3 levels: Obvious, Subtle, Prior), and Frequency (5 levels: F0, F0.5, F1, F1.5, and F2), an Interaction factor between Category and Frequency, and a random intercept factor for participant ID to model individual differences. Akaike Information Criterion (AIC) (Akaike, 1974) was used to estimate the goodness-of-fit including a penalty for the number of parameters included in the model, where a smaller AIC represents a better fit.

To investigate whether the category, frequency, and interaction factor contributed significantly to the fit of the mixed model, the full model was compared to a trimmed model in which one of these factors was removed. This was analysed using a log likelihood ratio test with the analysis of variance (ANOVA) function in R. If the full model was significantly better than the trimmed model, this provided evidence that this factor contributes significantly. For each factor that contributed significantly, post-hoc comparisons of the model estimates were used to investigate which specific levels of the factors differed from each other. These comparisons used Tukey corrections for multiple comparisons and Kenward-roger's degrees-of-freedom method.

3.5. Results

Overall performance

Overall performance was above chance, replicating previous findings: Average D' was above 0 and the AUC was above 0.5. Criterion values above 0 show that participants were biased towards conservative ratings. Estimated means from mixed models illustrate how these estimates follow the same patterns as the real data (Table 1). Performance was above chance for most participants across filter conditions for obvious and subtle abnormalities, shown by individual ROC curves above the chance line (Fig 2). However, for priors, performance was markedly lower or at chance for some participants in some filter conditions. Overall, participants could extract the gist of abnormality across all filter conditions but regularly struggled with prior cases, which will be further explored in the mixed models.

Table 3.1: Group average and mixed model estimated mean of D', criterion, and AUC for unfiltered mammograms and each high-pass filter frequency.

FREQUENCY	D'		AUC		CRITERION	
	Average	Estimated	Average	Estimated	Average	Estimated
F0	0.685	0.715	0.645	0.640	0.193	0.184
F0.5	0.937	0.897	0.657	0.665	0.469	0.27

F1	0.390	0.297	0.557	0.551	0.538	0.514
F1.5	0.666	0.708	0.611	0.617	0.790	0.879
F2	0.277	0.318	0.562	0.542	0.301	0.664

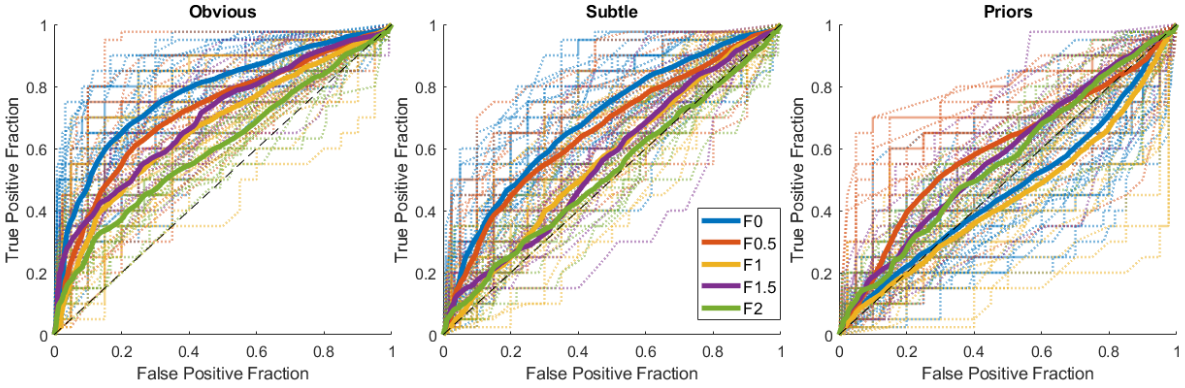


Figure 3.2: *ROC curves per image category. Average and individual (dotted) ROCs per frequency condition (0, 0.5, 1, 1.5, and 2 cpd) for each abnormal mammogram category (obvious, subtle, and prior). The black dashed line represents chance levels, with anything above it being above chance.*

Univariate analysis of variance showed no significant effect of group (in-person/online) on D' for unfiltered mammograms when accounting for number of cases read in the previous year (covariate) (corrected model F(2,31)=2.198, p=.128). This supports the decision to combine the data from the two groups for the main analyses.

Factors influencing D' performance measure

For D', linear mixed model analysis showed evidence for significant contributions of Category, Frequency, and an Interaction (intercept: 1.264, random effect of ID:0.071, AIC: 406.62). An ANOVA comparing log-likelihoods of the full model to one without the category factor showed a significant contribution of category to the model fit ($\chi^2(2)=127.14$, p=<.001). Similarly, frequency contributed significantly to the model fit ($\chi^2(4)=43.514$, p=<.001), as did the interaction factor ($\chi^2(8)=51.655$, p=<.001).

Pair-wise comparisons were performed for frequency (Fig 3A), and mammogram category (Fig 3B). Based on these comparisons, specific interaction effects were reviewed, comparing the unaltered mammograms to the 0.5 and 1.5 cpd high-pass filters that showed no significant difference in overall D'. For priors, D' was significantly higher for F0.5 (estimated difference=0.646, t(263)=5.566, p=<.001) and F1.5 (estimated difference=0.499, t(268)=3.443, p=.006) than unfiltered (F0) mammograms. Meanwhile, there was no significant difference in D' between F0 and F0.5 for obvious (estimated difference=0.091, t(264)=0.781, p=.936) or subtle (estimated difference=0.011, t(263)=0.098, p=1.000) mammograms. For F0 versus F1.5, there was no difference in D' for obvious mammograms (estimated difference=0.068, t(268)=10.467, p=.990), but F1.5 reduced D' for subtle mammograms (estimated difference=0.453, t(268)= 3.128, p=.017). The same pattern of results was observed for AUC (appendix A).

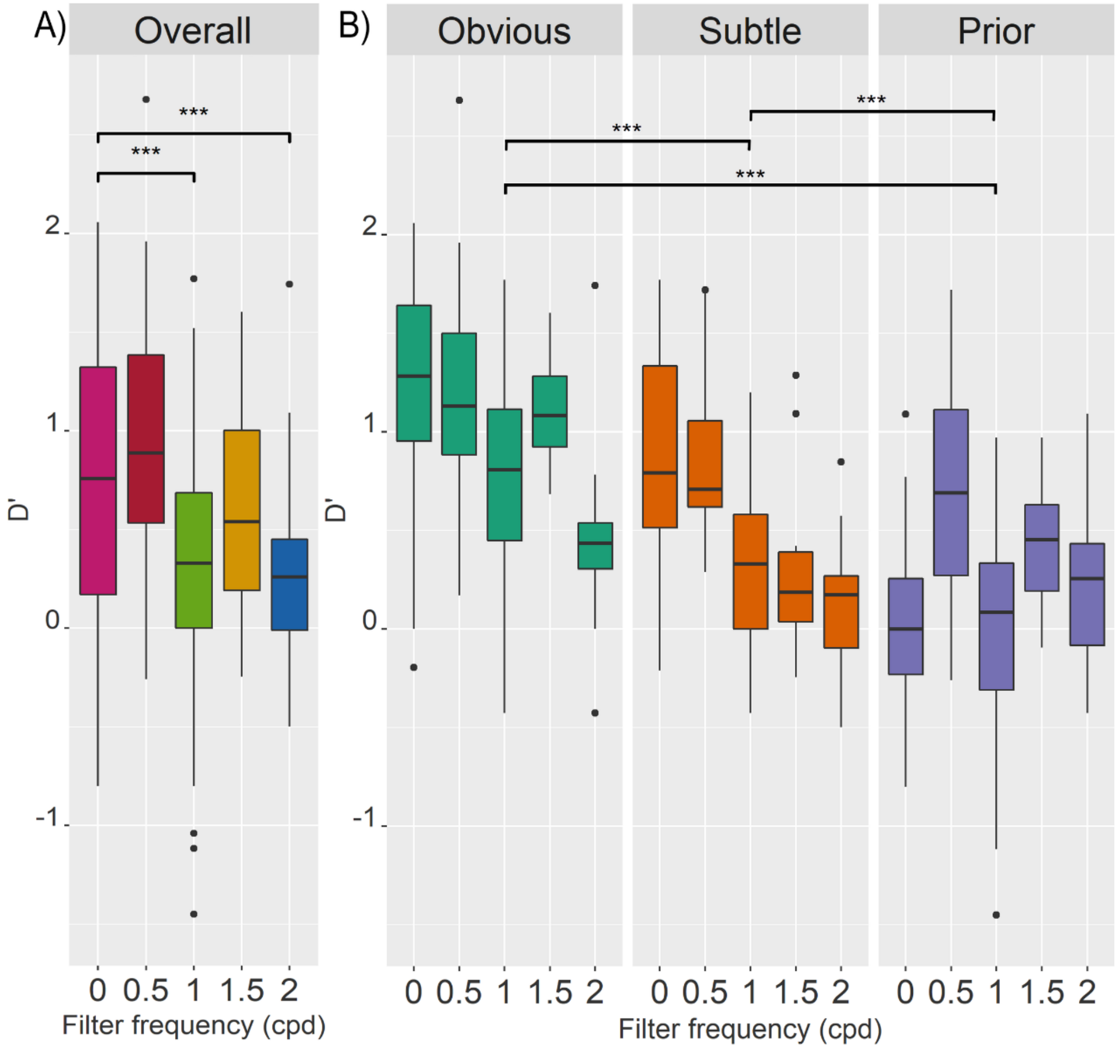


Figure 3.3: *Boxplots of D' across conditions. Each boxplot shows the median as the line within the coloured box containing the 25th and 75th percentiles, with extending whiskers containing the 95% CI, with any outliers plotted as dots. Significance of pairwise comparisons is indicated in the figure with *=p<.05, **=p<.01,***=p<.001. (A) Boxplots showing D' across frequency conditions (0, 0.5, 1, 1.5, and 2 cpd). Pairwise comparisons of frequency showed that D' was significantly higher for F0 than F1 (estimated difference=0.482, t(271)=6.080, p=<.0001) and F2 (estimated difference=0.397, t(282)=4.502, p=<.0001), but did not differ significantly from F0.5 (estimated*

difference=-0.181, t(273)=-2.637, p=.067), and F1.5 (estimated difference=0.007, t(282)=0.083, p=1.000) –and even trended towards a higher D’ in F0.5. (B) Boxplots showing D’ for each mammogram category (obvious, subtle, and prior) and frequency, to illustrate mammogram category and interaction effects. Pairwise comparisons of mammogram category showed that D’ was significantly higher for obvious than subtle (estimated difference=0.462, t(258)=7.278, p=<.0001) and prior mammograms (estimated difference=0.683, t(258)=10.763, p=<.0001), and higher for subtle than prior mammograms (estimated difference=0.221, t(258)=3.485, p=<.001).

Factors influencing the bias in rating measure

For criterion, linear mixed model analysis showed evidence of significant contributions of Category, Frequency, and an Interaction (intercept:-0.108, random effect of ID: 0.323, AIC:356.35). An ANOVA comparing log-likelihoods of the full model to one without category showed a significant contribution of category to model fit ($\chi^2(2)=48.458$, p=<.001). Similarly, frequency ($\chi^2(4)=53.488$, p=<.001) and the interaction effect ($\chi^2(8)=16.563$, p=.035) contributed significantly to the model fit. Pairwise comparisons of main effects can be observed in Fig 4, showing that participants became more conservative for all filter conditions except F0.5.

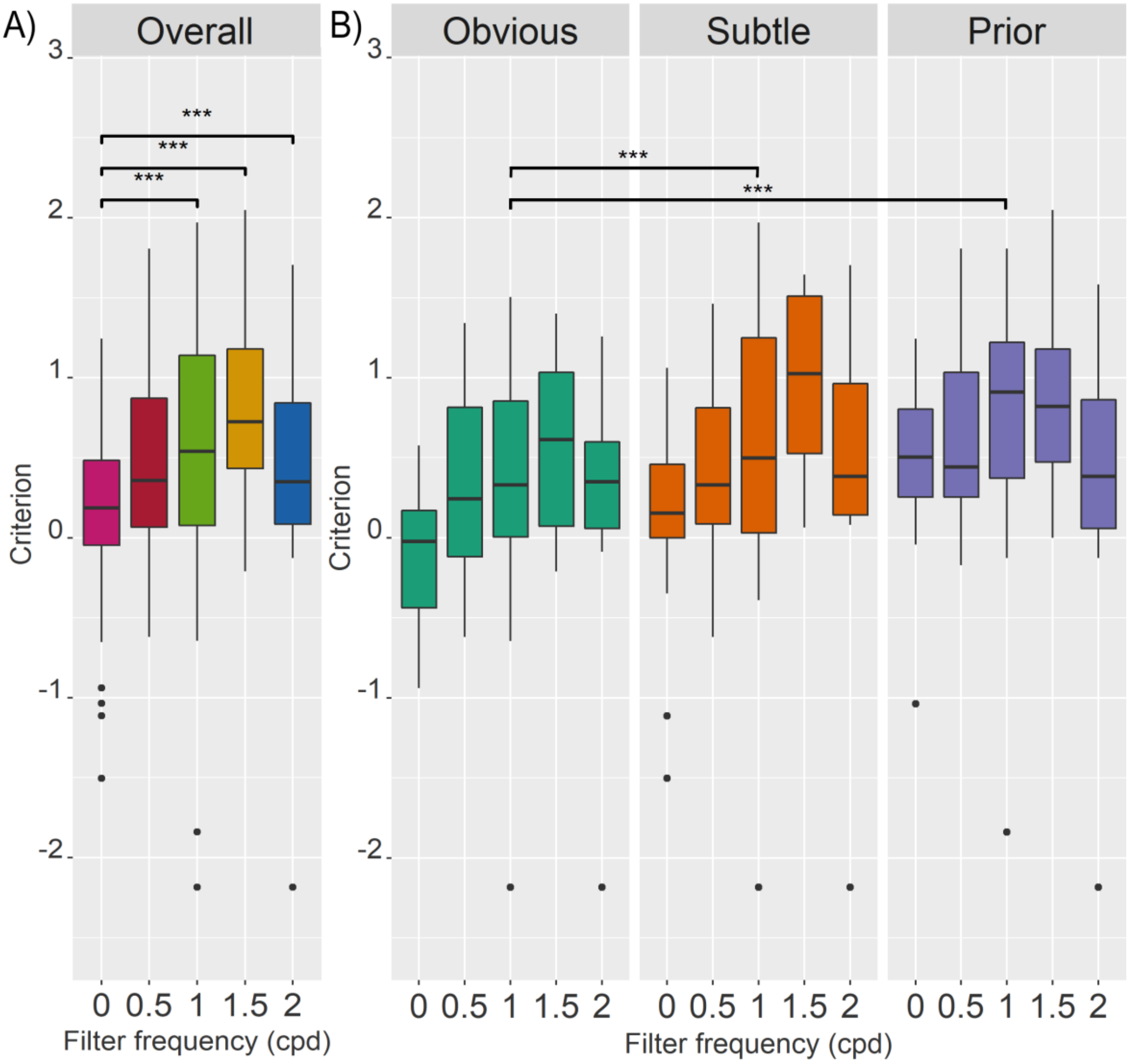


Figure 3.4: *Boxplots of criterion across conditions. Each boxplot shows the median as the line within the coloured box containing the 25th and 75th percentiles, with extending whiskers containing the 95% CI, with any outliers plotted as dots. Significance of pairwise comparisons of main effects is indicated in the figure with *=p<.05, **=p<.01,***=p<.001. (A) Criterion across frequency conditions (0, 0.5, 1, 1.5, and 2 cpd). Pairwise comparisons of frequency showed that criterion was significantly higher for F0 than F1 (estimated difference=-0.3295, t(261)=-5.640, p=<.0001), F1.5 (estimated difference=-0.695, t(264)=-9.176, p<.0001), and F2 (estimated difference=-0.480, t(264)=-6.329, p=<.0001), but did not differ significantly from F0.5 (estimated difference=-0.086, t(261)=-1.477, p=.579). (B) Criterion for each mammogram category (obvious, subtle, and prior) and frequency, to illustrate mammogram category and interaction effects. Pairwise comparisons of mammogram category showed that criterion was significantly lower (less conservative) for obvious than subtle (estimated difference=-0.232, t(258)=-4.345, p=<.0001) and prior mammograms (estimated difference=-.347, t(258)=-6.517, p=<.0001), but did not differ significantly between subtle and prior mammograms (estimated difference=-.116, t(258)=-2.172, p=.078).*

3.6. Discussion

D’ and AUC mixed model findings demonstrate that F0.5 and F1.5 high-pass filters significantly increased gist extraction performance in mammograms acquired years prior to onset on any visible cancerous lesions: D’ was boosted by 0.646 for F0.5 and by 0.499 for F1.5 respectively, a considerable increase. Additionally, 0.5 cpd high-pass filters did not impact radiologists’ performance on obvious or subtle mammograms. This strongly suggests that removing the lowest frequencies in mammograms can enhance the gist of abnormality for current presence or future risk of cancer in cases that do not yet show any visibly actionable signs of cancer, while retaining the signal of current abnormalities.

Radiologists rated mammograms that maintained only frequencies over 1, 1.5 and 2 cpd more conservatively compared to those with frequencies above 0.5 cpd or those with full spectrum. Thus, filtering out spatial frequencies below 0.5 cpd would be the most suitable, as it did not significantly affect observer’s decision criterion, retained performance for obvious and subtle mammograms, and enhanced it for priors. Gist ratings for these high-pass filtered mammograms could be used to flag missed current cancers for a second opinion and for enhanced screening when no abnormalities are found.

Out of the tested filter conditions, two (F0.5 and F1.5) showed retained overall performance and increased performance on priors. However, the other two filter conditions (F1 and F2) showed an overall drop in performance without increasing performance for any sub-types. This pattern could be explained by different effects influencing performance. Firstly, frequencies below 0.5 cpd might mask gist signals, especially in priors, resulting in an increase in performance when a F0.5 filter is applied, perhaps because this removes widespread ‘blur” from breast density. While breast density can be a risk factor for breast cancer, previous research found no correlation between BIRAD density and gist of abnormality ratings (Evans, Birdwell, et al., 2013; Evans et al., 2019; Evans, Georgian-Smith, et al., 2013; Evans et al., 2016). Secondly, intermediate frequencies between F0.5 and F1 might include some important aspects of the gist signal, causing a significant drop in performance when filtering below F1. Thirdly, increased performance on priors with a slight decrease for subtle abnormalities when removing signal between F1 and F1.5 suggests that this frequency band contain some gist signal, but also contributes noise that might obscure global signals of (future) cancer. Lastly, reduced performance when spatial frequencies below 2 cpd are removed from mammograms points to the importance of F1.5 – F2 cpd for the gist signal. Together, these findings suggest the gist of abnormality is contained mainly in 0.5 to 1 cpd and 1.5 to 2 cpd spatial frequencies, with a mix of signal and noise in 1 to 1.5 cpd. Further research would be needed to test these predictions in detail.

The combined effect of high-pass filtering and contrast normalization in increasing the performance of radiologists matches previous findings in both behavioural and neuroimaging work on spatial frequency. Our results match the previous observation that low-pass filtering strongly reduced gist of abnormality performance, while high-pass filtering without contrast normalization had a much less pronounced effect (Evans et al., 2016). Similarly, in scenes gist performance on HSF scenes was reduced without contrast normalization, but contrast normalization equalized performance between LSF and HSF scene images (Perfetto et al., 2020). Our findings match this retention of overall performance with HSF with contrast normalization, combined with a novel enhancement of global abnormality signals in priors.

What is more, recent neuroimaging work shows that many scene-selective areas respond preferentially to HSF rather than LSF. Activity in the parahippocampal place area (PPA) was higher for HSF than LSF checkerboards, scenes, and faces (Rajimehr et al., 2011). Similarly, contrast-equalized HSF scenes activated the PPA and the occipital place area (OPA) more than LSF equivalents, although there was no difference in the retrosplenial cortex (RSC) (Kauffmann et al., 2015). Going beyond simple levels of activation, computational models can decode scene categories from BOLD signals in the PPA, RSC, and lateral occipital complex (LOC) of viewing photographs and line drawings (=HSF) (Walther et al., 2011). Similarly, scene category could be decoded from HSF photographs viewed for 800 ms in the PPA, RSC, LOC, and OPA, while LSF photographs could only be decoded in the posterior PPA (Berman et al., 2017). This increased activation and decoding in response to HSF demonstrate the important role of HSF’s contours and edges in rapid scene category processing. This fits with our behavioural findings of importance for HSF for mammogram-category extraction. There might be a similar role for HSF in both scene and medical abnormality gist extraction, again strengthening our belief that mammogram perception closely resembles scene perception.

Our filtering protocol included a brightness increase and contrast normalization. This method made the fine detail more visible in the filtered mammograms. A minor disadvantage is that this makes the data less informative for understanding the role of high spatial frequencies in conventional mammograms, as boosted brightness increased the weight given to the high frequency information. However, these stimuli remain ecologically valid, as no mammogram is ‘unaltered’. X-ray methodology creates a 2D representation of 3D tissue density, and the visibility of specific tissues depends on the specific machine, settings, image processing used (Cole et al., 2005), and even the practitioners’ preferential compression force (Mercer, Hogg, Szczepura, & Denton, 2013). What’s more, programs used for viewing medical cases often contain options to change the contrast or brightness. Thus, a brightness increase would not make the mammogram more or less 'naturalistic', it simply increased the chance of finding high-pass filters that enhanced detection rates, which was the main objective of this study.

Future research could focus on more fine-tuned enhancements by delving into the role of specific spatial frequency bands using bandpass or bandstop filters, which combined low- and high-pass filters to selectively retain *or* filter out a small band of frequencies. This would allow for more controlled adjustment of frequency content and could help identify the exact combination of spatial frequencies that contain the gist of abnormality. This could for example be used to filter out F0 – F0.5 and F1-F1.5 to investigate whether this combination further enhances the gist signal.

It might also be worth considering whether these, or similar image enhancements have the same effects on different domains of medical imaging. Previous research has shown that a gist of abnormality signal is also detectable in various other imaging modalities, such as digital breast tomosynthesis (C. C. Wu et al., 2019), chest radiographs (Carmody, Nodine, & Kundel, 1981; Kundel & Nodine, 1975), and even pap test images (micrographs) of cervical cells (Evans, Georgian-Smith, et al., 2013). It is possible that a similar high-pass filter would increase the signals of abnormality in other medical images as well, especially for radiographs, but it is also conceivable that different tissues are differentially affected by the development of a cancerous abnormality and would require different spatial frequency filtering to enhance their gist of abnormality signals. By comparing effects on different imaging modalities future studies could investigate the best image enhancements for each, which could in addition give insight into the (dis)similarities in gist signal content between modalities.

3.7. Conclusion

In conclusion, we have shown that certain high-pass filters (F0.5 and F1.5 cpd) combined with brightness boosting and contrast normalization can retain overall performance while boosting the gist of abnormality signal in mammograms at future cancer risk. Especially the 0.5 cpd high-pass filter seemed promising in boosting the signal in priors, without reducing the signal in mammograms with obvious or subtle signs of cancer in mammograms, nor making the radiologists more conservative in their decisions. Future research could investigate the effects of image enhancements on additional medical imaging modalities, to explore whether these findings hold true across imaging types. Additionally, future experiments should use bandpass or bandstop filtering to selectively retain or remove spatial frequencies to further investigate the role of specific spatial frequency bands in mammograms. The approach could be used to inform about more subtle enhancements that could potentially further boost the gist signal allowing for even earlier cancer detection. Overall, our findings provide initial evidence for a viable solution to enhance the gist of abnormality in mammograms to use as a risk factor in the clinical toolbox for radiologists.

3.9. Declarations

Acknowledgements

This project was supported by Cancer Research UK (https://www.cancerresearchuk.org/) & EPSRC (https://www.ukri.org/councils/epsrc/) grant EDDCPJT\100027 awarded to KKE, while EMR was funded through a departmental scholarship from the department of Psychology at the University of York, United Kingdom for the 3-year duration of her PhD project.

We would also like to thank Lucy Starling and Lyndon Rakusen for their assistance with aspects of data collection.

Author contributions

EMR contributed significantly to conceptualization, methodology, software, investigation, formal analysis, visualization, project administration, and writing and editing of the manuscript.

KKE contributed significantly to conceptualization, methodology, software, investigation, supervision, and review and editing of the manuscript.

Declaration of interests

The authors declare no competing interests.

Data and code availability

Data from the experiment is available on OSF under Creative Commons Attribution (CC BY) license at https://osf.io/t8cdr/, as well as the main analysis scripts needed to extract the data and to perform the statistical analysis.

3.10. Appendices

Appendix A: Factors influencing AUC performance measure

For AUC, the linear mixed model analysis showed evidence of significant contributions of Category, Frequency, and an Interaction factor. The full model had an intercept of 0.760, and a random effect of ID intercept of 0.002, and an AIC of -589.77. An ANOVA comparing the log-likelihoods of the full model to the model without the category factor showed a significant difference ($\chi^2(2)= 168.97$, p=<.001), showing that category significantly adds to the model fit. Similarly, the frequency factor contributes significantly compared to a model without this factor ($\chi^2(4)= 46.627$, p=<.001). Lastly, the interaction factor was also significant ($\chi^2(8)=75.396$, p=<.001).

Pair-wise comparisons were performed for the frequency (Fig 4A), as well as mammogram category factors (Fig 5). Again, interaction effects were reviewed with a special focus on the F0.5 and F1.5 groups that showed no significant difference in overall AUC compared to F0. These comparisons showed that AUC for prior mammograms was significantly higher for F0.5 (estimated difference=0.134, t(264)=5.844, p=<.001) and F1.5 (estimated difference=0.110, t(270)=3.843, p=<.001) than the unfiltered F0 group. Meanwhile, there was no significant difference in AUC between F0 and F0.5 for obvious (estimated difference=0.036, t(264)=1.568, p=.519) or subtle (estimated difference=0.024, t(264)=1.033, p=.840) mammograms. On the other hand, for F0 versus F1.5, there was no difference in AUC for obvious mammograms (estimated difference=0.055, t(270)=1.931, p=.303), but there was a reduction in AUC for subtle mammograms at F1.5 (estimated difference=0.122, t(270)=4.292, p=<.001). These interactions can also be observed in Fig 5.

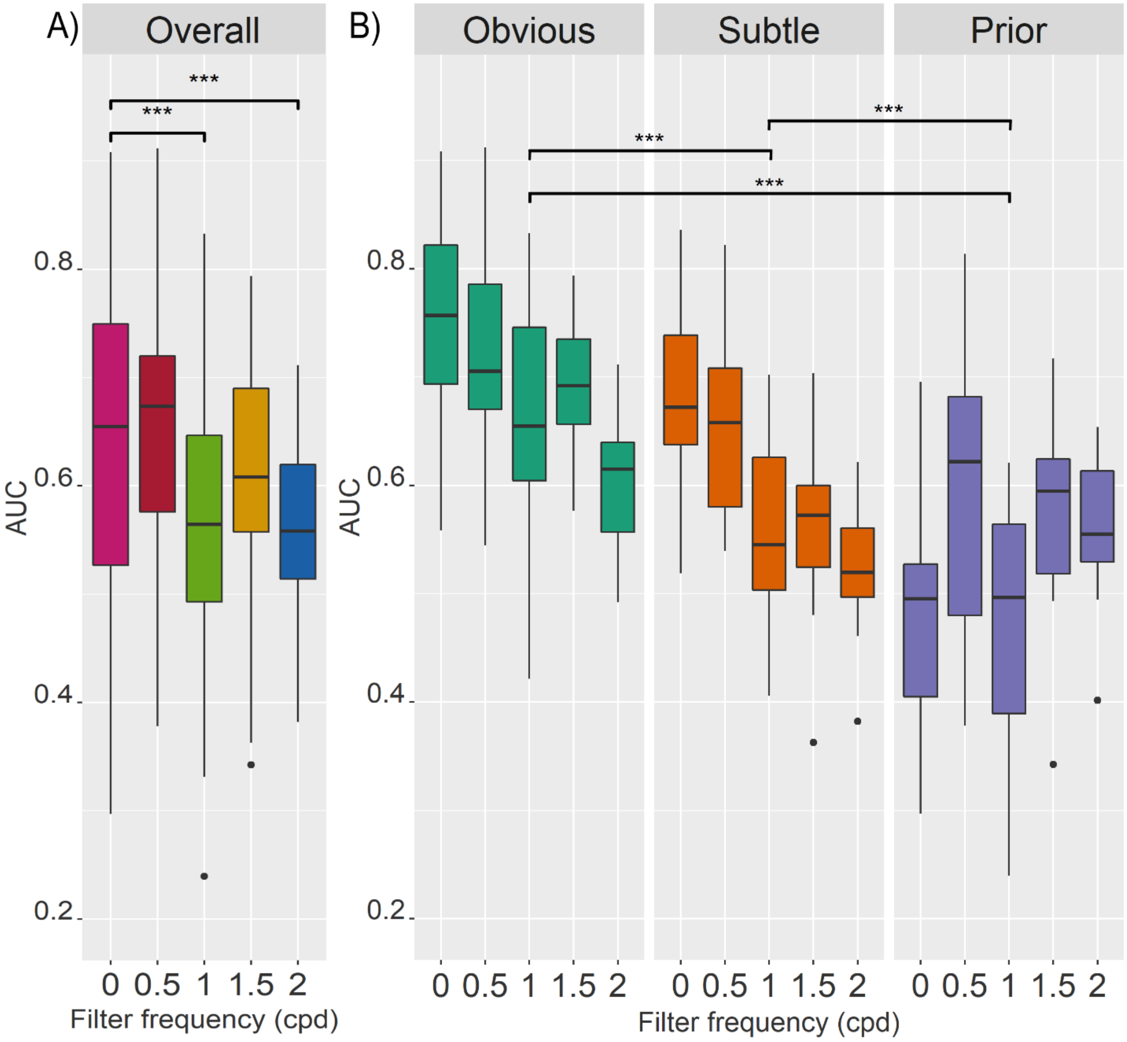


Figure 3.5. *Boxplots of AUC across conditions. Each boxplot shows the median as the line within the coloured box containing the 25th and 75th percentiles, with extending whiskers containing the 95% CI, with any outliers plotted as dots. Significance of pairwise comparisons is indicated in the*

*figure with *=p<.05, **=p<.01,***=p<.001. (A) Boxplots showing AUC across frequency conditions (0, 0.5, 1, 1.5, and 2 cpd). Pairwise comparisons of frequency showed that AUC was significantly higher for F0 than F1 (estimated difference=0.089, t(273)=6.533, p=<.0001) and F2 (estimated difference=0.098, t(284)=5.647, p=<.0001), but did not differ significantly from F0.5 (estimated difference=-0.025, t(273)=-1.827, p=.360) and F1.5 (estimated difference=0.023, t(284)=1.306, p=.688). (B) Boxplots showing AUC for each mammogram category (obvious, subtle, and prior) and frequency, to illustrate mammogram category and interaction effects. Pairwise comparisons of mammogram category showed that D' was significantly higher obvious than subtle (estimated difference=0.089, t(258)=7.058, p=<.0001) and prior mammograms (estimated difference=0.156, t(258)=12.368, p=<.0001), and higher for subtle than prior mammograms (estimated difference=0.068, t(258)=5.310, p=<.001).*

3.11: Commentary: Image processing

The main aim of this study was to find ways to boost the gist of medical abnormality, which was successfully achieved with the combination of image enhancements utilized. As described in the materials & methods, the high pass filters in this study were followed by a 3.5x brightness increase and then contrast normalisation with the unfiltered mammograms. The brightness increase improved the visibility of the high spatial frequencies, as the high-pass filtered images had low brightness after the high-pass filters were applied. This drop in brightness was caused by removing the 0 Hz frequency that consists of the average brightness of an image, as well as removing the brightness contained in lower frequencies. The latter is comparable to natural images, in which the low frequencies also contain more contrast energy than the high ones, following an inverse power law (Perfetto et al., 2020).

The brightness increase factor was sufficient to brighten the images without reaching ceiling levels: across all mammograms, less than 0.25% of the pixels that were increased in brightness were capped at a ceiling level of brightness (255) for the lowest filter level of F0.5, with this proportion decreasing further with increasing filter levels. Additionally, the images were contrast normalised with the unfiltered images afterwards, meaning that there were no differences in overall contrast energy of the stimuli used in the experiment.

Chapter 4: Future studies using bandstop filters to selectively remove specific frequency bands likely will not need to use brightness boosting, as they would not remove 0 Hz. These bandstop filters could be helpful in further isolating the role of specific frequencies in carrying or masking the gist of medical abnormality. Using global feedback to induce learning of gist of abnormality in mammograms

Chapter 4 was published as Raat, E. M., Kyle-Davidson, C., & Evans, K. K. (2023). *Using global feedback to induce learning of gist of abnormality in mammograms* in Cognitive Research: Principles and Implications, 8(1), 1-22. This version of the article has been accepted for publication after peer review but is not the Version of Record and does not reflect post-acceptance improvements. It also has edited headers and figure numbers to fit with the thesis format. The Version of Record is available online at: doi.org/10.1186/s41235-022-00457-8

4.1. Declarations

Ethics approval and consent to participate

All participants provided informed consent before participating in this study. The study received ethical approval from the Departmental Ethics Committee of the Department of Psychology, University of York, United Kingdom (ID: 881).

Consent for publication

The mammograms used in the figures were sourced from the OPTIMUM database and have been fully anonymised.

Availability of data and material

The raw data generated and analysed during the current study are available on our OSF repository, https://osf.io/mv47p/

This data is available under Creative Commons Attribution-NonCommercial-ShareAlike 2.0 UK: England & Wales (CC BY-NC-SA 2.0 UK).

Competing interests

No competing interests to disclose.

Funding

EMR is funded through a departmental scholarship from the department of Psychology at the University of York, United Kingdom for the 3-year duration of her PhD project.

This project is supported by Cancer Research UK & EPSRC grant EDDCPJT\100027 awarded to KKE.

These funding bodies had no influence on the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

EMR made substantial contributions to the conception and design of the study, to the acquisition of the data, to the analysis and interpretation of data, and drafted the work.

CKD made substantial contributions to the analysis and interpretation of the data, and to the revising of the work.

KKE made substantial contributions to the conception and design of the study, to the analysis and interpretation of the data, and to the editing and revising the work.

Acknowledgements

We thank Isabel Farr for her help in preparing the stimuli, and for her assistance in the acquisition of the data. We thank Dr Roisin Bradley (York and Scarborough Teaching Hospitals) for providing radiological perceptual ratings for the testing set of mammograms.

4.2. Abstract

Extraction of global structural regularities provides general ‘gist’ of our everyday visual environment as it does the gist of abnormality for medical experts reviewing medical images. We investigated whether naïve observers could learn this gist of medical abnormality. Fifteen participants completed nine adaptive training sessions viewing four categories of unilateral mammograms: normal, obvious-abnormal, subtle-abnormal, and global signals of abnormality (mammograms with no visible lesions but from breasts contralateral to or years prior to development of cancer) and receiving only categorical feedback. Performance was tested pre-training, post-training, and after week’s retention on 200 mammograms viewed for 500 ms without feedback. Performance measured as d’ was modulated by mammogram category, with highest performance for mammograms with visible lesions. Post-training, twelve observed showed increased d’ for all mammogram categories but a sub-set of nine, labelled learners also showed a positive correlation of d’ across training. Critically, learners learned to detect abnormality in mammograms with only the global signals, but improvements were poorly retained. A state-of-the-art breast cancer classifier detected mammograms with lesions but struggled to detect cancer in mammograms with the global signal of abnormality. Gist of abnormality can be learned through perceptual/incidental learning in mammograms both with and without visible lesions, subject to individual differences. Poor retention suggests perceptual tuning to gist needs maintenance, converging with findings that radiologists’ gist performance correlates with number of cases reviewed per year, not years of experience. The human visual system can tune itself to complex global signals not easily captured by current Deep Neural Networks.

Key words: gist of abnormality, gist extraction, medical image perception, medical expertise, medical imaging, perceptual learning, implicit learning, statistical learning, deep neural network

4.3. Significance statement

Breast screening plays a vital role in early diagnosis of breast cancers, which is essential for improving patient outcomes. Correct interpretation of mammograms relies on both medical knowledge and perceptual expertise. Perceptual expertise is thought to increase effectiveness of gist extraction: the ability to recognise global properties of an image after brief exposure. Indeed, expert radiologists can detect a global 'gist of abnormality' from mammograms after just 250 milliseconds with above chance accuracy, even when no visible lesions are present, for example in breasts contralateral to breast with cancer, or breasts that will develop cancer in the nearby future (Evans et al, 2016, 2019). This suggests that the gist of abnormality could be of clinical use as a risk factor. However, gist extraction performance varies between radiologists, correlating with number of mammograms screened within a year, suggesting an important role of perceptual exposure. How human observers develop the ability to extract the gist of a new categories is unknown. Understanding the development of perceptual expertise for gist extraction could be leveraged to enhance training of radiology residents and could be used to train perceptual experts for the purpose triage or evaluating risk assessment. The current work provides a proof-of-concept training paradigm that was able to induce the learning of the gist of abnormality in naïve observers without any medical training, using perceptual exposure and global feedback. Our findings support the idea that gist extraction abilities can develop separately from medical knowledge and can be developed through simple, perceptual training paradigms.

4.4. Introduction

Medical experts often report having a gut feeling about the state of a radiograph when briefly looking at certain medical imaging cases, where they get the impression that something might be wrong but are not able to pinpoint the exact image elements that made them feel that way. These anecdotes suggest medical experts might rapidly access first impressions of abnormality. However, there is more than just anecdotal evidence for this notion: it is also supported by human observer studies, which have shown that radiologists are able to discriminate between normal and abnormal medical images with above-chance accuracy within 200-500 ms for chest radiographs (Kundel & Nodine, 1975), pathology images, or mammograms (Evans, Georgian-Smith, et al., 2013), the latter of which will be the focus of the current study. Thus, medical experts indeed possess the perceptual ability to rapidly extract a signal that indicates abnormality from images in their field of expertise.

This shows an incredible perceptive power, which is furthered by research demonstrating that the ability does not rely on the presence of a localizable signal like a lesion. Indeed, radiologists can recognise this gist of abnormality in patches of the abnormal mammogram that do not contain a lesion, or even from the breast contralateral to the abnormality (Evans et al., 2016), both of which do not contain any localizable abnormalities. Even more striking, when normal mammograms from women who went on to develop cancer in the next two to three years were intermixed with normal and abnormal mammograms, they were rated as significantly more abnormal than the normal images (Patrick C. Brennan et al., 2018; Evans et al., 2019). Thus, the gist of abnormality signal can be detected without localizable abnormalities. For mammograms containing a single mass, it has been suggested that radiologists can sometimes access coarse location information (Carrigan et al., 2018), although this study did remove image artefacts and large calcifications from the breast tissue. Together, these findings point to a rapidly extracted global signal of image statistics that allows medical experts to detect whether the imaged tissue is normal or abnormal, which might provide access to coarse location information, but does not require local information to function. This description fits closely with the process of gist extraction that has been widely described in scene processing literature.

Gist extraction is a perceptual process that allows observers to quickly retrieve the global meaning, or gist, of visual input. After as little as 20-30 ms, humans can accurately discriminate between man-made and natural environments, so-called superordinate categories (Joubert et al., 2009), recognize forests, fields, rivers, and other basic scene categories (Greene & Oliva, 2009), or determine the presence or absence of broad categories such as animals (Bacon-Macé et al., 2005) or vehicles (VanRullen & Thorpe, 2001). Indeed, there is a wide range of research showing that humans can extract surprisingly complex information from rapidly presented visual information, which fits closely with the observations in rapid medical image perception.

The key characteristics of gist extraction are that it occurs rapidly, globally (across the whole image) with loss of specific local information and does not require focused attention. Instead, it occurs without prior location of items and in a non-selective manner. For example, gist can be extracted from scenes in the periphery in parallel with a demanding foveal letter discrimination task (F. F. Li et al., 2003) or from two, or even four scenes in parallel with minimal drops in performance (Rousselet et al., 2004) or scenes presented in medium to far periphery (Boucart et al., 2013; Larson & Loschky, 2009), clearly showcasing the global and non-selective nature of the process. In addition, gist extraction does not require prior configuration of the visual system: it occurs when monitoring for multiple cue categories simultaneously (Evans, Horowitz, et al., 2011), or even when the target category is post-cued after a rapid serial visual presentation (Evans, Horowitz, et al., 2011; Potter et al., 2014). However, it also means that information about locations of specific elements that make up the scene is not consciously accessible (Evans & Treisman, 2005). Overall, scene gist extraction clearly occurs rapidly, globally, and without the need of focused attention or pre-selection, which fits closely with the observations of what we will refer to as the *gist of (medical) abnormality.*

But which signals are extracted by this global, rapid process to contribute to the formation of our gist understanding? As every image is built up from spatial frequencies at various orientations, shared categorical regularities between a gist category might be captured in similarities in spatial structural regularities, as described by Portilla and Simoncelli (2000)'s statistics. The statistic they define are extracted using spatial filters of specific sizes and orientations and are applied to noise to create an artificial 'metamer', that contains the same spatial structural regularities, but no recognizable objects. Such a metamer is indistinguishable from the original in two alternative forced choice task (2-AFC) at 200 ms viewing time (Freeman & Simoncelli, 2011), suggesting that spatial structural regularities capture essential aspects of scenes that are accessed during gist extraction. The idea of a statistical signature of an image fits with the Efficient Coding Hypothesis (Simoncelli, 2003), as reducing an image to its spatial structural regularities would allow efficient encoding of its essential information. Mammogram content is even more closely related to its spatial frequency content than scene images, due to most of the content being textural. For example, previous research has shown that low-pass filtering strongly reduced gist extraction, while high-pass filtered mammograms retained most gist information (Evans et al., 2016). Spatial structural regularities might be more similar between images from the same category and thus allow for flexible perceptual rules for gist categorization.

Oliva and Torralba (2001) further explained these spatial structural regularities with a focus on human perception through gist descriptors, which similarly captured spatial frequency patterns on a global spatial scale, the global spatial envelope. Gist descriptors can be represented as scores on scales such as expansiveness and openness. Patterns in these feature scores have been shown to be more similar within than between scene categories. Additionally, false alarms made by observers could often be predicted by similarities in gist predictors (Greene & Oliva, 2009). This supports the idea that shared patterns of frequencies and textures could play an important role in the flexible, yet reliable gist categorization of scenes, which could reasonably be extended to mammograms.

To allow for its non-selective and global nature, gist extraction must be highly flexible, especially as it must generalize across a wide range of exemplars that all fall under one gist category. For example, we can recognise the gist category of a scene environment in a variety of conditions, such as viewing angles, lighting, and specific objects (figure 1A, 1B), and the same applies to mammograms, as these can also vary widely in their appearance, size, shape, density, and texture (Fig. 1C, 1D). However, previous experience influences our ability to extract gist accurately, as human observers performed considerably worse on scene gist extraction for photographs from aerial compared to terrestrial viewpoints (Loschky, Ringer, Ellis, & Hansen, 2015). Thus, our brain might develop a set of general perceptual rules of expected spatial regularities for each gist category, based on previous experience, that are flexible enough to generalize across variations, but specific enough to allow it to distinguish a beach from a river, or a normal from an abnormal mammogram.

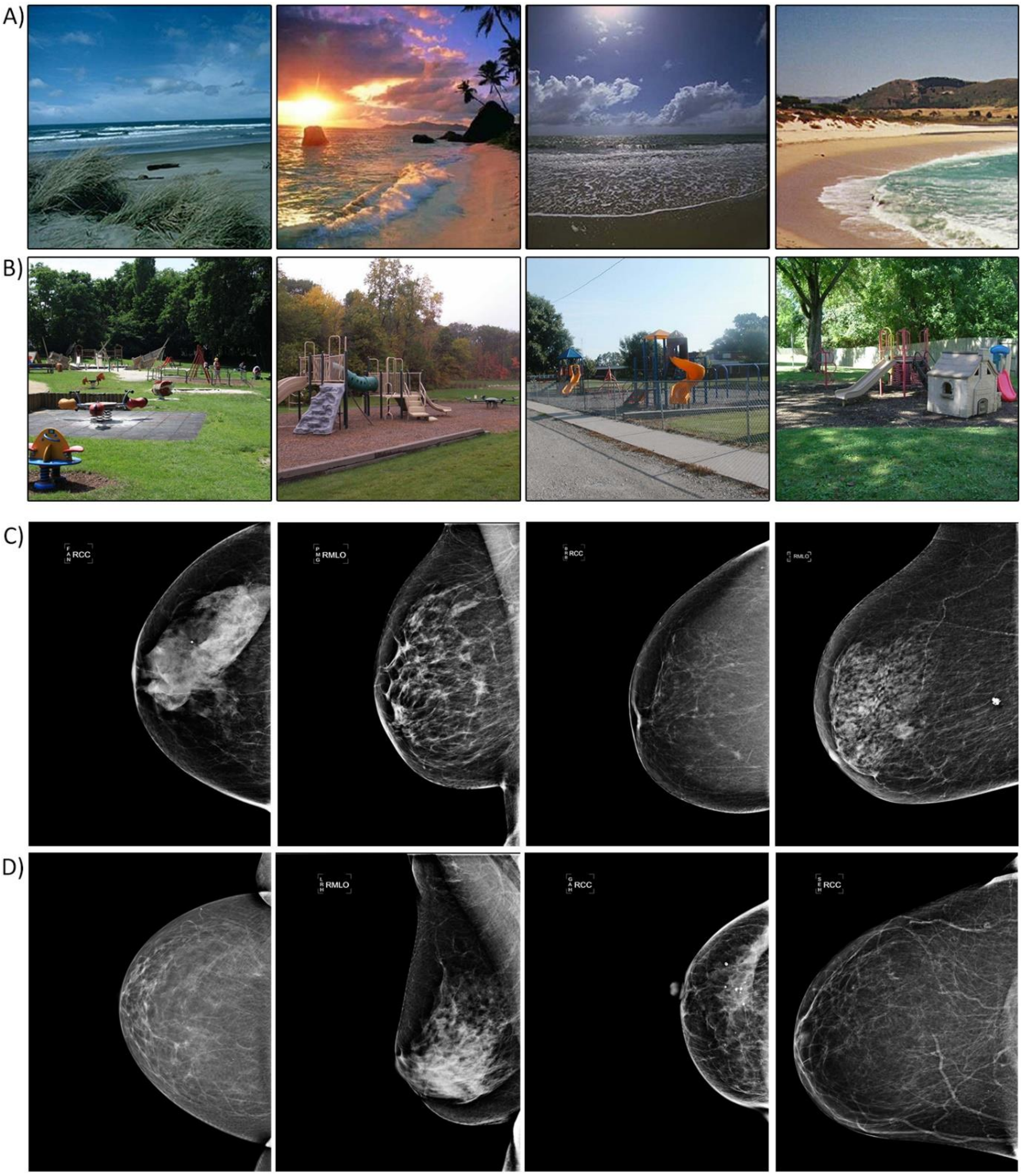


Figure 4.1: *Scene exemplars for beaches (A) and playgrounds (B) that illustrate the variation in viewing angle, lightning, configuration, and specific objects. Mammogram exemplars containing subtle abnormalities (C) or no abnormalities (D) illustrating the variation in shape, size, and textural patterns.*

However, it is not yet known how people acquire these sets of expectations or sensitivity to emergent statistics needed to extract the gist of novel categories, whether that is a natural scene category, or a more abstract categorisation of a medical image. Since the learning of natural scene categories happens during normal development, this learning must be able to occur under natural viewing conditions and should not rely on detailed feedback that explicitly explains which features make the scene a beach. Rather, the learning would be expected to reliably occur with broad feedback consisting of just categorical information ('We are at a beach'). This learning would be in line with the principles of statistical learning, the process through which humans can extract naturally occurring statistical patterns in space and/or time (Turk-Browne et al., 2005).

Indeed, statistical learning leads observers to recognise temporal or spatial statistical regularities and patterns in auditory or visual stimuli after a multitude of exposures without explicit instructions on what to learn (Turk-Browne, 2012). For example, passively viewing a stream of symbols produced strong familiarity feeling for viewed patterns (Fiser & Aslin, 2002a). Interestingly, children as young as 9-months old pay more attention to arrays containing previously seen shape arrangements than new arrangements (Fiser & Aslin, 2002b), suggesting that statistical learning takes place from early on in our development. While the previous examples used simple shapes, statistical learning also extends to more complex stimuli, such as scene images. Observers report more familiarity to scene sequences, such as a kitchen followed by a forest, that were previously seen in a visual stream (300 ms each) without being instructed to pay attention to the order of scene categories (Brady & Oliva, 2007).

Statistical learning is often investigated in the context of temporally separated stimuli, but like previously stated, it also occurs over spatial regularities, which would form the basis for gist category learning. Indeed, observers become familiar with the configurations of complex objects in a grid through repeated exposure (Fiser & Aslin, 2001), and they can decrease their reaction time in a search task due to repeated configurations of distractor arrays without recognition of repeated arrays occurring (Chun & Jiang, 1998), as they implicitly learn to recognise the regularities in contextual cues, or in other words, invariant visual properties, allowing them to interact with the environment more efficiently (Chun, 2000). Similarly, someone might learn to recognise the invariant global properties of a forest, beach, or even an abnormal mammogram through statistical learning of spatial regularities. Statistical learning with global feedback allowed observers to recognise camouflaged objects by learning the general statistics of the background (X. Chen & Hegdé, 2012). Thus, in our definition of statistical, implicit learning, no assumptions are made about the unconscious nature of the learning or complete lack of awareness of learned patterns, but only that it consists of learning through repeated exposure without explicit instructions or feedback on which features or patterns to extract. We expect that statistical learning through repeated perceptual exposure to novel categories and their group labels would allow observers to acquire the gist of a new category.

To investigate the learning of gist signals, a category is needed in which observers can be trained to improve. Previous training research has shown that the speed of gist extraction from natural scenes is already optimized and at ceiling levels, as extensive training across 15 days did not significantly speed up reaction time of a 2-AFC animal absent/present task (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001). While accuracy increased slightly and average reaction time decreased slightly for familiarized stimuli, this did not transfer to new stimuli, and was mostly driven by an increase in speed/accuracy for the most difficult familiarized targets with RTs above 400 ms. Thus, the processes underlying gist extraction for scenes of categories are already highly efficient in adults, and do not seem to be able to be further compressed or enhanced. Thus, scenes cannot be used to investigate the processes involved in the learning of a new category of gist. However, it does underline the fact that scene categories must be deeply familiar to the average human observer, which would only be possible if the global gist is learned through the rare instances of explicit feedback (‘these exact features make this a beach/forest/mountain’) or, as we hypothesize, is largely learned through the frequent global feedback moments we encounter in daily life (‘you are in a forest’). Interestingly, expertise within a specific object category, such as cars, will increase the ability to rapidly detect scenes containing that object category, but not others (e.g. humans), in a simultaneous presentation of two scenes (Reeder, Stein, & Peelen, 2016), adding support to the idea that expertise in a category might influence rapid detection of that category, similar to what is seen in medical experts.

For the gist of medical abnormality, previous research has repeatedly shown that, as expected, naïve observers are unable to extract this signal (Evans, Georgian-Smith, et al., 2013; Raat et al., 2021), showing that the general population is not familiar with this gist signal representing a medical abnormality. Interestingly, however, a recent study trained naïve observers to recognise obviously visibly abnormal mammograms (microcalcifications/breast mass) with above-chance accuracy after approximately 600 cases of training (Hegdé, 2020), showing that non-medically trained observers can develop the perceptual ability to recognise obvious abnormalities on free-viewing tasks. This indicates that naïve observers can at the very least learn to recognise perceptual characteristics of lesions in mammograms a localized signal, which suggests they might also be able to be trained to recognize the gist signals of abnormality in the overall tissue.

Thus, this study’s aims are twofold: to investigate whether/how people can learn the categorisation of a new gist signal (medical abnormality), and to explore which perceptual features in mammograms might drive this gist signal. We will evaluate whether naïve observers can learn to rapidly recognise the gist of a new category after repeated perceptual exposure through training with global feedback, and if this learning is retained after the end of training. Global feedback is defined as the ground truth of the trial, without additional instructions on location of abnormalities or potential features that might indicate the ground truth. In other words, the task and label are both made explicit, but since no further guidance on which content to use is provided, only implicit/statistical learning can be used. Since gist of abnormality is a global signal, learning to recognise the gist of abnormality should improve performance on not only mammograms with visible abnormalities, but also on mammograms with only global signals of abnormality, such as contralateral mammograms or those taken prior to the development of localizable cancer, similar to the ability of trained medical experts (Patrick C. Brennan et al., 2018; Evans et al., 2019; Evans et al., 2016). Based on the framework of gist development, and the previous findings of Hegdé (2020), training is expected to induce learning of the gist of medical abnormality, and this is expected to improve performance for mammograms with and without local abnormalities.

As an extension to the training findings, we will also evaluate the performance of a state-of-the-art machine learning model on the same images and compare it to human perception. Human statistical, implicit learning shares key similarities with the concept of deep learning, a computational method where each decision is compared to feedback of a simple label, inducing learning through backpropagation of the error between the decision and ground truth, which can lead to tuning towards statistical regularities in the input (Voulodimos, Doulamis, Doulamis, & Protopapadakis, 2018). Both describe conceptually similar processes that could underlie learning without explicit rules or instructions. As one type of computational modelling, deep learning, was developed based on observed brain architecture and processing (Voulodimos et al., 2018). Deep learning models can capture complex visual patterns, allowing for object (Ouyang et al., 2016; Simonyan & Zisserman, 2014) and facial recognition (Taigman, Yang, Ranzato, & Wolf, 2014).

By comparing human and machine performance on specific images, we can learn more about whether these models capture the same image features that humans might be using – which in turn can be informative for human perception. The single breast classifier (SBC) version of N. Wu et al. (2019) deep neural network (DNN) for breast cancer screening predicts probability of both benign and malignant abnormalities for individual unilateral mammograms, and reaches a high performance (AUC malignant: 0.84-0.90, AUC benign: 0.74-0.76) on detecting visible abnormalities in a large screening dataset, which make it suitable for our purposes. We will use both the SBC and SBC heatmap (SBC+HM) version, which adds heatmaps generated via a secondary network which examines smaller pixel patches for their malignancy probability. These heatmaps provide additional scrutiny of local information that is expected to improve performance, while the SBC without heatmaps would be more equivalent to the global information used in gist extraction. Comparing the probability scores from both the SBC and

SBC+HM network to human rating scores will allow us to investigate whether they capture similar information used by human gist extraction of medical abnormality.

4.5. Methods

Participants

19 adults without previous medical training or experience with viewing mammograms took part in this multi-session experiment, of which 4 withdrew their participation during the training phase. The remaining 15 participants were included in the final dataset (aged 20-38, average age 23, 11 female) as they all passed the pre-determined exclusion criteria. Exclusion criteria were pre-defined in order to exclude participants if there was significant evidence to suggest inattention, defined as 1) having missed more than 30 out of 144 attention trials in total across the 9 training sessions, 2) having failed more than 6 out of 16 attention trials in one training session, or 3) having rated 85% or more of the trials as 50 in any testing session or more than 1 training session. Attention trials which were randomly interspersed across different points in the training sessions, briefly showed an image of a beach or forest, which the participant was asked to categorize, a task that should be trivial if the screen was attended.

Participants received a compensation of 50 pounds for their time (~5 pound per hour) after completing all 10 sessions and they receive a bonus payment of 10 pounds if they passed 95% or more of the attention checks, as an incentive for them to pay close attention to each trial. Participants all had normal or corrected-to-normal vision. All participants had completed at least their A-levels or equivalent. The sample size was based on the work by Hegdé (2020), which reported significant learning during an untimed mammography training experiment with 11 and 14 general population participants in two separate experiments.

Stimuli and apparatus

The stimuli used in this experiment were 8-bit PNG images of four categories of anonymized unilateral mammograms in mediolateral oblique (MLO) or craniocaudal (CC) view: normal mammograms of healthy women (normal), mammograms with obvious cancerous abnormalities (obvious), mammograms with subtle cancerous abnormalities such as architectural distortions (subtle), mammograms without visibly actionable lesions that are thought to contain global features of abnormality (either contralateral to a breast with a cancerous abnormality (contralateral), or mammograms taken one to six years prior to visible actionable sign of abnormality appearing in a subsequent scan (priors)). The labels ‘obvious’ and ‘subtle’ were categorised as such by an experienced radiologist for the Complex Cognitive Processing Laboratory of the University of York. Further information about cancer type descriptors can be found in Appendix A. Contralateral and prior cases were combined into one category, as both contain global signals of abnormality and lack any localizable lesions. The normal, obvious, subtle, and contralateral cases were sourced from the OPTIMAM database. The priors were sourced from the Complex Cognitive Processing Lab database in collaboration with Dr. Bradley of the York Hospital for this study. The majority of selected mammograms were acquired with Lorad Selenia (75.4%) and Selenia Dimensions (13.5%), with a smaller portion of mammograms acquired with Senographe Essential (8.9%) and the L30 (1.8%), and a minority taken by MammoDiagnost DR (0.3%) and Mammomat Novation DR (0.1%). All mammograms that are part of the Complex Cognitive Processing lab database of stimuli can be shared with other researchers upon reasonable request to the senior author (K.K. Evans), while the OPTIMAM database is also available for research purposes through an application process (https://medphys.royalsurrey.nhs.uk/omidb/getting-access/).

The training set was composed of 5668 unilateral mammograms, consisting of 1558 normal, 1019 obvious, 899 subtle, and 2192 global (1868 contralateral, 324 prior) images, so approximately 72% of the available stimuli contained the gist of abnormality. This large dataset ensured that participants were trained on a wide range of mammograms and reduced the number of repetitions. Some repetitions occurred randomly across the 36 blocks, but never within a block: on average, normal mammograms were repeated 0.9 times, obvious, subtle, and contralateral mammograms were repeated <0.1 times, priors were repeated 2 times.

The testing set consisted of 200 unilateral mammograms: 80 normal, 30 obvious, 30 subtle, 30 contralateral, and 30 prior mammograms, meaning 60% of the stimuli contained gist of abnormality signals. The same images were used for each test session to equate the difficulty level across participants and testing phases, and these were not used during training phases. Previous research has shown very low recognition memory in both general population (d' prime=.36) and radiologists (d'=.86) when tested on recognition directly after exposure to 72 mammograms viewed for 3 seconds each (Evans, Cohen, et al., 2011). Since we use a larger number of mammograms shown for shorter durations and with longer inter-exposure intervals no significant memory effects were expected, especially since no feedback was given on the test cases.

To further characterize the test cases, an experienced mammogram reading radiologist assessed each mammogram on radiological perceptual features. The following radiological features were rated: 1) four-point BI-RAD breast density scale (D'Orsi, Bassett, & Feig, 2018) as I) fatty, II) mixed but predominantly fatty, III) mixed but predominantly glandular, and IV) extremely dense), 2) breast pattern as normal or complex, and 3) level of concern/suspicion on a five-point scale from I) normal, II) benign, III) indeterminate, IV) suspicious, V) malignant. Chi-square tests of independence showed no significant association between density and image type ($X^2(12)=9.63$, *p*=.648). Associations between image type and both breast pattern ($X^2(4)=11.50$, *p*=.021) and level of concern ($X^2(16)=138.05$ *p*<.001) underline that an experienced radiologist could detect radiological perceptual differences in our cases, but that these signals were not driven by density. Thus, simply becoming sensitive to the density of mammograms would not result in significant increases in performance. This is in line with previous studies, that also showed a lack of correlation between BIRAD density and gist of abnormality ratings (Evans, Birdwell, et al., 2013; Evans et al., 2019; Evans, Georgian-Smith, et al., 2013; Evans et al., 2016).

The experiment took place on a computer or laptop screen, with the participant using a mouse and keyboard to submit rating responses. Since the experiment took place online, the exact apparatus varied between participants. However, physical stimulus size was equated by using a screen calibration method using either diagonal screen length or a credit-card size matching task inspired by the method proposed by Q. Li et al. (2020) to ensure the images were displayed as 12.8 cm/5 inches high by 15.75 cm/6.2 inches wide across all sessions and participants. The experiment was accessed via a website optimized for Firefox and Chrome browsers, where participants could log in for each session according to the scheduling rules, using their unique user ID.

Procedure

This study used a multi-session within-subject repeated measures design. It consisted of a total of 9 training phases and three testing phases completed across 10 sessions spread out over multiple days, as is summarized in the flowchart of Figure 2. Before the first session, each participant joined an individual video conferencing call via Zoom with the experimenter to guide them through the instructions and check for any questions or technical difficulties. During this conference call, the participants also watched a pre-recorded instruction video, explaining what a mammogram is and what the experiment task is, to ensure all participants received identical instructions. The first session started with a pre-training test phase to establish baseline of performance. After the pre-training baseline, participants immediately performed the first training phase, which was followed by 7 subsequent sessions consisting of training phase each, separated by at least 1 and at most 3 days each. The 9th session consisted of the last training phase and a subsequent post-training test phase to measure potential improvements in performance. The tenth and last session took place 7 to 10 days after the last training session and consisted of a retention test of performance. Participants scheduled their own sessions according to these scheduling rules but received regular reminder emails to inform them when their next session was due.

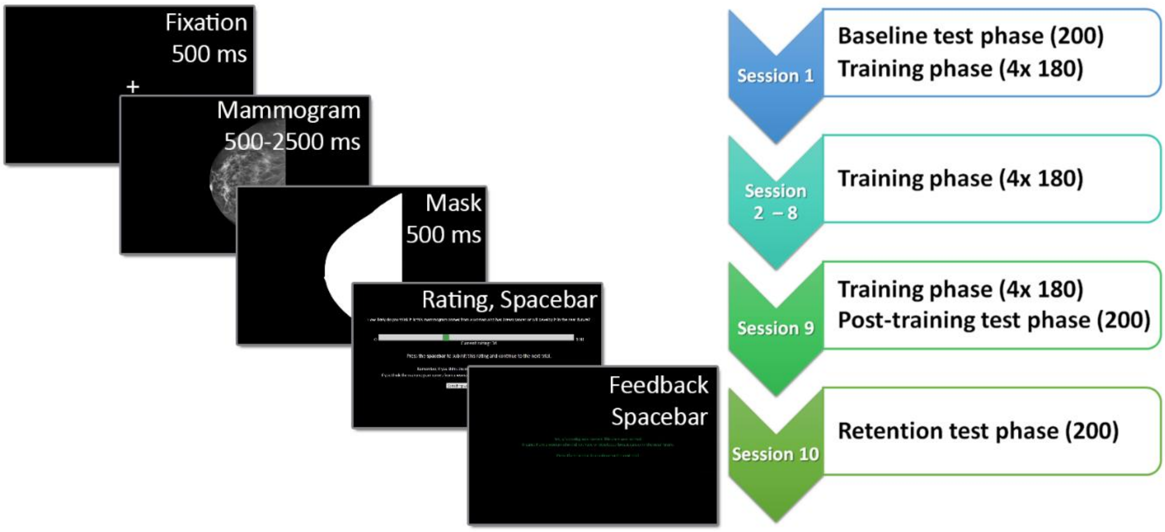


Figure 4.2: *Overview of the experimental procedure and flow-chart schedule of the experiment. The screens show the presentation order within a training trial and the duration or button press to continue. Test trials always showed mammograms for 500 ms and omitted the feedback screen but were otherwise identical. The flow-chart schedule shows the order of experimental phases for each session, and the number of unilateral mammograms viewed per session. In the test phases, 200 mammograms were viewed, while the training phases had 4 blocks with 180 mammograms each. Session 1 to 9 were separated by 1 to 3 days each, while session 10 was delayed by 7 to 10 days after session 9.*

Both test and training trials followed a similar format (Fig. 2). They each consisted of a fixation cross (500 ms), the mammogram (500 ms or 500-2500 ms), a mask of the filled shape of the mammogram (500 ms), followed by a rating scale between 0 and 100 (self-paced). Participants were asked to give their decision by adjusting a curser on a rating scale that would indicate how sure they were that a unilateral mammogram was normal of abnormal. This rating was then used as a performance measure applying signal detection theory methodology. In the training trials, this was followed by a feedback screen (self-paced). Feedback was based on the rating decision and ground truth, e.g., if the ground truth was abnormal, ratings above 50 were counted as correct, and ratings of 50 or below were counted as incorrect. The feedback screen informed participants whether their submitted rating was correct or incorrect, and whether the ground truth for the trial was normal or abnormal. The colour of the text was green for correct and red for incorrect ratings. Participants received no feedback during the test phases.

Each test phase consisted of 203 trials: 3 practice trials with feedback to familiarize them with the task, then 200 test trials showing the pre-selected test set in a randomized order. The test set consisted of 80 normal mammograms, and 30 each of the four abnormal categories (see stimuli and apparatus for more details). Each mammogram was shown for 500 ms before the mask and then rating screen appeared.

Each training phase consisted of a total of 736 trials, split into 4 blocks of 184 trials each: 180 mammograms, and 4 attention trials dispersed throughout each quarter of the block. The 180 mammograms were randomly selected from the training set to show 72 normal mammograms, 27 obvious, 27 subtle, and 54 global abnormal mammograms. More global abnormal mammograms were shown because these are thought to be both the most difficult, and the most likely to contain the global gist signal, on which we would expect increased performance if indeed a gist signal was learned. The attention trials showed easily recognisable colour photographs of either a forest or a beach, and had an alternative rating instruction to rate beaches as 0 and forests as 100. These trials also showed feedback based on the response, however, if the answer was incorrect, the feedback screen was shown for at least 10 seconds before they could continue, and the attention trial was repeated until they answered correctly. Participants were encouraged to take self-paced breaks in between each block.

During the training session, the maximum viewing time for the mammogram started at 2500 ms in the first block to familiarize the participants with the procedure and task. Participants were encouraged to press the spacebar as soon as they had a first impression to continue to the mask, then rating screen (minimum viewing time 500 ms). However, this was not required, and the mammogram would automatically be replaced by the mask at the maximum viewing time. In subsequent blocks, maximum viewing time was adapted based on performance: if the total d' prime for the block was above 0.2, max viewing time was decreased to 90% of the average actual viewing time of that block, but if d' prime was below 0.05, it instead increased to 105% of the current maximum viewing time to a maximum of 2500 ms.

Data analysis

Signal detection measures were used for analysing observers' performance, as these can differentiate performance (d') and response biases (criterion) in a binary classification task, calculated from proportions of hits and false alarms. D' characterizes accuracy of performance, with a d' of 0 representing chance and higher values representing better performance. Criterion characterizes response bias, with a criterion of 0 being unbiased, a negative criterion is liberal, meaning that any random trial the participant is more likely to label it as abnormal than normal, and the opposite is true for a positive criterion, which is conservative, leaning towards rating trials as normal.

First, proportions of hits and false alarms were calculated from the rating and ground truth (normal or abnormal) of the trials for each mammogram category. The numerical rating for a trial was compared to the set threshold of 50 for d’ and criterion: the binary rating decision was considered “normal” if below, or “abnormal if above the threshold. D’ was then calculated by subtracting the z-transformed false alarms from the hits (d’ = z(hits) – z(false alarms)). A d’ of zero represents chance performance, with positive values representing above-chance accuracy. Criterion on the other hand adds the z-transformed hit and false alarms rates and divides them by -2 (c= (z(hits) + z(false alarms))/-2). As the task explicitly instructed participants to rate normal trials below 50 and abnormal trials above 50, and to rate more extreme values the more confident they were, d’ and criterion at threshold 50 were the primary outcome measures of performance.

To further characterize the shape of the rating curves at different points of the experiment, area under the curve (AUC) measurements of Receiver operating characteristic curves (ROC) were used. ROCs were constructed by repeating the division of trials into proportions of hits and false alarms using a sliding value of normal/abnormal rating thresholds (1-99) and plotting all data points, from which the AUC was then calculated in Python. AUC ranges from 0 to 1, and represents the probability that a randomly chosen abnormal trial will be rated higher than a randomly chosen normal trial (Hanley & McNeil, 1982), with chance performance in a raw rating experiment yielding an AUC of 0.5 and higher AUCs representing more accurate performance.

The average and median viewing time of different screens were also calculated for the mammogram screen (training phases only), rating screen, and feedback screen (training phases only) for each of the sessions. Outlier rating times (outside of mean plus/minus 3 STD of the individual session) were excluded.

The main research question of whether naïve observers can learn a new category of gist through perceptual training was evaluated using 3-by-3 two-way repeated measures ANOVAs with 2 factors: testing moment (3 levels, pre-test, post-test, and retention test), and image type (3 levels, obvious, subtle, global) for d’ prime and criterion. To evaluate whether participants were engaged with the task, attention checks and feedback viewing time were evaluated with descriptive statistics. Additionally, to investigate potential differences in rating speed, which might signify elements of decision-making speed, before and after training, a 4-by-3 two-way repeated measures ANOVA was performed on rating time across the testing sessions (pre, post, retention) and image types (normal, obvious, subtle, global). For any repeated measures ANOVA with a significant effect of testing moment, planned simple contrasts were performed comparing between the pre-test and post-test, and the pre-test and retention test, as this was the primary research interest. Pearson’s correlations were calculated for d’ across the training phases, to evaluate whether individual performance improved throughout the training period. Based on the correlation coefficient, participants could be divided into learners (above 0 coefficient) and non-learners (below 0 coefficient), which were investigated with the main aim to explore the main effect of testing phase on performance. This method was also used on a bootstrapped simulation of a population making random rating decisions, to ensure that any learner vs non-learner effects were not caused by chance.

As an additional means of assessing whether participants outperformed chance, alternative Log-Linear-Likelihood ROCs and AUCs were calculated and compared to chance levels. This was based on methodology suggested by (Semizer, Michel, Evans, & Wolfe, 2018) to handle potential bimodal distributions that can result from raw rating experiments more accurately. ROC curves were smoothed with a Gaussian kernel, width 10, after which log likelihood ratios were calculated to compute the area under the curve (AUC). ROC curves and their AUCs are calculated for the real data and 100 randomly bootstrapped samples (with resampling). If the AUC of the real ROC was higher than the 95th percentile of the randomly bootstrapped AUCs, this strongly suggests that the participant outperformed chance.

Lastly, as exploratory analysis, we compared the ratings by human observers to the probability scores of benign/malignant findings from a deep neural network (DNN). Single unilateral mammograms were evaluated using the single breast classifier (SBC) and SBC plus heatmap (SBC+HM) version of N. Wu et al. (2019) DNN for breast cancer screening. 16-bit PNG versions of each unilateral mammograms were pre-processed to remove annotations and then run through the SBC and the SBC+HM. DNN inference was accomplished on Cloud Viking, a University of York HPC cluster. The compute nodes used were equipped with a NVIDIA V100 GPU. Stimuli supplied to the SBC had higher pixel dimensions than those shown to human observers, and a greater bit-depth, due to the requirements of the SBC. The output consisted of prediction scores for benign and malignant findings for each mammogram, ranging from 0 to 1, which were transformed to 0 to 100 scale to match the human rating scale. AUCs were calculated for the SBC and SBC+HM to evaluate overall performance. Image-level and category-level comparisons between human and SBC scores were made using Spearman's rank correlations, to investigate the level of agreement. These correlations were compared before and after training, to see if training increased the level of agreement between human and machine scores.

4.6. Results and Discussion

Human observer performance in training to detect cancer

Attention and task engagement

Participants were highly attentive during the training phases, as indicated by the very low number of incorrectly answered attention check trials (median: 0, mean: 0.93, std: 1.24, max: 4) across the 144 total checks in the 9 training phases. Additionally, participants actively used the spacebar to continue to the rating screen, meaning both their average and maximum viewing time rapidly decreased from 2500 ms, with all participants showing below 600 ms average maximum viewing time during the fourth training phase (see appendix B for more details on engagement and viewing times).

Effect of training on performance measures

Figure 3 shows the mean d', criterion and AUC for each image type pre-training, post-training, and at retention. Averaged over image types, d' increased after training in 12 out of 15 participants, with a mean d' of 0.274±0.058 prior to and 0.378±0.079 after training, and 0.255±0.086 at retention. Compared to pre-training, rating criterion became more liberal after training in 14 participants, and remained more liberal at retention in 13, with a mean criterion of -0.0377±0.073 prior to, -0.356±0.112 after training, and -0.284±0.114 at retention. Meanwhile, AUC was higher than pre-training in 9 out of 15 after training, and in 6 out of 15 at retention, with a mean of 0.582±0.016 prior to, 0.589±0.016 after training, and 0.568±0.018 at retention. Similarly, Log-Linear-Likelihood AUCs were compared to bootstrapped chance levels, which showed a sizeable increase in participants performing above chance levels after training (see Appendix C). Additionally, analysis of average and median rating times showed that participants took significantly less time to make rating decisions after completing their training (see Appendix D).

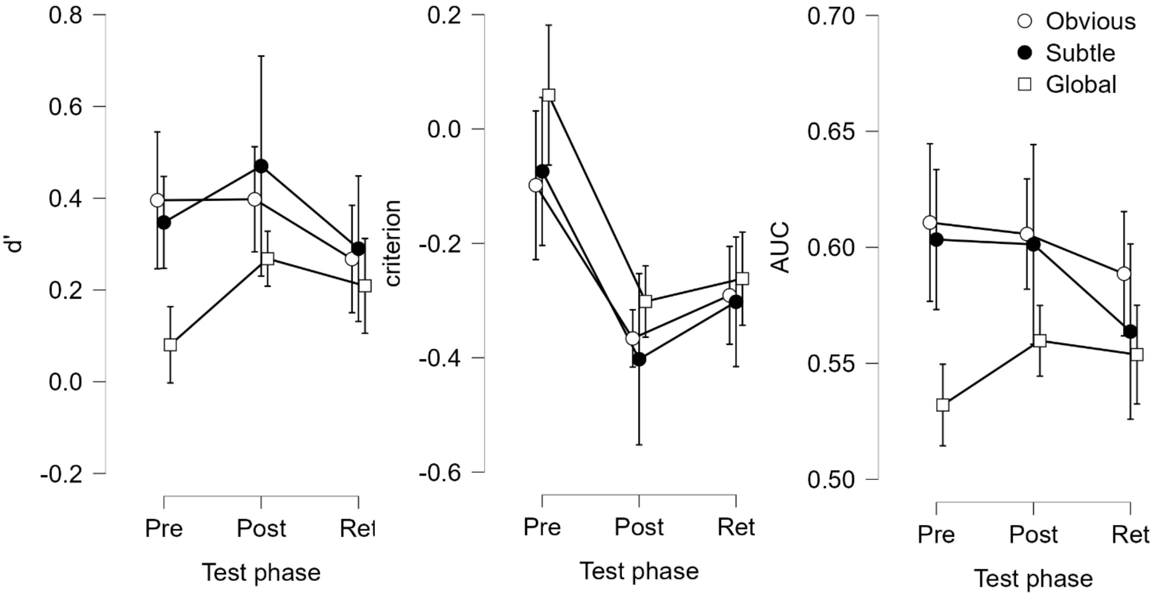


Figure 4.3: *Mean d', criterion, and AUC across test phases (± 95% confidence intervals) for all participants (n=15), plotted separately for each abnormal image type (○ Obvious, ● Subtle, □ Global).*

3x3 repeated measures ANOVAs with the factors testing phase (pre, post, retention) and image type (obvious, subtle, global) were used to investigate the effect of training on d', AUC, and criterion. For d', this showed evidence of an image type effect (F(1.433,20.066)=7.451, p=.007, ηp^2=.347 with Greenhouse-Geisser correction), while the testing phase effect was trending towards significance (F(2,28)=2.816, p=.077, ηp^2=.167) and there was no significant evidence for an interaction effect (F(4,56)=1.455, p=.288, ηp^2=.094). The image type effect was also observed for AUC (F(1.292,18.088)=11.242, p=.002, ηp^2=.445), while there was no significant evidence for a testing phase (F(2,28)=1.191, p=.319, ηp^2=.078) nor interaction effect (F(4,56)=2.005, p=.106, ηp^2=.125). However, AUC was seen as less informative than d' in this experiment, as participants were explicitly instructed to rate trials below 50 for normal and above 50 for abnormal decisions, meaning the cut-off was fixed. Overall, there was no significant evidence of improvements as a result of training, but the trending p-value for d' suggests this might be due to individual variation in learning ability in the testing group, which will be further explored in the following section on performance throughout training.

On the other hand, for criterion, the 3x3 RM-ANOVA showed a significant effect of image type (F(1.433,20.066)=7.451, p=.003, ηp^2=.347 with Greenhouse-Geisser correction) and of testing phase (F(1.352,18.922)=11.501, p<.001, ηp^2=.451 with Greenhouse-Geisser correction), but no evidence for an interaction effect (F(4,56)=1.455, p=.228, ηp^2=.094). Overall, criterion differed significantly between baseline and both post-training (Estimate: -0.319, t(28)=-4.571, p<.001) and retention (Estimate: -0.247, t(28)=-3.542, p=.001). In summary, perceptual training made participants more likely to rate any given trial as abnormal. This could indicate that participants tended to put more weight on negative feedback when they missed a cancerous case than when they incorrectly labelled a normal case as abnormal, causing a shift towards liberal rating bias. Importantly however, participants were not instructed to preferentially avoid one type of error over the other.

Performance throughout training

To investigate performance improvements across training phases, linear Pearsons' correlations were calculated between d' across image types and training phase, numbered 1 through 9 (figure 4). Correlation coefficient varied considerably across participants, with an average of 0.109±0.239. Notably, a positive correlation was found between d' and training phase for 9 participants (average 0.418±0.172), and a negative correlation of the remaining 6 (average -0.357±0.245). This indicated that in the training groups there might be learners and non-learners when dividing participants based on their ability to improve their performance on this specific perceptual learning task.

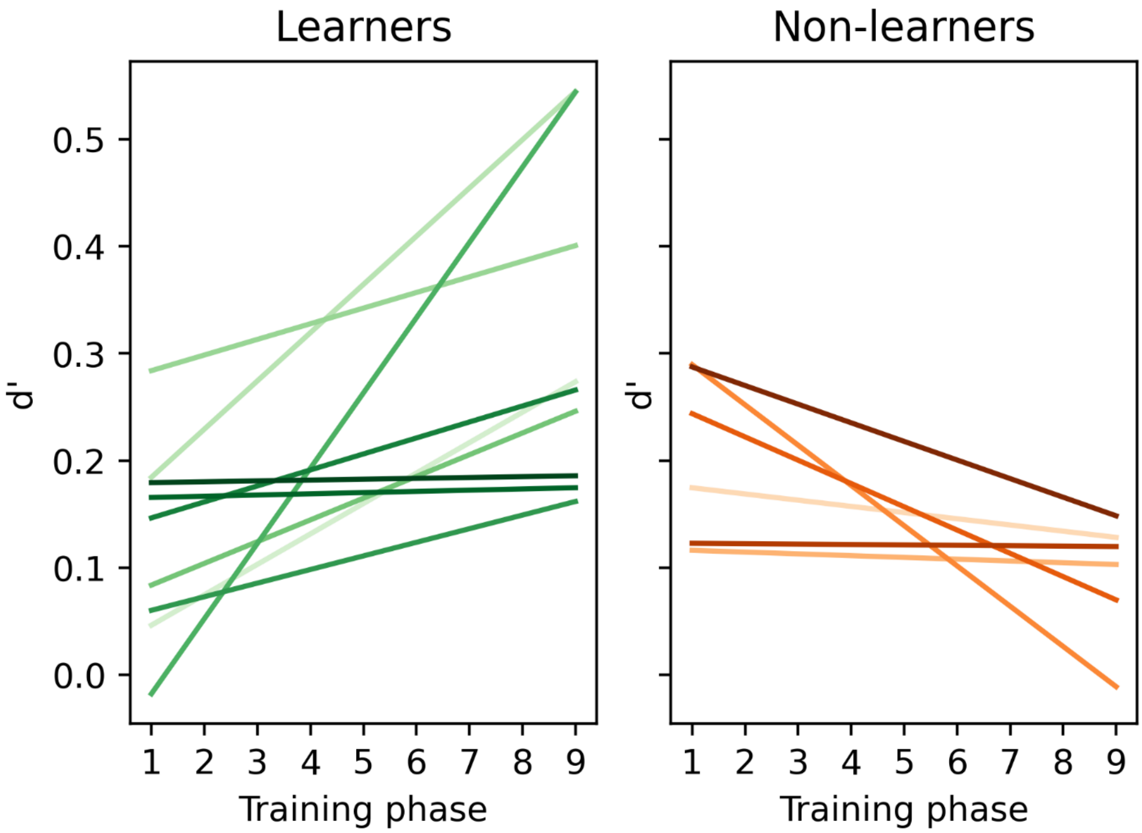


Figure 4.4: *Individual progression of d’ across the 9 training phases, with the learners in green hues in the left plot and the non-learners in orange hues in the right plot.*

To further explore this, analysis of performance measured by d’ was repeated separately for learners and non-learners, to see if the learning during the training phases translated to improved performance on the test phases. For learners, it showed that d’ was affected by both image type (F(2, 16)=13.169, p<.001, ηp^2=0.622) and testing phase (F(2,16)=4.597, p=.026, ηp^2=0.365), without interaction effect (F(4,32)=0.223, p=.924, ηp^2=0.027). Planned comparisons for the testing phase effect with a simple contrast showed that post-training d’ was significantly higher than pre-training levels (difference: .209, t(16)=2.971, p=.009), while this was not the case at retention (difference: .068, t(16)=0.962, p=.350) (see figure 5). On the other hand, for non-learners, d’ was not significantly affected by image type (F(1.091, 5.455)=3.409, p=0.118, ηp^2=0.405) or testing phase (F(2,10)=2.184, p=.163, ηp^2=0.304), but did show evidence for an interaction effect (F(4,20)= 4.254, p=.012, ηp^2=0.460). Post-hoc comparisons for this interaction effect with Holm correction showed that this was driven by significant differences between obvious and subtle pre-training (d’ difference: 0.579, t=4.438, p=.005), and between obvious pre-training and global at retention (d’ difference:4.165, t=4.165, p=.008), both of which do not signify learning of the gist signal. Thus, for learners, d’ improved significantly after training and returned towards baseline levels at retention, suggesting that the learning period was not sufficient for long-term retention. The fact that these effects were not found for the non-learners suggests there is individual variation in people’s ability to obtain the gist of a new category through this type of online training. Analyses for criterion can be found in appendix E.

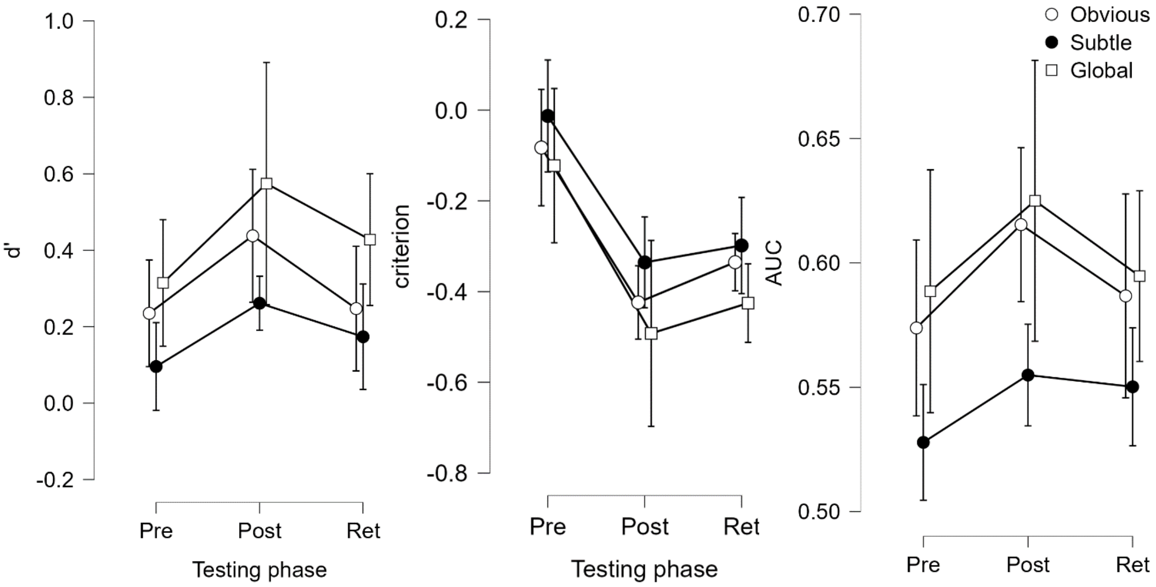


Figure 4.5: *Mean d’, criterion, and AUC across test phases (± 95% confidence intervals) for the learners (n=9), plotted separately for each abnormal image type (○ Obvious, ● Subtle, □ Global).*

These results were compared to those expected under random chance to further ascertain that the split in learning effect was caused by individual differences, rather than any selection bias caused by applying a criterion based on Pearsons’s correlation coefficients. Random rating decisions were simulated across 1000 runs of 15 participants each, calculating their performance on the pre-training and post-training test phase, and each of the 9 training phases, and splitting them into learner and non-learner categories with the same Pearson’s correlations as used for the real observers. The difference between pre- and post-training d’ for ‘learners’ was on average 0.001±0.006, while for the ‘non-learners’ this was 0.002±0.006 (95%CI). This clear lack of improvement in both simulated groups demonstrates that the observed split in learners and non-learners cannot be explained by random effects.

Our results show that nine sessions of perceptual training with global feedback were sufficient to induce a small, but robust increase in gist recognition across all mammogram categories that was significant in the subset of learners. Importantly, this included mammograms that did not contain any localizable lesions, as they were contralateral or prior to the development of a visible lesion, supporting the notion that this was a global signal, and not only the local signal that was captured by the learners. Thus, perceptual exposure paired with global feedback was sufficient to learn the gist of a new category in a group of learners.

However, performance returned towards baseline levels after 7 to 10 days of retention without exposure to mammograms, indicating that the learned signal is poorly retained. While this in itself might seem unfortunate, it is evidence that participants underwent perceptual learning of the global gist signal rather than following any rating strategy based on simpler specific local features, as a strategy would be expected to be retained. Instead, this ‘use it or lose it’ aspect fits with the view of perceptual tuning of the visual system to regularly occurring image statistics in the mammogram texture that must be actively maintained. This finding also converges with findings that radiologists’ gist performance correlates with cases reviewed in a year, not years of experience (Evans et al., 2019). Thus, showing it is recent, continued perceptual experience, and not only (medical) knowledge that allows gist extraction to occur.

Further underlining the importance of perceptual experience rather than knowledge for detection tasks is previous research that showed that pigeons could be trained to recognise cancer-relevant microcalcifications in small patches with above-chance accuracy (Levenson, Krupinski, Navarro, & Wasserman, 2015). The findings give supporting evidence that mammograms contain perceptual features that can be learned through global feedback in implicit learning. However, importantly, the pigeons could not learn to differentiate benign from suspicious masses nor could they detect cancer before onset of any visibly actionable lesions, suggesting a limitation of their perceptual capabilities. Thus, while pigeons could potentially be used as a cost-effective medical image observer to for example investigate impact of technical aspects such as spatial frequency, colour, or other display parameters on performance, as suggested by Levenson et al. (2015), our research instead suggests that training naïve human observers might be a more viable alternative, especially for more complex medical imaging categorisation tasks, as humans can learn a complex gist of abnormality, and are arguably easier to instruct.

Our findings suggest an important role for individual differences in the ability of a participant to learn the gist of abnormality, resulting in a group of learners and of non-learners. This can be compared to the variability in gist extraction performance between individual radiologists, which partially but not fully correlates with recent perceptual exposure, suggesting there are additional individual factors influencing radiologist performance. What’s more, while the learner and non-learner groups were identified based on their learning rate across the nine training phases, further investigation showed that the learner group had an above-chance performance on identifying global abnormalities even before any training had taken place. This is striking, as no local abnormalities are present in these mammograms. Thus, learner participants might already have been more sensitive to disruption of image statistic regularities pre-training than their non-learner counterparts. Previous literature contains numerous examples of individual differences in perceptual sensitivity. Individual differences in performance or sensitivity have been reported across many perceptual domains: in visual search tasks (Brock, Xu, & Brooks, 2011; Sobel, Gerrie, Poole, & Kane, 2007; Wang, Lin, & Drury, 1997), face processing (White & Burton, 2022), scene processing (Pringle, Kramer, & Irwin, 2004), or even low-level visual properties such as colour sensitivity (Emery & Webster, 2019), or auditory temporal processing (Shinn-Cunningham, Varghese, Wang, & Bharadwaj, 2017). In this context, it is not surprising that our participants also showed a range of initial sensitivity to the task.

Furthermore, the observed variability in learning rates between participants in this study matches previous literature. Learning rates differ significantly between individuals across seven perceptual tasks in the visual and auditory domain, such as Vernier acuity, face view discrimination, auditory frequency discrimination (Yang et al., 2020). Importantly, the contribution of participant-specific (36.8%) factors is approximately equal to the task-specific (~38.6%) factors influencing learning rate, underlining the large impact individual differences can have on learning rates across tasks. Individual differences in learning rates have also been demonstrated in spatial learning in virtual environments (Waller, 2000).

So, learners might have been predisposed to have enhanced sensitivity to structural regularities, resulting in above-chance pre-training performance, and subsequently further improved their performance after training. This predisposition might be innate, or due to previous experiences. Innate factors can influence performance and learning, as shown by positive correlations between learning rates and cortical thickness in the posterior parietal cortex (PPC) and motion sensitive area MT+ of the V5 for a motion discrimination visual search task (Frank, Reavis, Greenlee, & Tse, 2016), and similarly for the left fusiform face area in a face view discrimination task (Bi, Chen, Zhou, He, & Fang, 2014). Furthermore, previous experiences such as gaming activity might influence brain plasticity and increase general perceptual learning ability (Bavelier, Green, Pouget, & Schrater, 2012; Bejjanki et al., 2014). Another factor that might have made learners more likely to learn the gist signal could be differences in strategy. It is possible learners were tuned to a more global strategy compared to non-learners who might have focused more on local signals. Previous research suggested that learners and non-learner groups utilized different strategies while being trained on a difficult grating orientation task (Dobres & Seitz, 2010). Further research could further explore differences in initial sensitivity, neural markers, and strategies employed by learners and non-learners in a gist learning task.

The fact that non-learners did not show improvement in their ability to detect the gist of abnormality might also be related to the duration of training. Perhaps these non-learners would have shown improvement after additional training sessions, where this was not the case after nine sessions, for example due to a slower learning rate or an initial maladaptive learning strategy. Interestingly, in Hegdé's (2020) design participants trained until a predefined performance level, which took anywhere between 288 to 936 trials, a factor 3.25 difference, providing evidence for the existence of a range in individual learning times. However, they also reported that 4 participants left part-way through the experiment, leaving it up to question if/when these participants would have reached the predefined performance level. Thus, while non-learners in the current study might have lacked the aptitude or capacity to learn the new gist category in the task format, they might have simply required further perceptual training before they would have been able to increase their performance. Future research could employ a predefined performance threshold similar to Hegdé's (2020) design to gain further insight into the variation in perceptual exposure needed to learn the gist of a new category.

As briefly discussed above, our results corroborate the main findings of a previous training study that showed that implicit learning through auditory global feedback could induce learning of visual patterns of medical abnormality in a free-viewing task (Hegdé, 2020). Notably, however, the learning described by Hegdé occurred much faster, after an average of ~600 trials, and resulted in a higher performance of d' 2.5. One factor that might explain the difference in performance is differences between the stimuli. The abnormal mammogram cases used by Hegdé and colleagues contained localizable, and obvious abnormalities with one region of interest at least 200 pixels wide, whereas the current study used a larger variety of mammograms, containing obvious or subtle abnormalities, or even only global signals of abnormalities with no visible lesions. Another factor is likely the difference in tasks, as free-viewing tasks are generally easier than rapid gist extraction tasks. The same effect can be observed for medical experts, as their performance in laboratory free-viewing experiments reached d' of 2.5 for chest radiographs (Kundel & Nodine, 1975), and d' of 1.9 for mammograms (Evans, Birdwell, & Wolfe, 2013), whereas gist extraction performance reached a d' of 1 for chest radiographs (Kundel & Nodine, 1975), and a d' of 1 (250 ms) and 1.14 (500 ms) for mammograms (Evans, Georgian-Smith, et al., 2013). Thus, while the current performance did not reach the same levels as observed by Hegdé, this can be explained by differences in task and stimuli.

A general limitation of the current study was the duration of the perceptual training. This had to be limited for viability of the research, but consequently, naïve participants did not reach the same performance levels as expert radiologists. After training, learners reached an overall average d' of 0.43, which is close to a medium effect size. Learners did not quite reach the d' of 0.88-1.14 reported for expert radiologists on obvious/subtle lesions in similar experiments (Evans et al., 2019; Evans, Georgian-Smith, et al., 2013; Evans et al., 2016) , but learners' post-training performance on mammograms with global abnormalities (d' 0.57) was remarkably similar to performance of expert radiologists on comparable cases in different experiments, such as a reported d’ of 0.59 on contralateral mammograms (Evans et al., 2016) and a d’ of 0.21 on priors (Evans et al., 2019), demonstrating the validity of the learning. The difference in performance on visible actionable lesions difference could be partially the result of specific medical knowledge, or it could reflect the differences in the magnitude and duration of perceptual training. While medical experts do not routinely perform gist rating tasks, they have years of real-world exposure to the stimuli with on average of up to 4000 read mammograms a year in which they focus on detecting visible abnormalities, which would involve an early non-selective stage of visual processing shaping their knowledge of the gist of abnormality.

In the current study, participants became significantly more liberal in their ratings after training, meaning they were more likely to label any given mammogram as abnormal than before. This could potentially reflect a self-imposed criterion in which participants tried to avoid missing any cancerous cases at the cost of more false alarms – although it is important to note that no such instruction was given in the experiment. A move to a more liberal decision criterion may indicate the participants’ feeling of familiarity with images after training and thus more willingness to report a signal but it is more likely a result of early stages of learning-related changes in developing perceptual expertise as observed in some perceptual training studies (Aberg & Herzog, 2012; Palmeri, Wong, & Gauthier, 2004; Xu, Rourke, Robinson, & Tanaka, 2016).

Another interesting observation was the change in rating time, as participants became significantly faster after training. This increase in rating speed could potentially be a marker of the development of expertise. Decreases in reaction times have previously been described to occur in naïves learning to categorize aerial photographs (Lloyd, Hodgson, & Stokes, 2002) and training on face-like artificial object categorisation (Wong, Palmeri, & Gauthier, 2009). However, other studies reported no consistent changes in reaction time after training subordinate and superordinate level bird categorization (Devillez et al., 2019; Jones et al., 2020). Additionally, interpretation of our findings is complicated by the fact that this study used a 0-100 rating scale, operated using a mouse. Thus, it is also possible that participants habituated to using the slider and became faster at reaching their desired rating score. Overall, this increase in rating speed is an interesting observation, but a different design is needed to be certain that this effect is caused by changes in decision making time rather than adeptness at the rating task.

Deep Neural Network performance in detecting cancer

With the aim to further understand how gist expertise develops we examined whether a DNN, analogous to human implicit learning, was able to capture the same image statistics that humans might be using when learning to detect gist of the abnormal. We use a DNN specifically developed for malignancy detection, which was pre-trained on mammograms, to evaluate its performance on the mammograms we used for training and testing our human learners. This is assessed using the DNN's calculated malignancy probability scores (Wu et al., 2019), the probability that that mammogram contained a malignant abnormality. Each unilateral mammogram in the training image set and test image set were scored by both the single breast classifier image-only (SBC) and SBC + heatmaps (SBC+HM) DNN. The DNN also provided benign probability scores, the probability that a mammogram contained a benign abnormality, which showed the same pattern of results as discussed below (see appendix F).

Histograms of DNN malignancy probability scores show more overlap between the normal and global cases, than between the obvious/subtle and normal cases (Fig. 6), indicating that both the SBC and SBC+HM were less able to distinguish global and normal from each other. The finding illustrates an apparent difficulty for the SBC and SBC+HM to distinguish the global gist signal of cancer compared to the visible obvious and subtle cancers.

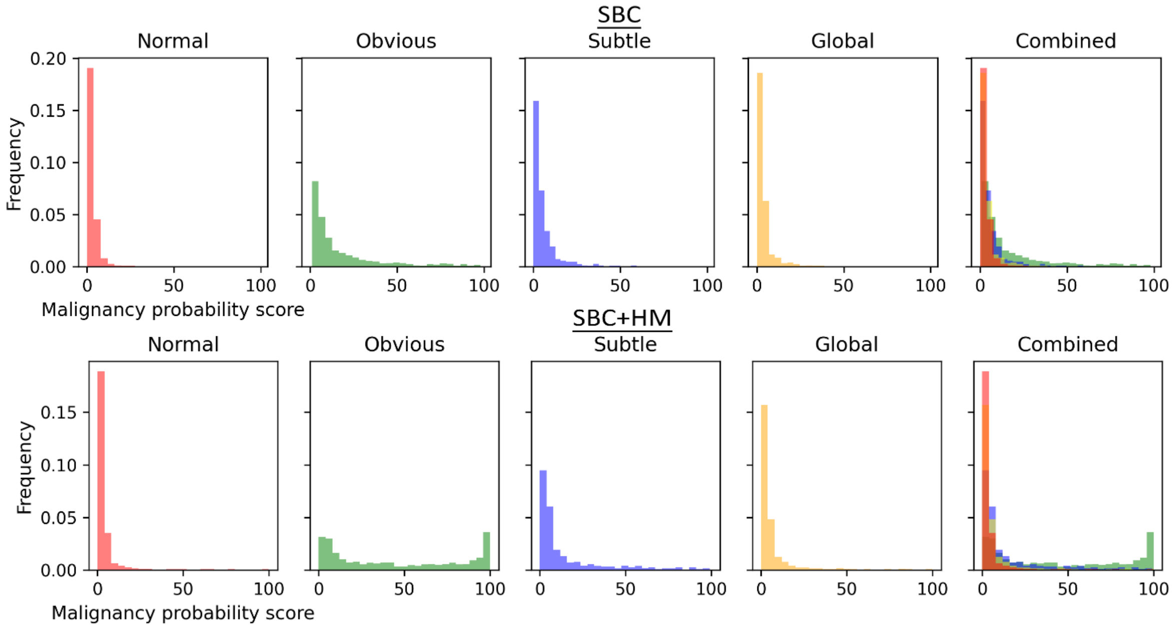


Figure 4.6: *Distribution of Single Breast Classifier (SBC) and SBC+Heatmap (SBC+HM) malignancy probability scores on the full image set of mammograms split into 25 bins for each of the image type categories, with a combined plot showing the overlap between normal (red), obvious (green), subtle (blue), and global (yellow) scores.*

Similarly, AUC calculations (table 1) show that the SBC and SBC+HM both performed well in discriminating the obvious and subtle mammograms from the normal mammograms on malignancy probability, whereas AUC dropped considerably for the global mammograms, although it did remain above chance levels for all except the malignancy-SBC on the global mammograms in the test set. The increase in AUC for SBC+HM shows that heatmaps improved the DNN's ability to detect the probability of malignancy in mammograms, especially in more subtle cases. These results on our mammography image lend support to the reported increase in performance with the added heat-map described in the original publication (Wu et al., 2019).

Table 4.1: *AUCs for malignancy probability scores for the SBC and SBC+HM for obvious, subtle, and global mammograms versus the group of normal mammograms. This is calculated for the training set and the test set separately. Square brackets contain the lower and upper bands of 95% CIs.*

	Training set		**Test set**	
	SBC	**SBC+HM**	**SBC**	**SBC+HM**
Obvious	0.839 [0.842-0.854]	0.897 [0.885-0.909]	0.844 [0.772-0.916]	0.885 [0.824-0.946]
Subtle	0.689 [0.668-0.710]	0.738 [0.719-0.757]	0.701 [0.599-0.603]	0.803 [0.720-0.886]
Global	0.582 [0.563-0.601]	0.598 [0.579-0.617]	0.505 [0.408-0.602]	0.683 [0.596-0.770]

Most critically, the low or even at-chance performance (AUC: 0.505 SBC on test set) on the globally abnormal mammograms shows that mammograms with the global signal of abnormality are especially obscure and difficult to detect. This adds to the significance of our finding that human observers were able to learn to detect abnormalities in these mammograms, performing above chance on the test set with which the SBC struggled severely. It also demonstrates that the chosen test set was representative of, or potentially even more difficult than, the overall mammography dataset, and learning was not a result of coincidentally easier stimuli in the test set.

Next, a direct comparison of human and SBC scores was made to see if similar image statistics might be used by human observers and machine learning models. This was done by correlating the average rating from the ‘learner’ group of observers to the malignancy probability scores of the SBC and SBC+HM. Spearman’s rank correlations were performed between the DNN malignancy probabilities and the average of the human learner scores given pre- and post-perceptual training (table 2). Before perceptual training, the correlation between SBC malignancy and human scores was non-significant (p=.137), while the correlation between SBC+HM and human scores was (p=.005). At post-training test, the average human score across the 200 test mammograms correlated significantly with both the SBC and SBC+HM malignancy and benign scores (all p<0.01, see table 2). Comparing pre- and post- perceptual training correlations showed that the correlation coefficient increased after the human observers completed their perceptual training. After training, human scores more closely agreed with the classifier judgements - mammograms that were judged as more abnormal by humans also received higher malignancy probability scores.

Table 4.2: *Spearman's rank correlations between the average human learner score pre- and post-training of human observers, and the SBC/SBC+HM malignancy probability scores.*

		Pre-training		**Post-training**		
		Correlation	**p-value**	**Correlation**	**p-value**	**Difference**
SBC	Malignant	0.105	0.137	0.207	0.003	0.102
SBC+HM	Malignant	0.198	0.005	0.318	0.000	0.119

The finding that agreement between human and SBC scores increased after training has interesting implications. It suggests that the gist of abnormality signal learned by human observers during perceptual training is partially captured by the DNN as well. This adds validity to our findings, as the human observers learned signals that were also detected by an 'expert' in the form of a DNN, demonstrating they were able to learn image features of abnormality. Additionally, the finding that correlation coefficient was markedly higher for the SBC+HM (0.318) than SBC (.207) suggests that the added heatmap might capture additional perceptual features used by the trained human observers. This suggests that the SBC+HM and similar deep neural networks could be used to investigate the perceptual features in mammograms contributing to the gist signal, for example by performing network dissection, a technique where layers of the network are investigated to extract the content that is activating nodes in these layers (Bau et al., 2020).

4.7. Conclusion

In conclusion, perceptual training with global feedback can result in the learning of the gist of a new category, although there are individual differences in both pre-training sensitivity to global structural regularities and ability to further learn the gist signal, and the new gist signal is poorly retained if exposure is not maintained. This suggests that gist categorisation might be a case of 'use it or lose it', although retention or complete tuning of the visual system to a new category might be obtained after extended exposure. The exposure in our study only amounted to approximately 9 hours task time, and 6470 instances viewed with feedback, which is substantially less than in real world learning of gist categories.

Furthermore, human perceptual expertise on difficult, ambiguous cases containing only global signals of abnormality (contralateral, prior) is still not matched by state-of-the-art neural networks, as indicated by the markedly lower, or even at-chance performance of the DNN on mammograms with global abnormalities that human observers were able to learn in our perceptual training paradigm. The global signal of abnormality is extremely difficult to detect and requires considerable perceptual expertise. On the other hand, we also observed an increase in agreement between the human observers and DNN after perceptual training, which indicates a potential overlap in image statistics used to classify mammograms as normal or abnormal. Finding out what these image statistics are could teach us more about the gist of abnormality and could help find ways to improve image filtering for both human observers and machine learning models. Together, these findings solidly emphasize the need for continued research into medical perceptual expertise with human observers in its own right, especially into more ambiguous global signals that would be vital for early cancer detection. But it also reinforces the need of combining these lines of research with the thriving field of machine learning research, especially since recent research has suggested benefits of combining radiologists' gist ratings with machine learning models to reach higher levels of performance than either could on their own (Gandomkar et al., 2021; Wurster et al., 2019).

We based our study on drawing a clear parallel between scene gist and the gist of abnormality in radiographs, and it would be beneficial to generalize the current results on learning to a wider area of gist extraction. The parallels between the two types of gist extraction would imply that the current findings of implicit learning should generalize to the learning of scene gist as well. However, as far as the authors are aware, this area has not yet been investigated in the known literature. A potential avenue to answering this question for scene gist could be developmental research with young children, especially as previous research has shown that infants already exhibit signs of statistical learning (Fiser & Aslin, 2002b). However, previous research on the development of rapid perceptual processing is very limited (but see Sweeny, Wurnitsch, Gopnik, and Whitney (2015). Overall, developmental research often suffers from complications, such as communication of task instructions or difficulties in directing attention, a lack of control over previous exposure, individual differences, and other developmental processes occurring at the same time (Johnson, 2011; Maurer, 2013). These factors make it less suitable to investigate the acquisition of the gist of a novel category.

Overall, the current study shows a strong case for how implicit learning would allow the learning of a new category of any gist, including scenes. What is more our finding that gist extraction abilities can develop separately from medical knowledge reinforces the viability of the idea, suggested by Voss, Kramer, Basak, Prakash, and Roberts (2010), of using trained naïve observers, not to 'usurp' radiologists' ratings, but to create a more accessible 'model observer' to use for further dissemination of the gist of abnormality signal. This training regime can be used for training of novice radiologists and screening radiographers or even as a refresher training for expert radiologists who over their careers see a considerable reducing of cases they read. Further research is needed to measure the effectiveness of our training paradigm on these populations, and to explore explanatory parameters for individual differences in pre-training performance, learning ability, and learning rate/speed, for example by investigating the potential variation in length of perceptual training required to achieve perceptual learning across different participants.

4.7. List of Abbreviations

AUC: Area Under the Curve

ROC: Receiver Operator Curve

SBC: Single breast classifier

SBC+HM: Single breast classifier plus heatmap

4.8. Open Practices Statement

The datasets generated and analysed during the current study are available on our OSF repository, together with the Python scripts needed to extract the performance measures. The data is available under Creative Commons Attribution-NonCommercial-ShareAlike 2.0 UK: England & Wales (CC BY-NC-SA 2.0 UK).

Mammograms were selected from the Complex Cognitive Processing lab database of stimuli, which can be shared with other researchers upon request to the last author (K.K. Evans), and from the OPTIMAM database, which can be accessed by requests for research purposes (medphys.royalsurrey.nhs.uk/omidb/getting-access/).

Experiments were not preregistered.

4.9. Appendices

Appendix A: Mammographic descriptors of obvious and subtle cases

Ductal Carcinoma In Situ (DCIS) grade can be classified as high, intermediate, or low. Percentages of DCIS grades in obvious and subtle mammograms can be found in table 3). Tumour surfaces can be positive, negative, or borderline (not strongly + or -) for Human Epidermal Growth Factor Receptor 2 (HER-2), and positive or negative for Estrogen and Progesterone receptors (see table 4 for percentages in the obvious and subtle cases). The presence or absence of these receptors in the tumour can impact both the cancer severity and viable treatment options. For example, so-called triple-negative cancers, without HER-2, Progesterone, and Estrogen Receptors, currently lack of approved targeted therapy and overall have poorer long-term outcomes (Sharma, 2016).

Table 4.3: *Percentage of mammograms with a high, intermediate, low, or unassessed DCIS grade for the obvious and subtle subsets of the image set. Where descriptors were not available in the OPTIMAM database, the mammogram was classified as unassessed.*

	Obvious	**Subtle**
High	26.3	14.7
Intermediate	25.2	21.8
Low	8.7	7.9
Unassessed	39.7	55.6

Table 4.4: *Percentage of mammograms that were positive, negative, borderline, or unassessed for different receptor groups, HER-2, Progesterone, and Estrogen receptors for the obvious and subtle subsets of the image set. Where descriptors were not available in the OPTIMAM database, the mammogram was classified as unassessed.*

	HER-2		**Progesterone Receptor**		**Estrogen Receptor**	
	Obvious	**Subtle**	**Obvious**	**Subtle**	**Obvious**	**Subtle**
Positive	5.4	3.5	66.5	55.8	58.4	49.7
Negative	58.8	51.1	6.3	3.3	11.9	7.4
Borderline	0.5	0.1	N/A	N/A	N/A	N/A
Unassessed	1.2	1.7	1.4	1.6	1.6	1.7

Appendix B: Engagement and attention in training phases

As mentioned in the main document, participant routinely used the spacebar to manually continue to the rating screen before reaching maximum viewing time (2.5 seconds) in the first training phase. This occurred on average on 234 out of 720 trials (95% CI: 138-330), indicating active engagement with the task instruction to view the mammogram until they formed a first impression to base their rating on. As a result, both average and maximum viewing time rapidly decreased, as is plotted in figure 7.

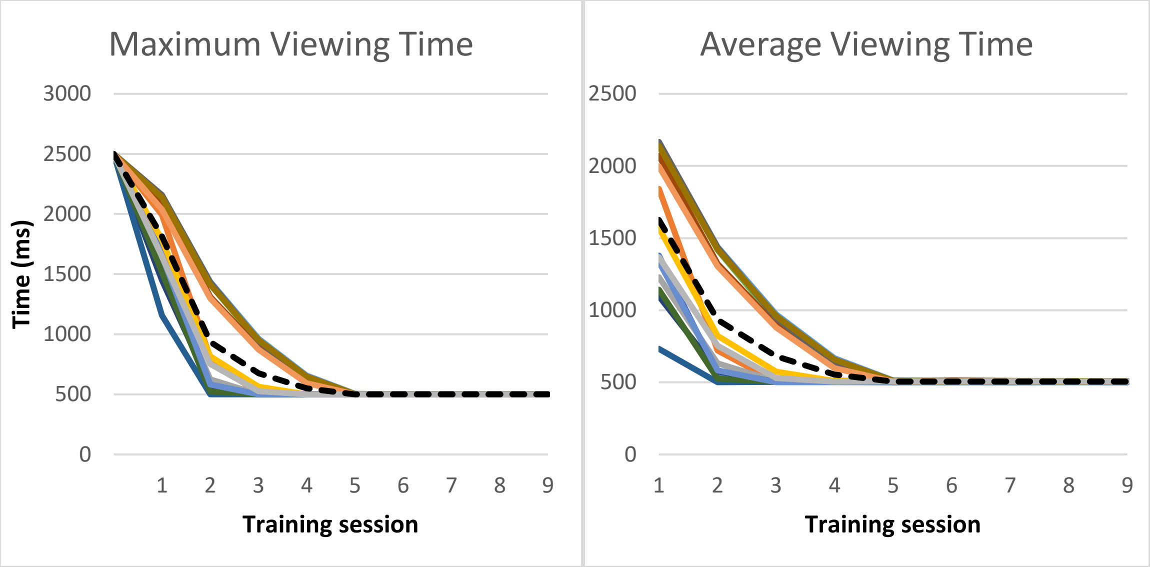


Figure 4.7: *Maximum and average viewing time in milliseconds per participant at the end of each training phase. Maximum viewing time is calculated for the fourth block of the session. Individual lines are plotted, while the dashed black line represents the group average.*

Additionally, participants viewed the feedback screen for an average of 741±72.4 ms per trial across the 9 training phases, which is estimated to be sufficient to perceive the "right or wrong" global feedback, due to the colour-coded and regular nature of the feedback text combined with the recency of the rating choice as feedback was shown immediately after confirming the rating. In conclusion, there was clear evidence of attention to and engagement with the training phases.

Appendix C: Log-linear likelihood ratios ROC curves

To evaluate whether individual participants' performance was significantly above chance the AUCs of log-linear likelihood (LLR) ROCs were compared to the AUC of the 95th percentile AUC of simulated ROCs. As shown in table 5, the number of participants that performed above chance increased from 5 to 11 overall after training, an increase driven by an increase from 1 to 7 out of 9 learners, while no change was observed for non-learners. This analysis shows that training caused most participants to outperform a very strict definition of chance levels, especially the sub-group of learners, in line with the significant testing phase effect observed for d'.

Table 4.5: *Number of participants performing at above chance levels (real AUC > 95th% simulated AUC) at each testing phase, split up for learners, non-learners, and total.*

	Pre-training	Post-training	Retention
Learners	1 (11.11%)	7 (77.77%)	5 (55.55%)
Non-learners	4 (66.66%)	4 (66.66%)	2 (33.33%)
Total	5 (33.33%)	11 (73.33%)	7 (46.66%)

Appendix D: Effect of training on rating time

To evaluate if perceptual training affected participants' decision-making speed, a 4x3 repeated measures ANOVA was conducted on the average rating time with the factors image type (normal, obvious, subtle, global) and testing phase (pre-training, post-training, and retention). Average rating time was significantly affected by test phase (F(1.08,15.10)=25.590, p=<.001 with Greenhouse-Geisser correction, η_p^2=.646), but not by image type (F(3,42)=1.631, p=<.001, η_p^2=.104), nor was there evidence for an interaction effect (F(6,84)=0.594, p=<.001, η_p^2=.041). Rating time went down significantly after training compared to pre-training (difference=-1291 ms, p<.001) and remained that way at retention (difference=-1158 ms, p<.001), as shown by a simple contrast planned comparison. Due to the lack of evidence for an image type effect, the main effect of testing phase on average rating time is visualized in the bar graphs in figure 8. The same pattern persisted for median rating time.

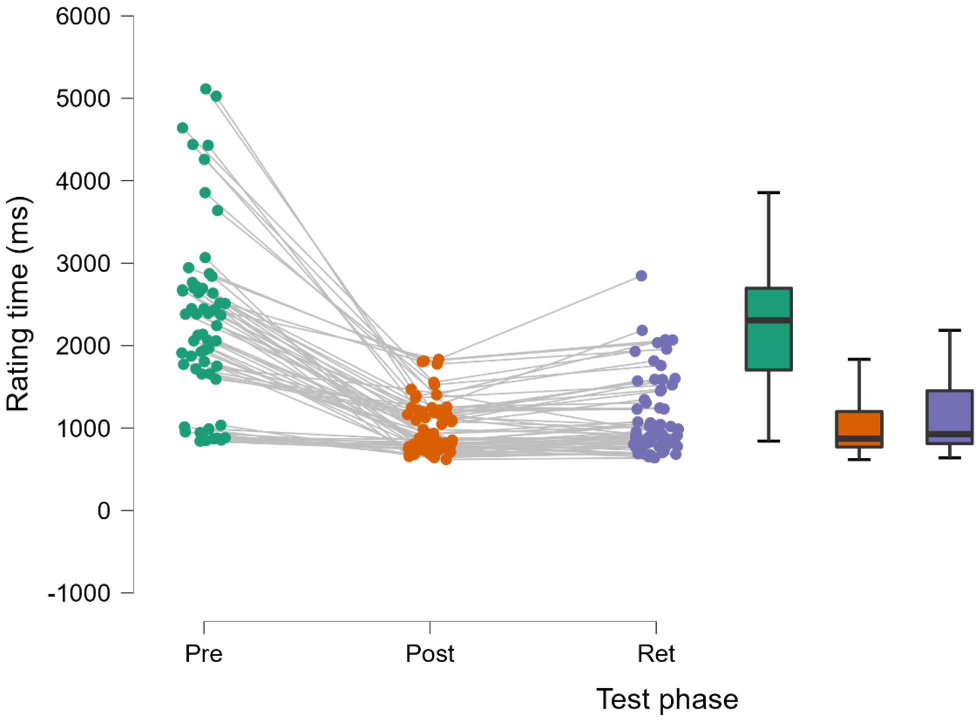


Figure 4.8: *Individual average rating times are shown at pre-training (pre, green), post-training (post, orange), and retention (ret, purple) testing phases, as connected dot-clouds per participant-image type combination and boxplots to show both individual patterns and the population distributions.*

Median rating time was also evaluated using a 4x3 repeated measures ANOVA with the factors image type (normal, obvious, subtle, global) and testing phase (pre-training, post-training, and retention). Median rating time was significantly affected by test phase (F(1.04,14.49)= 24.590, p=<.001 with Greenhouse-Geisser correction, η_p^2=.637), but not by image type (F(3,42)=1.307, p=.285, η_p^2=.085), nor was there evidence for an interaction effect (F(6,84)=0.284, p=.943, η_p^2=.020). Rating time went down significantly after training compared to pre-training (difference=-1160 ms, p<.001) and remained that way at retention (difference=-1069 ms, p<.001), as shown by a simple contrast planned comparison and visualized in figure 9. Thus, participants took significantly less time to make rating decisions after completing their training.

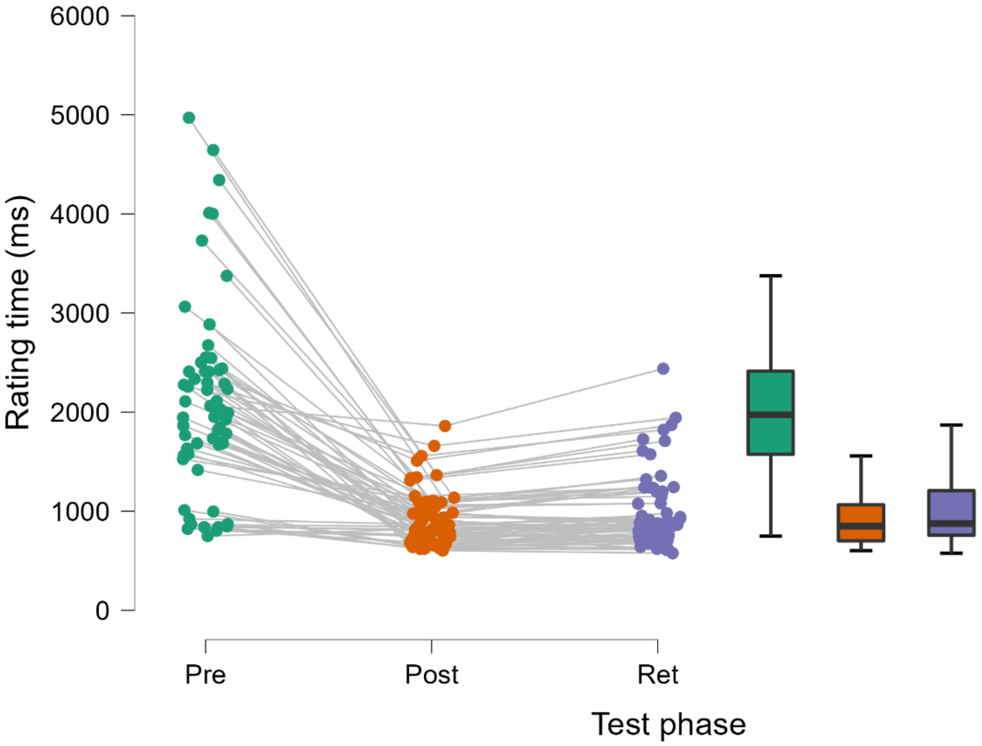


Figure 4.9: *Individual median rating times are shown at pre-training (pre, green), post-training (post, orange), and retention (ret, purple) testing phases, as connected dot-clouds per participant-image type combination and boxplots to show both individual patterns and the population distributions.*

Appendix E: Criterion for learners and non-learners

For learners, it showed that criterion was affected by both image type (F(2, 16)=13.169, p<.001, ηp2=0.622) and testing phase (F(2,16)=12.509, p<.001, ηp2=0.610), without interaction effect (F(4,32)=0.223, p=.924, ηp2=0.027). Planned comparisons with a simple contrast showed that post-training criterion was significantly lower (more liberal) than pre-training levels (estimate: -.345, t(16)=4.703, p<.001), and remained this way at retention (estimate: -.280, t(16)=3.826, p=.001). However, for non-learners, criterion was not affected by image type (F(1.091, 5.455)=3.409, p=.118, ηp2=0.405) nor testing phase (F(2,10)=2.002, p=.186, ηp2=0.286), but did show an interaction effect (F(4,20)=4.254, p=.012, ηp2=0.460).

Appendix F: DNN probability of benign abnormality

Histograms of DNN benign probability scores show more overlap between the normal and global cases, than between the obvious/subtle and normal cases (Fig. 10), indicating that both the SBC and SBC+HM were less able to distinguish global and normal from each other. Similar to the malignancy probability scores, this again illustrates difficulty for the SBC and SBC+HM to distinguish the global gist signal of cancer compared to the visible obvious and subtle cancers.

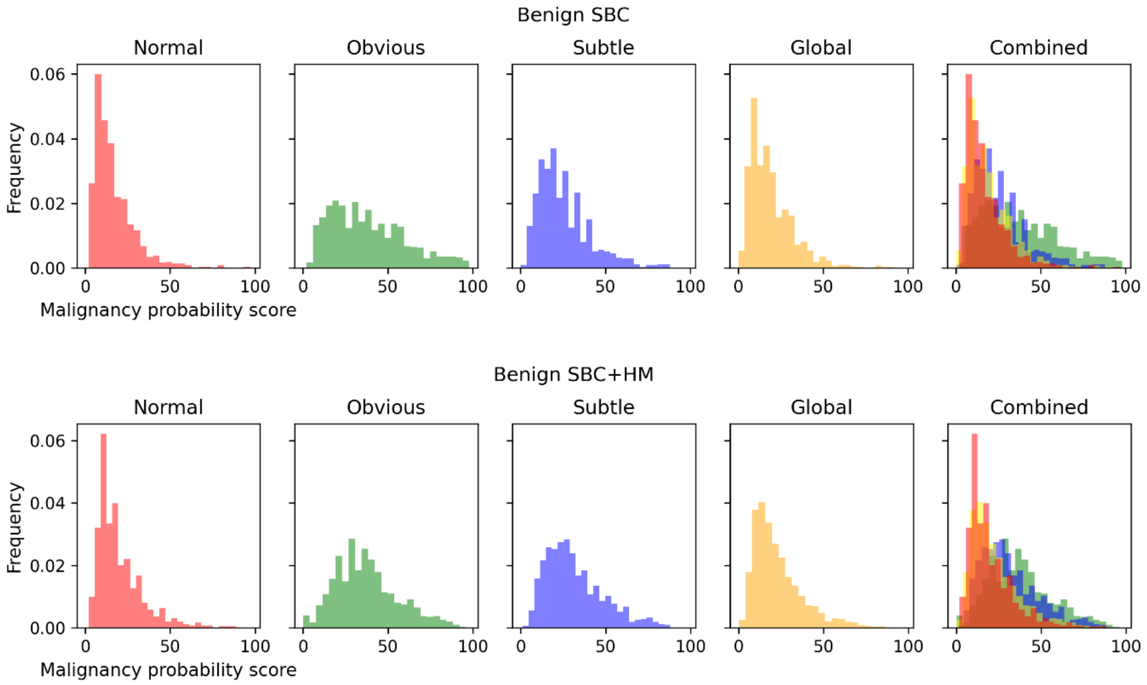


Figure 4.10: *Distribution of Single Breast Classifier (SBC) and SBC+Heatmap (SBC+HM) benign abnormality probability scores on the full image set of mammograms split into 25 bins for each of the image type categories, with a combined plot showing the overlap between normal (red), obvious (green), subtle (blue), and global (yellow) scores.*

AUC calculations for benign probabilities (table 6) show that the SBC and SBC+HM both performed well in discriminating the obvious and subtle mammograms from the normal mammograms on malignancy probability, whereas AUC dropped considerably for the global mammograms, although it did remain above chance levels (~0.55).

Table 4.6: *AUCs for the probability of benign abnormality for the SBC and SBC+HM for obvious, subtle, and global mammograms versus the group of normal mammograms. This is calculated for the training set and the test set.*

	Training set		**Test set**	
	SBC	**SBC+HM**	**SBC**	**SBC+HM**
Obvious	0.817 [0.801-0.833]	0.818 [0.802-0.834]	0.818 [0.739-0.897]	0.785 [0.698-0.872]
Subtle	0.701 [0.681-0.721]	0.670 [0.649-0.691]	0.670 [0.563-0.777]	0.764 [0.673-0.855]
Global	0.569 [0.550-0.588]	0.555 [0.536-0.574]	0.555 [0.459-0.651]	0.547 [0.451-0.643]

Spearman's rank correlations between the DNN malignancy probabilities and the average of the human learner scores given pre- and post- perceptual training (table 7) showed a marked increase in correlation after perceptual training. After training, human scores more closely agreed with the classifier judgements - mammograms that were judged as more abnormal by humans also received higher benign abnormality probability scores.

Table 4.7*: Spearman's rank correlations between the average human learner score pre- and post-training, and the SBC/SBC+HM probabilities of benign abnormality.*

		Pre-training		**Post-training**		
		Correlation	**p-value**	**Correlation**	**p-value**	**Difference**
SBC	**Benign**	0.286	0.000	0.373	0.000	0.087
SBC+HM	**Benign**	0.280	0.000	0.402	0.000	0.122

Appendix G: DNN correlation with non-learners

Correlating SBC scores with the average ratings of the learner group showed that the correlation went up post-training. The same correlations were performed for the average ratings of the non-learners (table 8).

Table 4.8: *Spearman's rank correlations between the average human non-learner score pre- and*

		Pre-training		**Post-training**		
		Correlation	**p-value**	**Correlation**	**p-value**	**Difference**
SBC	***Malignant***	0.158	0.026	-0.038	0.592	-0.196
SBC+HM	***Malignant***	0.301	0.000	0.099	0.161	-0.201
SBC	***Benign***	0.310	0.000	0.104	0.142	-0.206
SBC+HM	***Benign***	0.342	0.000	0.105	0.140	-0.237

post-training, and the SBC/SBC+HM probabilities of malignant or benign abnormality.

These results show two things. Firstly, before training, the correlation between SBC+HM malignancy predictions and the non-learners was 0.301, compared to the 0.198 of learners. This suggests that the non-learners might have started out sensitive to part of the same signals used by the SBC, and especially the SBC+HM. Potentially, this could be caused by more focus on localized signals, as implied by the increased correlation with the added heatmap – which adds scrutiny to local features. Secondly, the correlation between non-learner and SBC goes down after training, and becomes non-significant for all four comparisons. This was unexpected, and could be the result of a maladaptive learning strategy, where non-learners incorrectly establish certain perceptual features as normal/abnormal and this leads them to not only fail at learning, but additionally diverge from the SBC predictions. However, since this dataset only contained six non-learners, a larger, more structured approach would be needed to further investigate potential maladaptive strategies in such a perceptual learning task.

Chapter 5: The neural signature of the gist of medical abnormality

5.1. Abstract

Rapid, global extraction of visual information gives observers access to the gist (the general categorical information) of an image, usually a scene. Through the same gist extraction mechanism, medical experts can rapidly distinguish abnormal from normal medical images. Previous research showed importance of the occipito-parietal regions and perhaps the visual P2 event related potential (ERP) for scene gist extraction, however, it is unknown whether medical gist extraction evokes similar activity patterns. In this exploratory study, five experienced radiologists performed 2-AFC ratings on normal and abnormal mammograms, while EEG was recorded. Activity patterns were investigated across a wide range of ERPs and brain areas using single-subject bootstrapping. Differential activity between frontal cluster, suggesting a whole-brain representation of the gist of medical abnormality. The involvement of areas beyond the occipital and parietal regions, suggest that the neural signature of medical gist is characterized by distributed activity. Differential activity amplitude correlated positively with performance, overall medical experience, and recent perceptual experience. This suggests that the neural signature of medical gist categories in an individual might be influenced by their medical and perceptual experience, and it might be associated with performance. There was also evidence for individual differences across radiologists, most notably in the direction of differential activity in the N1, P2, and P600. Overall, the findings suggest that the gist of abnormality is extracted in a network of regions, with potentially individual differences in how medical gist categories are represented. Further research is needed into individual differences in (medical) gist extraction and the functional role of areas in the observed network of activity.

Key words: Gist extraction, medical expertise, EEG, neural correlates, visual processing, differential activity, individual differences, mammography

5.2. Introduction

Gist extraction is a process in which global information is rapidly and non-selectively used to inform a general sense (gist) of a visual stimulus. Indeed, people are able to detect the general category of a scene within 30 ms (Joubert et al., 2009; Joubert et al., 2007), without needing selective attention or focal vision (Boucart et al., 2013; Larson & Loschky, 2009; F. F. Li et al., 2003; Rousselet et al., 2004) (see Chapter 1 for more detail). Gist extraction is generally thought to occur through a combination of processes such as the extraction of spatial structural regularities, summary statistics, and distributed features.

Interestingly, research with medical experts has shown that they are able to detect the (ab)normal nature of a medical image within 100-250 ms in chest x-rays (Kundel & Nodine, 1975), PAP smears (Evans, Georgian-Smith, et al., 2013), and mammograms (Patrick C. Brennan et al., 2018; Evans, Georgian-Smith, et al., 2013; Evans et al., 2016), without being able to localize the abnormalities as gist extraction only provides global categorical information. The ability of medical experts to extract this gist of medical abnormality varies considerably across individuals, but correlates with their recent perceptual exposure rather than years of experience in the field (Evans et al., 2019), suggesting that it is indeed a perceptual ability rather than the result of medical knowledge. Furthermore, a perceptual training protocol was able to induce significant learning of the gist of abnormality in a sub-group of learners in Chapter 4, which fits with the perceptual nature of the gist of abnormality signal. Overall, previous findings draw clear parallels between the gist extraction processes for medical abnormalities and for scenes.

Early divergence in neural activity has been observed as a result of scene gist perception. When shown for 500 ms each, scenes evoked higher P2 (200-320 ms) amplitude than objects or faces, across anterior, central and posterior brain regions, with maximal scene selectivity in the posterior-lateral cluster (Harel et al., 2016). Magnetoencephalography (MEG) similarly showed higher M1(100-130 ms) amplitude for outdoor scenes than faces in the medio-occipital region (Rivolta et al., 2012). Thus, posterior, occipital regions might be important for scene gist categorisation, as they show differential activity between scenes and objects or faces.

However, directly comparing neural activity between scene categories gives more interesting insights into the potential neural correlates for the gist of abnormality. Various studies compared neural activity between animal and non-animal scenes. These studies observed early differences between animal and non-animal scenes, starting at 150-170 ms, characterized by frontal negativity, occipito-temporal positivity, and a potential temporal negativity for distractor/no-go scenes without animals (Antal et al., 2001; Delorme et al., 2004; Rousselet et al., 2002; Thorpe et al., 1996). Even when these animal/no-animal scenes were only flashed for 6.25 ms, this difference was observed at frontal and occipital sites (Bacon-Macé et al., 2005). Furthermore, a comparison between animal and vehicle-containing scenes showed differential activity explained by visual content starting at 75 ms, but found task-related categorical differences at 150 ms (VanRullen & Thorpe, 2001). However, these scenes often display the animal or vehicle as a large portion of the image, positioned near the centre, making categorizing these images more akin to object(-in-scene) detection. More interesting are studies investigating scene categories, such as the study by Harel et al. (2016), which found that lateral-posterior P2 amplitude was higher for natural than manmade scenes viewed for 500 ms. What's more, the P2 amplitude was also influenced by image properties such as openness, contrast energy, and spatial coherence, suggesting that the P2 is sensitive to low-level image statistics in scenes (Harel et al., 2016). Other studies found that natural and man-made scenes viewed for 100 ms first showed occipital differential activity from 70 to 250 ms, which was followed by parietal-occipital differential activity between 258 and 464 ms if the scenes were task-relevant (Groen et al., 2016). These findings suggest that early activity might capture differences in visual properties, especially from 75 to 150 ms, while later activity might be task- or attention-dependent, with a potential early scene-gist-selective role for the occipital P2. However, it is not yet known whether the same neural patterns occur during the gist of abnormality.

The current chapter is an exploratory study aimed at investigating the neural signature of extracting the gist of medical abnormality from mammograms, by measuring EEG activity in medical experts as they rapidly view and rate mammograms as normal or abnormal. As stated before, there are no previous studies on the neural correlates of the gist of medical abnormality, making this a novel dataset. However, this also makes it difficult to provide precise predictions. The aim of the study if to observe differences in neural signature between normal and abnormal mammograms, analogue to indoor vs outdoor or manmade vs natural scenes. However, since medical gist categorisation is much more difficult than regular scene categorisations where performance is often in the high 90%, differential activity effects might only be visible in the subset of correctly categorized mammograms (hits and true negatives). Based on results from manmade vs natural scene research, differential neural activity would be expected to be first detectable from 75 ms at the earliest in occipital regions, with further parieto-occipital activity from 250 to 470 ms. Especially the occipital P2 might be important as scene-selective activity has been found in this ERP. However, early differences (75-250 ms) might be limited, as the low-level visual differences between a normal and abnormal mammogram are expected to be more subtle than differences between e.g., manmade and natural scene categories. This is especially true for contralateral and prior mammograms that do not contain any localizable signals of abnormality. More widely, results in animal/vehicle-in-scene detection suggest effects might be observed in frontal, occipito-temporal, and perhaps parietal regions from 150 ms onwards, although these might be evoked by object-in-scene rather than gist categorisation and thus might not be present in mammograms. Lastly, differential activity in later ERP components might be stronger in radiologists with better performance or more (recent) perceptual experience, as their neural system might be more finely tuned to the gist of medical abnormality.

5.3. Materials & Methods

Participants

5 radiologists (mean age: 51, 5 female, 4 right-handed) took part in this study. They were very experienced at mammography reading, having read between 5000 and 10000 scans for an average of 6600±1693 (SD) scans read in the previous year, with on average 94±11 percent of

their total caseload consisting of mammograms, and had been practicing for between 5 and 18 years, for an average of 12.2±4 years.

This study was approved by the Psychology Departmental Ethics Committee of the University of York (ID: 141), and all participants gave informed consent. Participants were compensated for their time. The number of participants was limited by participant availability, as this is a specialist population, that is hard to recruit, especially for an in-person study.

Stimuli and apparatus

The mammograms used in this study were unilateral and in mediolateral oblique (MLO) or craniocaudal (CC) view. These unilateral mammograms were divided into 4 categories: Normal, obvious abnormal, subtle abnormal, and global abnormal. Normal mammograms came from healthy women without any abnormalities. Obvious and subtle mammograms contain local abnormalities but were classified depending on the difficulty of recognising the abnormality in a normal screening based on the perceptual ratings of an experienced radiologist. Lastly, global mammograms were mammograms without any localizable abnormalities, which are thought to carry only global signals of gist of abnormality. These global signals are present in mammograms contralateral to a breast with a cancerous abnormality (contralateral), and in mammograms taken one to six years prior to visible actionable sign of abnormality appearing in a subsequent scan (priors). Normal, obvious, subtle, and contralateral cases were sourced from the OPTIMAM database, while priors were sourced from the Complex Cognitive Processing Lab database in collaboration with Dr. Bradley of the York Hospital. Researchers can request access to the OPTIMAM database through an application process (medphys.royalsurrey.nhs.uk/omidb/getting-access/), while the prior cases can be shared with other researchers upon reasonable request to the CCPL (ccpl.york.ac.uk, K.K. Evans).

This experiment used the same image set as the testing phases in Chapter 4, aside from a small modification to increase the number of trials: the 200 mammograms (80 normal, 30 obvious, 30 subtle, 30 contralateral, and 30 prior mammograms) were doubled by adding the mirrored equivalent of each. The mirrored mammograms increased the number of trials, while retaining the ability to compare performance with Chapter 4, as the difficulty of the mirrored and original images is expected to be equal. Additionally, in order to familiarize the radiologists with the trial structure, 30 practice trials were added, taken from a set of obvious abnormal and normal mammograms not in the testing set. Thus, there were 430 trials in total, 400 of which were used in data analysis.

Stimuli were presented to the radiologists on a VPixx 3D Lite monitor with a 120 Hz refresh rate, with a viewing distance of approximately 50 cm. A wrist rest was used to allow the participants to rest their fingers on the keyboard keys, minimising unnecessary hand movements. The experimental stimuli and behavioural responses were generated and received on a Mac Pro computer using MATLAB (Mathworks, MA, USA) with the PsychToolbox add-on (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). EEG data was collected using via a second computer using ASALab version 4.9.2.23537 (ANT Neuro, Netherlands), which was connected to the EEG cap via a high-speed 64-channel amplifier with a 1000 Hz sampling rate. EEG caps with a 64-channel layout according to the 10/20 system (WaveGuard original, ANT Neuro, Netherlands) were used, with three cap sizes (S/M/L) according to the participant's skull circumference. Vertical electro-oculogram (EOG) data was recorded using self-adhesive electrodes positioned above the left eyebrow and on the top of the left cheekbone, after cleaning the skin with an alcohol wipe. EOG data was sent to the same amplifier as the EEG data. Timing of stimuli behavioural responses, EEG, and EOG were synchronised using triggers that were generated by PsychToolbox and sent to the EEG/EOG amplifier.

Procedure

Each trial consisted of a fixation cross (1000 to 2000 ms, randomized interval), with a short 8kHz audio cue in the last 100 ms to alert the participant the stimulus would appear, after which the unilateral mammogram was shown for 500 ms, which was subsequently masked with a white outline of the same breast for 500 ms, after which the 2-AFC was shown for a maximum of 2000 ms, where the participant pressed either the left or right arrow key (Fig 1). The next trial started automatically 200 ms after the participant pressed one of the arrow keys. The 2000 ms window was set based on previously observed reaction times (e.g. Chapter 4) and piloting of the experiment. If no answer was given in the 2000 ms window, the trial was classified as not answered, and was not included in further analysis. The radiologists failed to respond to an average of 4 out of 400 trials, showing a high overall response rate.

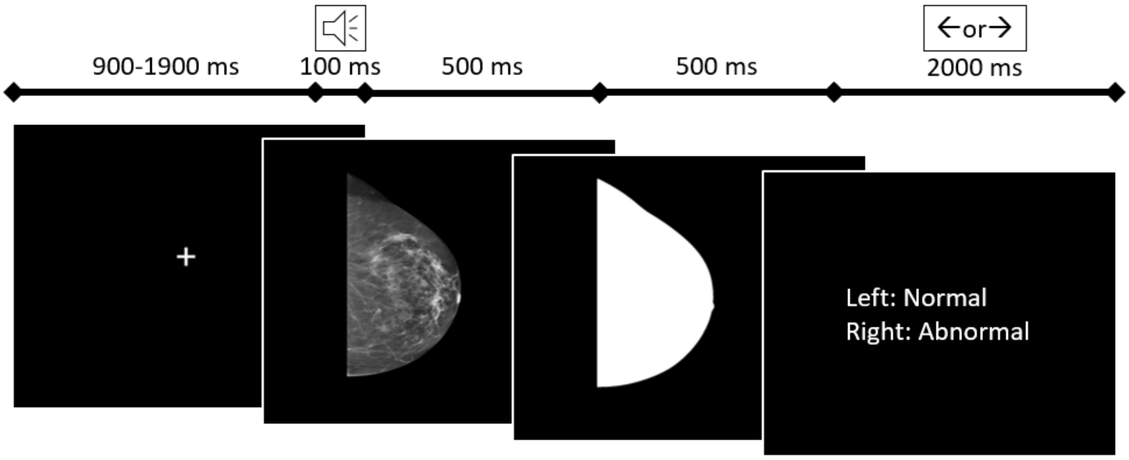


Figure 5.1: *The timeline of a trial, showing the fixation cross (1000 – 2000 ms total), auditory cue (100 ms), mammogram (500 ms), mask (500 ms), and 2-AFC answering window (2000 ms).*

As described in stimulus & apparatus, the experiment consisted of 30 practice trials, followed by 400 experimental trials. The left and right arrow keys were used to answer whether the mammogram was normal or abnormal. The initial binding of the arrows to normal/abnormal was counter-balanced between participants and flipped after 200 experimental trials, in order to counterbalance the directionality of responses both for any potential directional answering bias and EEG motor responses. Participants took self-timed breaks after the 30 practice trials, and after each 100 trials thereafter. Participants were shown their overall score during the break after 200 and 400 real trials, which aimed to increase motivation. Due to technical difficulties, no audio cues were available for radiologist 4.

The procedure for the EEG experiment differed from the testing phases in Chapter 4 in a few ways to improve the EEG recordings. The main difference was the rating scale, which was a 2 alternative forced choice (2-AFC) between normal and abnormal instead of a 0-100 slider, in order to reduce movement to a single keypress, and to speed up the trials. Additionally, this study used the aforementioned answering window, whereas Chapter 4 had unlimited response duration.

Data analysis

Behavioural data was processed using signal detection theory to calculate d’ and criterion. D’ indicates performance, where higher values indicate better performance, with 0 being chance-level. Criterion indicates bias, where positive values indicate a conservative bias, where participants were more likely to mark mammograms as normal, and negative values indicate a liberal bias, where a participant was more likely to mark any given mammograms as abnormal. No AUC was calculated since ratings were 2-AFC (normal/abnormal). Trials where the participant did not respond in time were excluded from the calculations. For radiologist 4), arrow key responses were flipped for the first 10 trials of block 2, as the radiologist reported they had been answering the wrong way around up to that point during the measurement, as noted by the experimenter during the measurement. D’, proportion of Hits, proportion of False Alarms, and criterion were the main performance outcomes.

EEG data was analysed using MNE-Python (Gramfort et al., 2013). First, EEG data was pre-processed using MNE-FASTER (https://github.com/wmvanvliet/mne-faster) based on the Fully Automated Statistical Thresholding for EEG artifact Rejection (FASTER) technique (Nolan, Whelan, & Reilly, 2010), which automates data cleaning, reducing personal bias. Channel Cz was used as online reference, but as the first step of data processing, all data was re-referenced to average activity. Then it was bandpass filtered between 0.5 and 40 Hz to remove electricity and movement artifacts and resampled to 200 Hz to reduce file size. FASTER was run on this filtered data, which first detected bad or dead channels and interpolated these, after which signal components containing eye movements or other artifacts were removed using Independent Component Analysis. Lastly, individual epochs were cleaned, interpolating channels which contained temporary noise, and epochs that remained too noisy were rejected. After FASTER, each epoch was baseline corrected to the average activity -700 to -200 ms before stimulus onset. On average, FASTER interpolated 4.2±1.939 (2 to 7) channels, removed 2±1.095 (1 to 4) Independent Component Analysis components, and lastly, rejected 10.4 ±2.154 (8 to 13) out of 400 trials for remaining noise.

Events of interests were identified based on the behavioural data to structure the epochs. This resulted in each epoch labelled as a hit, miss, true negative, and false alarm depending on the rating and ground truth. Each epoch was constructed using a time window of -700 to +1000 around the appearance of the mammogram.

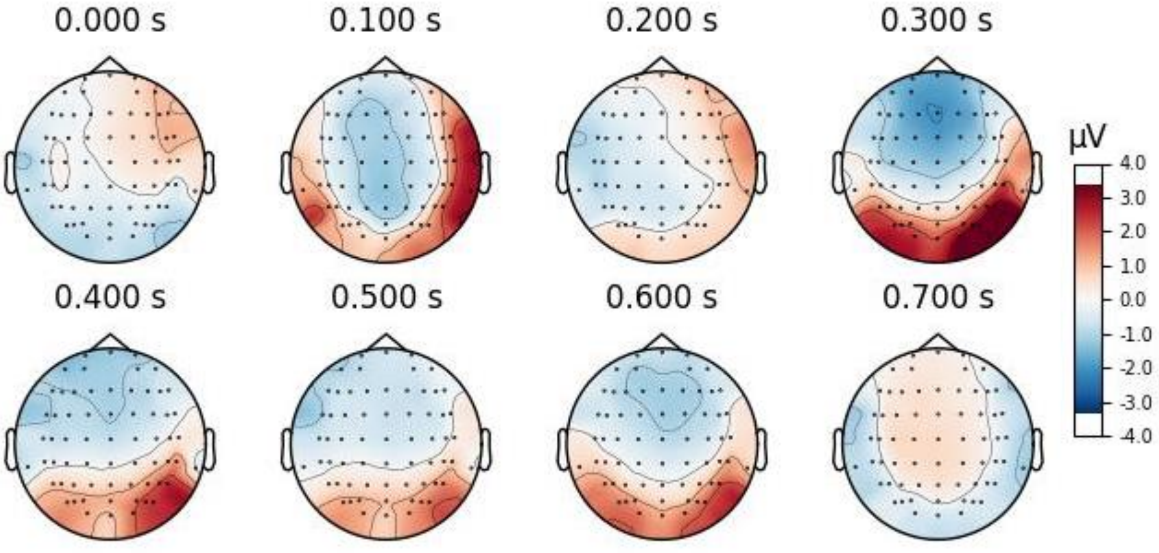


Figure 5.2: *Timeline of topography maps of averaged activity across the five radiologists.*

Five clusters of interest were identified based on both literature and analysis of a topography timeline across all trials and radiologists (Fig 2). This timeline showed early positivity in occipital, parietal and temporal regions, with negativity in central and frontal regions. The occipital activity persisted to about 600 ms, after which some frontal and central positivity appeared at 700 ms. Based on this, the following clusters were chosen in this study (Fig 3):

- Occipital: Oz, O1, O2, POz, PO3, PO4
- Parietal: Pz, P1, P2, P3, P4, P5, P6, P7, P8

- Temporal: FT7, FT8, T7, T8, TP7, TP8
- Central: FC1, FCz, FC2, C1, C2, CP1, CPz, CP2
- Frontal: Fpz, Fp1, Fp2, AF8, AF7, AF4, AF3, F8, F7, F6, F5, F4, F3, F2, F1, Fz

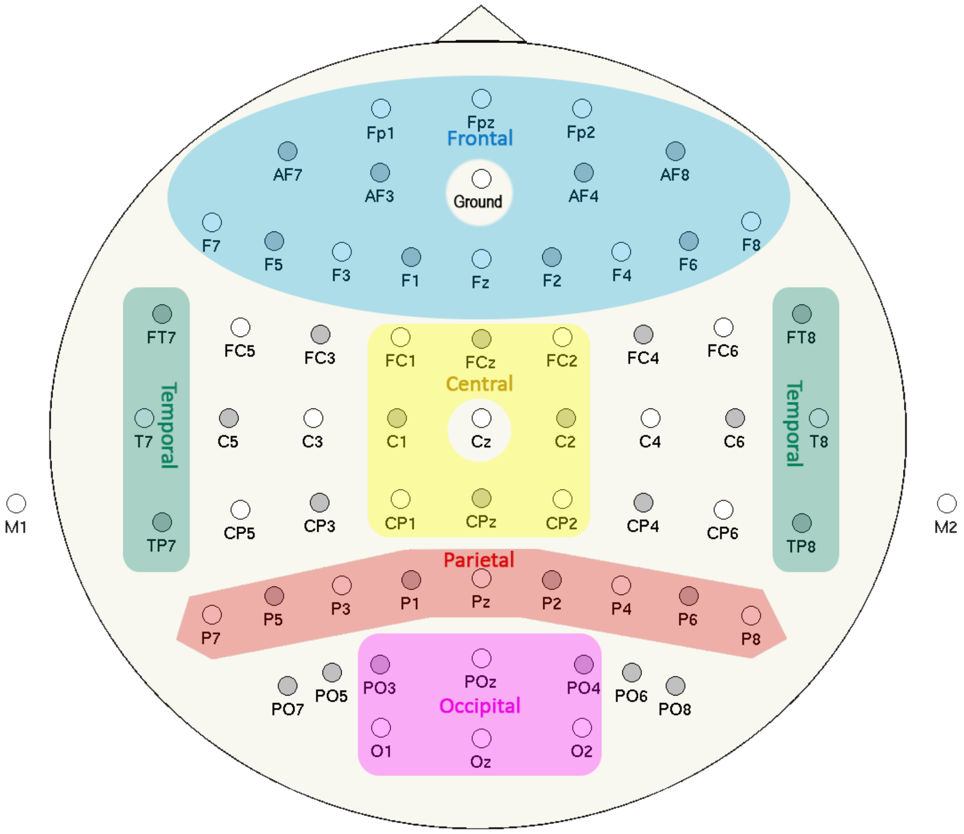


Figure 5.3: *Channel layout and chosen clusters of interest: Frontal (blue), central (yellow), temporal (green), parietal (red), and occipital (pink).*

The main effect of interest was that of different gist categories on neural activity, which was investigated by comparing the evoked activity in two different categories of mammograms. Two main comparisons were made, firstly between abnormal (cancerous) and normal mammograms, secondly between hits and true negatives. This second comparison was chosen as trials in which the participant correctly identified the mammogram as (ab)normal would be expected to have the strongest gist of (ab)normality processing. Because of the small sample size in this exploratory study, single subject bootstrap analysis was performed, following the methods described by Oruç et al. (2011). The bootstrap analysis was performed for each subject for the chosen ERPs for each cluster of interest (table 1), comparing the activity between two conditions (abnormal vs normal; hits vs true negatives).

For each cluster, the ERPs of interest and their respective search windows (table 1) were chosen by a combination of ERP descriptions in previous literature and inspecting the averaged time traces in our experiment (see for example Fig 6), of which a brief summary is given here. The chosen ERPs are a combination of traditional visual (P1, N1, P2, N2) and decision making (P3a, P3b) ERPs, and more exploratory choices (N400, P600). A wide range of ERPs and clusters was chosen, as this is an exploratory study in which both effects and any lack thereof are of interest, given that there is no previous research investigating the EEG patterns evoked by gist extraction in medical images.

Firstly, P1 amplitude is strongly influenced by spatial attention, but also by sensory information, responding more strongly to target stimuli (Hillyard & Anllo-Vento, 1998), making it potentially interesting for early differential effects. The P1 is characterized by posterior positivity from 70-90 ms after stimulus onset, peaking around 80 to 150 milliseconds, with maximal activity in the occipital region (Finnigan, O'Connell, Cummins, Broughton, & Robertson, 2011; Gonzalez, Clark, Fan, Luck, & Hillyard, 1994; Hillyard & Anllo-Vento, 1998). P1 activity was investigated in the occipital, parietal, and temporal clusters with a 60 to 150 ms search window.

Next, similar to the P1, N1 amplitude is also amplified by spatial attention and sensory information (Hillyard & Anllo-Vento, 1998), but attentional modulation of the N1 occurs only when further visual processing is required for a category response RT task (e.g. long vs short bar) rather than a simple stimulus-onset RT task (Mangun, 1995; Vogel & Luck, 2000). As our task requires further processing, the N1 might show effects of any early normal-abnormal discrimination. The N1 is observed in most brain regions, including anterior and posterior areas (Finnigan et al., 2011; Gonzalez et al., 1994; Hillyard & Anllo-Vento, 1998), with activity in an approximate time window of 130 to 210 ms. The peak of the N1 is often observed first in anterior regions, around 150 ms, followed by the posterior regions, around 170 ms (Gonzalez et al., 1994). While the N1 is not traditionally investigated in the temporal cluster, it has been reported in centro-temporal cluster (C3, C4, T3, T4 in the 10-20 layout) (Johannes, Münte, Heinze, & Mangun, 1995). For that reason, the N1 was tested across all five clusters, with later search windows for the posterior (120 to 200 ms) than anterior (75 to 150 ms) clusters.

The P2 is often thought to be involved with working memory and encoding (Finnigan et al., 2011). The current study had a special interest in the P2 as this ERP was suggested as a scene-selective marker with sensitivity to scene category (manmade/natural) and image statistics (e.g. spatial coherence) (Harel et al., 2016). The P2 is observed in various brain regions, including frontally and occipitally (Kanske, Plitschka, & Kotz, 2011). P2 latency differs between regions, for example around 190 to 290 ms frontally (Kanske et al., 2011), and 160 to 340 ms occipitally (Finnigan et al., 2011). The P2 was tested in the occipital and parietal (200 to 350 ms), and frontal and central cluster (150 to 250 ms).

N2 activity can observed across anterior and posterior areas, with different roles for each component. The focus in this study was on the fronto-central component. This fronto-central N2 is thought to be involved in cognitive control and in mismatch from a perceptual template (Folstein & Van Petten, 2008), which might result in differential activity in the current study. This fronto-central N2 is sometimes labelled as the N2c, or classification N2 (Näätänen & Picton, 1986; Patel & Azzam, 2005). The N2 plays a potential role in the classification of gist of abnormality. The fronto-central N2 typically peaks around approximately 180 to 325 ms (Folstein & Van Petten, 2008; Patel & Azzam, 2005). The N2 was tested in the frontal and central cluster with the 180 to 325 ms search window.

The P3a is thought to reflect passive comparison, influenced by attention and novelty, that plays a role in initial detection and attentional recruitment towards a target. In concert with the P3a, the P3b represents a (mis)match with a consciously held working memory trace, with amplitude indexing memory storage. However, the P3b is also involved in decision making (Twomey, Murphy, Kelly, & O'Connell, 2015), and its amplitude might represent cumulating evidence towards a threshold to trigger a categorization decision. P3b's role as a decision-making threshold is especially interesting for the current study. The P3a and P3b are part of the P300, which typically emerges approximately 300 to 400 ms after onset, but is known for a large range of possible latencies, from 250 to 900 ms (Patel & Azzam, 2005). The P3a typically has a shorter latency and a fronto-central distribution, and habituates faster than the P3b (Polich, 2003). The P3a was tested for the frontal, central, and temporal clusters with a search window of 200 (temporal) or 250 (frontal/central) to 350 ms. The P3b occurs more posteriorly and was tested for the occipital and parietal clusters with a search window of 350 to 500 ms.

The N400 is known to be involved in semantic context mismatch in reading, but has also been observed in non-semantic situations (Näätänen & Picton, 1986). For example, N400 amplitude was influenced by object-in-scene congruity (Ganis & Kutas, 2003), with a centro-parietal maximum. The N400 can be observed in mainly anterior and central regions, with the frontal N400 occurring between 460 and 590 ms (Kanske et al., 2011). The temporal cluster N400 was also investigated in this study as the average trace indicated a potential effect at the N400 time window, and some studies have reported a smaller, but measurable N400 trace in temporal electrodes (Kutas & Hillyard, 1982; Shin, Kang, Choi, Kim, & Kwon, 2008). Thus, N400 was investigated in the temporal, frontal, and central clusters with a search window of 350-500 (temporal) or 350-600 ms (frontal/central).

Lastly, the P600 has commonly been investigated in the context of syntactical grammatical errors during reading (Osterhout & Holcomb, 1992). However, increased P600 amplitude was also observed across the scalp when flashing an incongruent eye-region area in a face image (completed face shown for 200 ms), with maximal activity parietally at 620 ms (Jemel, George, Olivares, Fiori, & Renault, 1999). It is a more untraditional choice for a visual categorization study, but was included as an exploratory component, as a cancerous abnormality could be regarded as an incongruent area within a mammogram. The P600 has a latency of 500 to 800 ms with a midpoint around 600 ms and is widely distributed, with maximal activity fronto-centrally (Jemel et al., 1999; Osterhout & Holcomb, 1992). The P600 was tested in the parietal, temporal, central, and frontal cluster (500 to 700 ms).

Table 5.1: *List of the ERPs that were tested for each of the clusters, with the start and end (ms) of the search window for the peak activity, the direction of the peak, and the target window length (ms) around the peak. The table combines some clusters where the same ERPs were investigated, to reduce the length of the table.*

Cluster	*ERP*	*Search Start*	*Search End*	*Direction*	*Target Window*
Occipital & Parietal	P1	70	150	+	50
Occipital & Parietal	N1	120	200	-	50
Occipital & Parietal	P2	200	350	+	100
Occipital & Parietal	P3b	350	500	+	100
Parietal	P600	500	700	+	100
Temporal	P1	70	150	+	50
Temporal	N1	120	200	-	50
Temporal	P3a	200	350	+	100
Temporal	N400	350	500	-	50
Temporal	P600	500	700	+	100
Frontal & Central	N1	75	150	-	50
Frontal & Central	P2	125	250	+	50
Frontal & Central	N2	180	325	-	50
Frontal & Central	P3a	250	350	+	50
Frontal & Central	N400	350	600	-	50
Frontal & Central	P600	500	700	+	100

The single subject bootstrap method (Oruç et al., 2011) briefly consists of the following steps: Within each search time window, individual target time windows were drawn around the point where the average ERP for that cluster-condition-combination had the highest amplitude (e.g. for a Central P2, a positive peak at 173 ms would give a 50 ms target time window of 148 – 198 ms across which activity was averaged). This was done to account for individual differences in timing of ERPs, since the goal of this analysis was to identify differences in peak amplitude. The activity difference between conditions was calculated as the difference between the mean potential of the ERPs for condition A and B across their respective target time windows. Then, to check if this difference was significantly different from 0, non-parametric bootstrapping was performed and the activity difference was calculated on each random resample, creating a histogram of activity differences (Fig 4). The p-value of the differences is calculated as the (smaller) proportion of resamples overlapping with 0 times two, as the comparisons were all two-tailed, as there was no a priori expectation of which condition would be larger, for example for radiologist 5, 1.5% of the resampled differences in Temporal P600 were lower than 0, so the p-value would be .03, showing significant evidence for higher activity in Abnormal than Normal trials. Figure 5 shows the corresponding average activity, target time windows, and p-values for each radiologist for the Temporal P600. The goal of this exploratory study was to highlight potentially interesting activity patterns for future investigation, which was why a wide net was cast across potentially interesting brain regions and ERPs. Thus, p-values were not corrected for multiple comparisons.

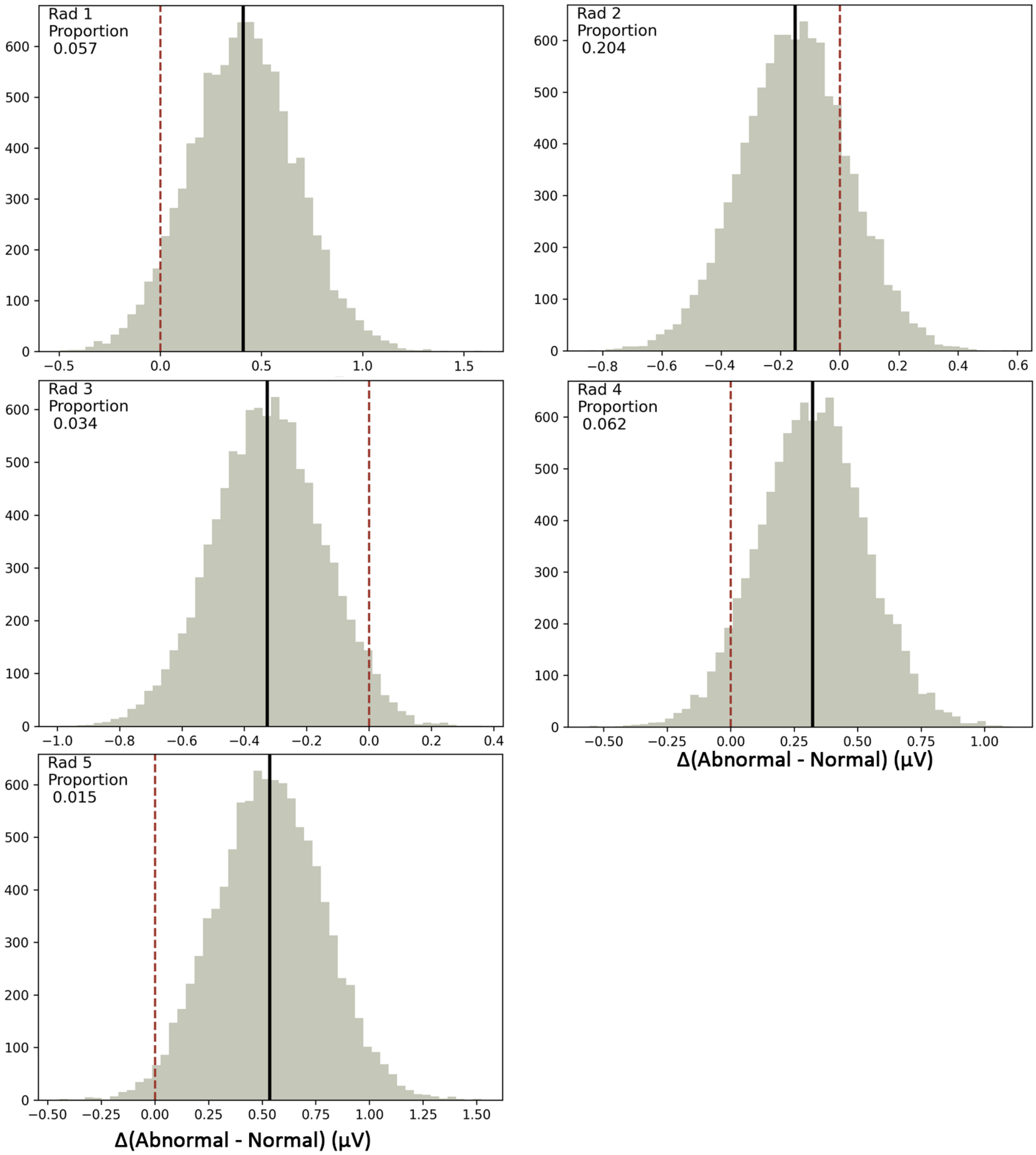


Figure 5.4: *Example histograms showing the distributions of 10.000 bootstrapped differences (μV) in Temporal P600 ERPs when viewing abnormal versus normal mammograms for each radiologist. The histograms show that the temporal P600 evoked by Abnormal mammograms was larger in amplitude for radiologists 1, 4, and 5, and smaller for radiologists 2 and 3. Each plot also shows the real observed difference (black line), zero difference line (red dashed line), and the proportion of samples below/above 0 (top left corner).*

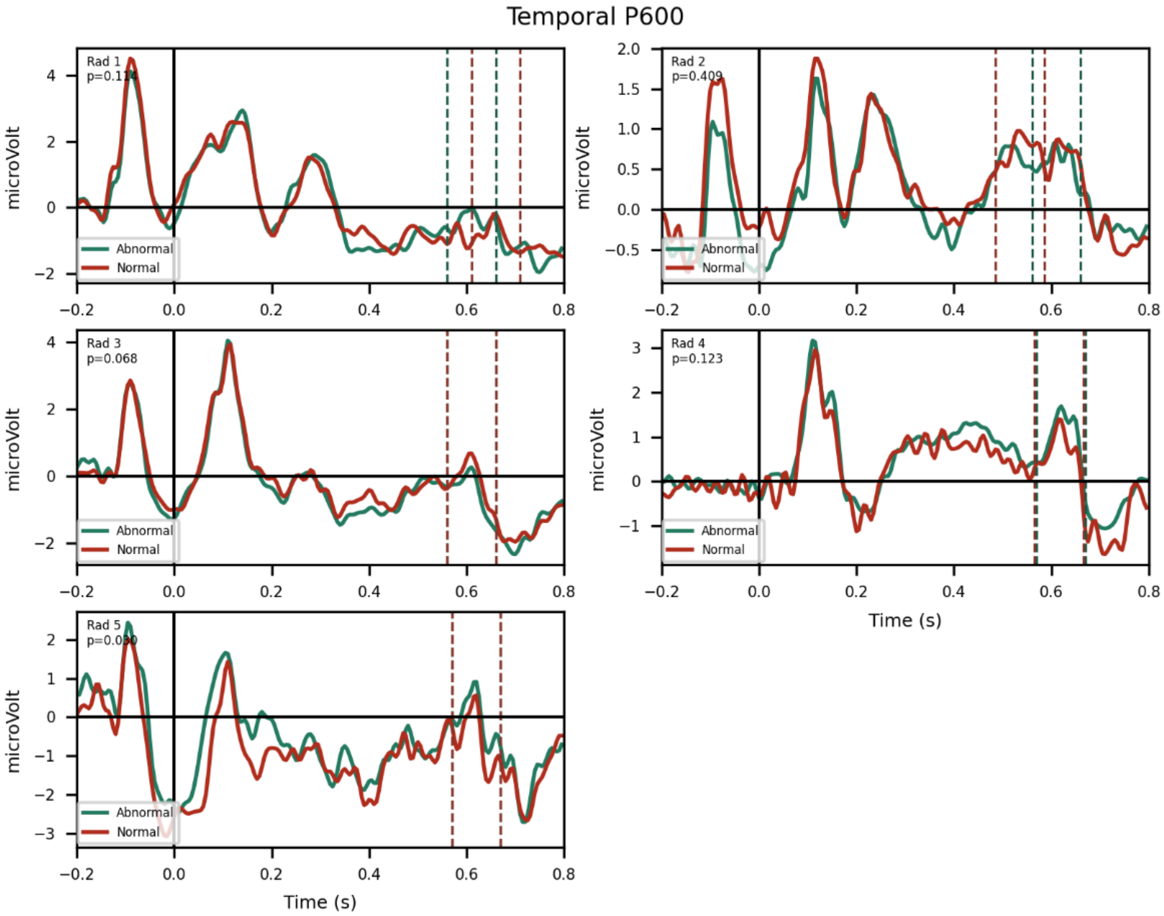


Figure 5.5: *Single subject bootstrapping results showing the Temporal P600 target time windows for abnormal (green dashes) and normal (red dashes), averaged activity, and p-values per radiologist.*

5.4. Results

Behavioural analysis

As expected, radiologists extracted the gist of abnormality with above-chance accuracy. Radiologists' performance was highest on mammograms with obvious abnormalities, followed by subtle abnormalities, with a substantial drop in performance for the global cases, while remaining above chance for four out of five radiologists (table 2). Looking at criterion, there was some variation, with three radiologists leaning towards a conservative rating strategy, while the other two had a slightly liberal rating strategy, but neither showed a large bias, with all values between -0.25 and 0.2.

Table 5.2: *Performance measures for each radiologist, showing their overall criterion, overall d' performance and d' separated for the obvious, subtle, and global mammograms. The years of experience and number of mammogram cases viewed in the previous year are also shown.*

RAD	CRITERION	D’	OBVIOUS	SUBTLE	GLOBAL	YEARS EXP	CASES VIEWED
1	-0.068	0.546	1.035	0.933	0.184	18	5000
2	0.058	0.622	1.210	1.152	0.137	5	6000
3	-0.118	0.424	0.878	0.619	0.136	14	5000
4	0.217	0.224	0.698	0.528	-0.117	14	7000
5	0.199	0.544	0.996	1.186	0.040	10	10000
AVERAGE	*0.058*	*0.472*	*0.963*	*0.884*	*0.076*	*12.2*	*6600*

Differential neural activity per cluster

To explore the neural correlates of the gist of medical abnormality, evoked activity is compared between trials showing abnormal and normal mammograms, and between hits and true negatives. Investigating this differential activity can show when and where the gist of medical abnormality appears. The average EEG traces associated with abnormal and normal mammogram trials (Fig 6) as well as hits and true negatives (Fig 7) are shown for each individual radiologists as well as the group average (Fig 6) to illustrate the overall activity patterns. These show considerable differences in timing and amplitude of EEG activity between the radiologists across the five neural clusters. However, visual inspection also shows potential areas of divergence in activity levels, where differential activity might occur between the abnormal and normal mammograms. In the next sections, detailed investigation of differential activity is described for each radiologist using single subject bootstrapping for each cluster.

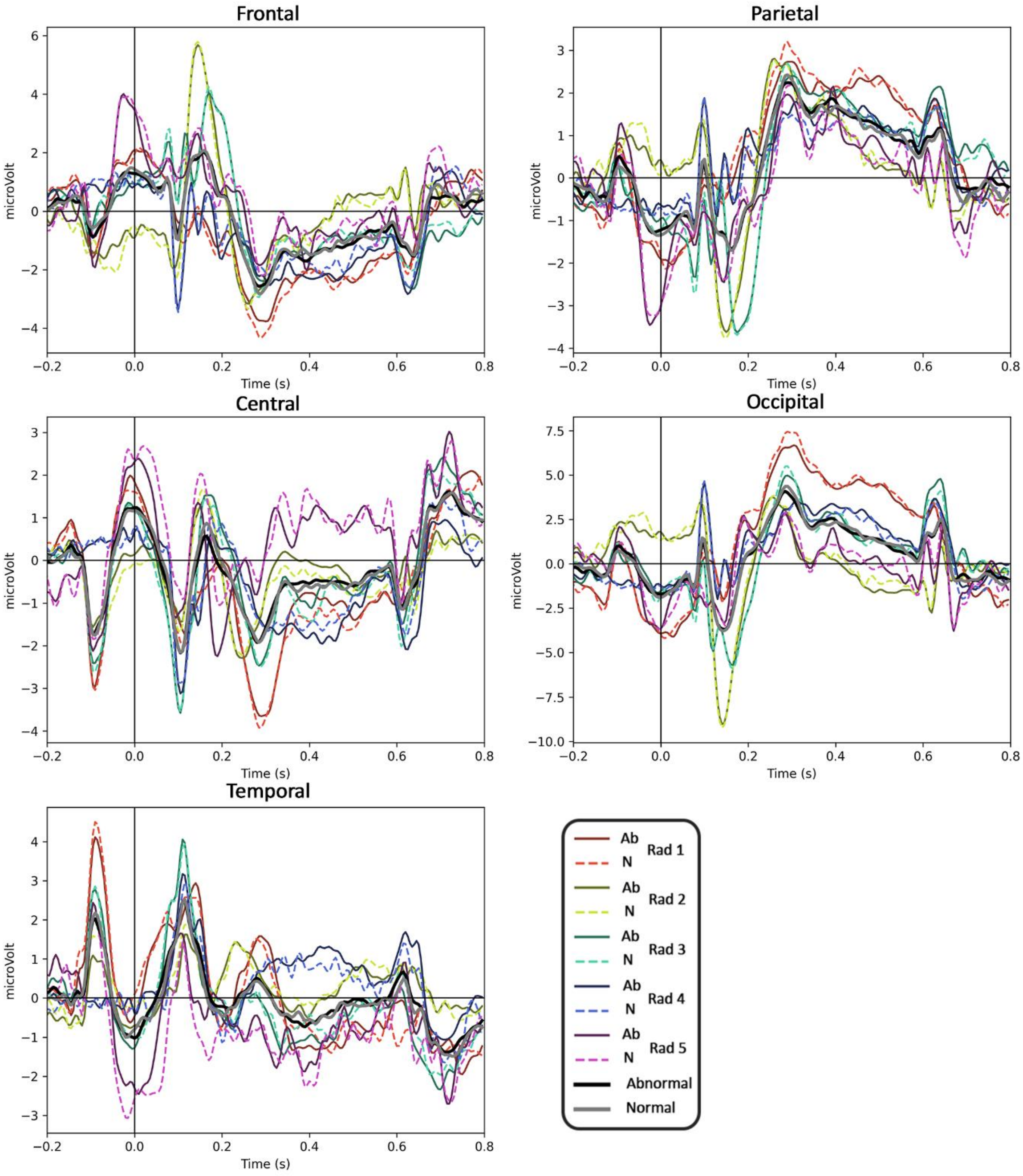


Figure 5.6: *Time trace of the average neural activity in Frontal, Central, Temporal, Parietal, and Occipital clusters for Abnormal (solid) and Normal (dashed) mammograms for individual radiologists, and group averages (Abnormal, black line; Normal, grey line).*

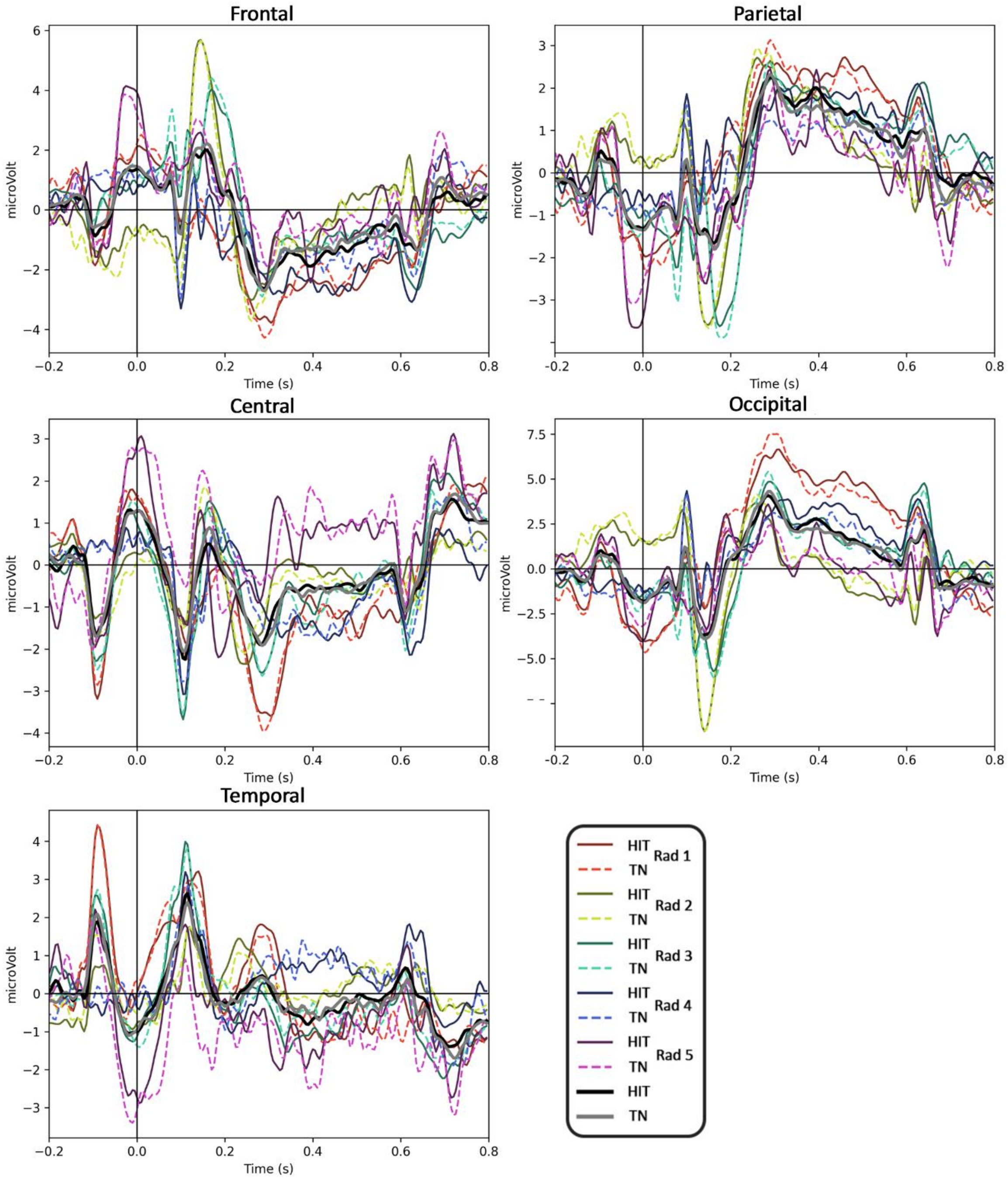


Figure 5.7: *Time trace of the average activity in Frontal, Central, Temporal, Parietal, and Occipital clusters for Hits (solid) and True Negatives (dashed) mammograms for individual radiologists, and group averages (Hits, black line; True Negatives, grey line).*

Occipital cluster

For the occipital cluster, there was no significant evidence for differential activity between abnormal and normal mammograms for any of the four investigated ERPs (Table 3). For hits and true negatives, on the other hand, there was evidence for differential activity in three ERPs (Table 4). While differential activity was only observed in two out of five radiologists in the occipital cluster, these had higher amplitudes for hits than true negatives.

Table 5.3: *Results of single subject bootstrapping for the Occipital cluster, showing the p-value and the average difference between abnormal and normal trials in μV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined.*

	P1		**N1**		**P2**		**P3b**	
	p	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**
1	0.109	-0.543	0.405	-0.293	0.073	-0.506	0.125	-0.449
2	0.628	0.172	0.686	0.155	0.764	-0.091	0.431	-0.196
3	0.632	0.169	0.747	0.115	0.544	-0.192	0.195	0.368
4	0.896	-0.042	0.920	0.035	0.355	0.277	0.226	0.329
5	0.495	0.327	0.984	0.025	0.941	-0.030	0.729	0.115

Table 5.4: *Results of single subject bootstrapping for the Occipital cluster, showing the p-value and the average difference between hits and true negatives in μV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined.*

	P1		**N1**		**P2**		**P3b**	
	p	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**
1	0.179	-0.562	0.600	-0.244	0.104	-0.581	0.625	-0.191
2	0.696	-0.171	0.773	-0.137	0.345	-0.356	0.453	-0.233
3	**0.017**	1.044	0.340	0.442	0.553	-0.264	0.438	0.291
4	0.910	-0.034	0.818	0.093	**0.034**	0.828	**0.023**	0.789
5	0.136	0.933	0.366	0.600	0.554	0.288	0.376	0.384

Parietal cluster

Next, for the Parietal cluster, there was significant evidence for differential activity between abnormal and normal mammograms for two of the five investigated ERPs in one radiologist each (Table 5). Differential activity was observed both early (P1) and late (P600), without significant effects in N1, P2, and P3b. Again, differential activity became more pronounced when comparing hits and true negatives: there were four comparisons with significant evidence for differential activity in three out of five ERPs, across two radiologists (Table 6). Overall, parietal activity was higher for abnormal/hits than normal/true negatives, except for the P1 where activity was higher for normal mammograms.

Table 5.5: *Results of single subject bootstrapping for the Parietal cluster, showing the p-value and the average difference between abnormal and normal trials in μV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined.*

	P1		**N1**		**P2**		**P3b**		**P600**	
	p	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**
1	**0.041**	-0.427	0.513	-0.133	0.080	-0.303	0.610	-0.090	0.993	-0.002
2	0.939	0.021	0.583	0.129	0.911	0.021	0.919	-0.019	0.775	0.043
3	0.258	-0.240	0.954	-0.009	0.692	0.083	0.558	0.108	**0.024**	0.369
4	0.807	0.052	0.722	0.074	0.123	0.294	0.325	0.173	0.137	0.228
5	0.788	-0.104	0.611	0.231	0.822	-0.057	0.691	0.100	0.093	0.393

Table 5.6: *Results of single subject bootstrapping for the Parietal cluster, showing the p-value and the average difference between hits and true negatives in μV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined.*

	P1		**N1**		**P2**		**P3b**		**P600**	
	p	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**
1	0.068	-0.468	0.726	-0.094	0.300	-0.222	0.223	0.267	0.180	0.315
2	0.629	-0.132	0.907	-0.041	0.550	-0.140	0.909	0.028	0.251	-0.235
3	0.691	-0.106	0.496	0.201	0.666	0.128	0.157	0.367	**0.036**	0.437
4	0.427	0.234	0.238	0.344	**0.006**	0.676	**0.022**	0.524	**0.000**	0.737
5	0.574	-0.296	0.795	-0.134	0.748	-0.102	0.835	0.071	0.695	0.113

Temporal cluster

For the temporal cluster, there was significant evidence for differential activity between abnormal and normal mammograms for three out of five investigated ERPs, all in the same radiologist (Table 7). Additionally, while not significant, differential activity trended towards significance in radiologist 3 for both the N1 and P600, but in the opposite direction to the significant effects observed in radiologist 5, as illustrated in Fig 8 (N1) and Fig 9 (P600). For hits and true negatives, there was significant evidence for differential activity for three out of five

ERPs across two radiologists (Table 8). N1 differential activity was significant for both radiologist 3 and 5, but again with opposite directionality (Fig 10). Overall, higher amplitudes were observed in the P1 and P3a for abnormals/hits, while the N1 and P600 showed opposite directionality of effect between the two radiologists with significant or trending effects. These reversals suggest that there might be individual differences in the representations of (ab)normality in the temporal region.

Table 5.7: *Results of single subject bootstrapping for the Temporal cluster, showing the p-value and the average difference between abnormal and normal trials in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined.*

	P1		**N1**		**P3a**		**N400**		**P600**	
	p	**Δ (µV)**	**p**	**Δ (µV)**	**p**	**Δ (µV)**	**p**	**Δ (µV)**	**p**	**Δ (µV)**
1	0.822	0.070	0.749	0.101	0.404	0.229	0.598	0.172	0.114	0.410
2	0.153	-0.408	0.983	-0.007	0.465	0.165	0.387	-0.204	0.409	-0.150
3	0.734	0.064	0.074	-0.351	0.922	0.017	0.073	-0.365	0.068	-0.326
4	0.933	-0.025	0.596	0.127	0.950	0.011	0.152	0.322	0.123	0.323
5	0.058	0.678	**0.003**	1.028	**0.011**	0.700	0.414	0.248	**0.030**	0.536

Table 5.8: *Results of single subject bootstrapping for the Temporal cluster, showing the p-value and the average difference between hits and true negatives in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined.*

	P1		**N1**		**P3a**		**N400**		**P600**	
	p	**Δ (µV)**	**p**	**Δ (µV)**	**p**	**Δ (µV)**	**p**	**Δ (µV)**	**p**	**Δ (µV)**
1	0.896	0.055	0.885	-0.050	0.080	0.551	0.870	-0.059	0.186	0.424
2	0.689	-0.139	0.359	0.307	0.180	0.351	0.260	-0.295	0.327	-0.223
3	0.995	-0.003	**0.020**	-0.582	0.711	-0.080	0.063	-0.525	0.457	-0.183
4	0.543	-0.191	0.328	-0.331	0.310	-0.312	0.525	-0.198	0.112	0.420
5	**0.049**	0.922	**0.003**	1.399	0.148	0.564	0.112	0.643	**0.005**	0.900

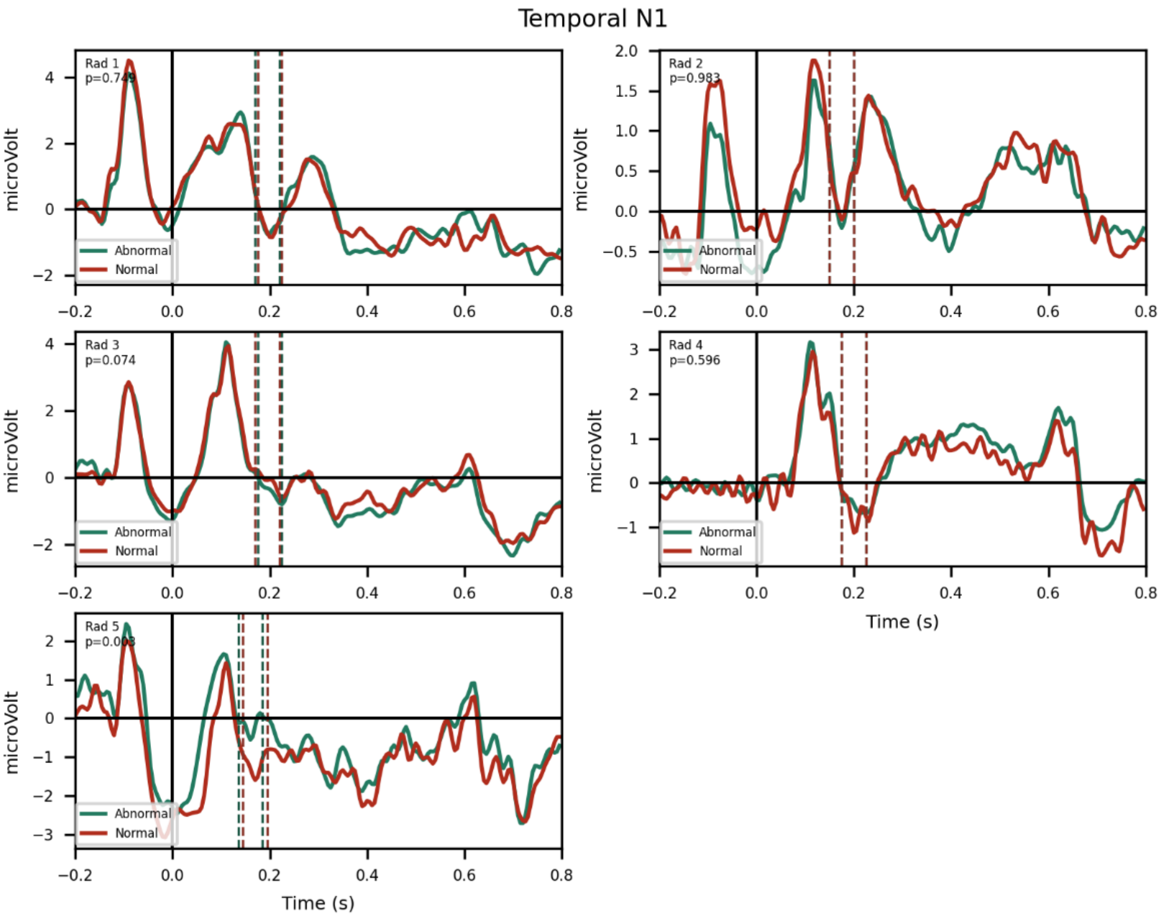


Figure 5.8: *Single subject bootstrapping results showing the Temporal N1 target time windows for abnormal (green dashes) and normal (red dashes), averaged activity, and p-values per radiologist.*

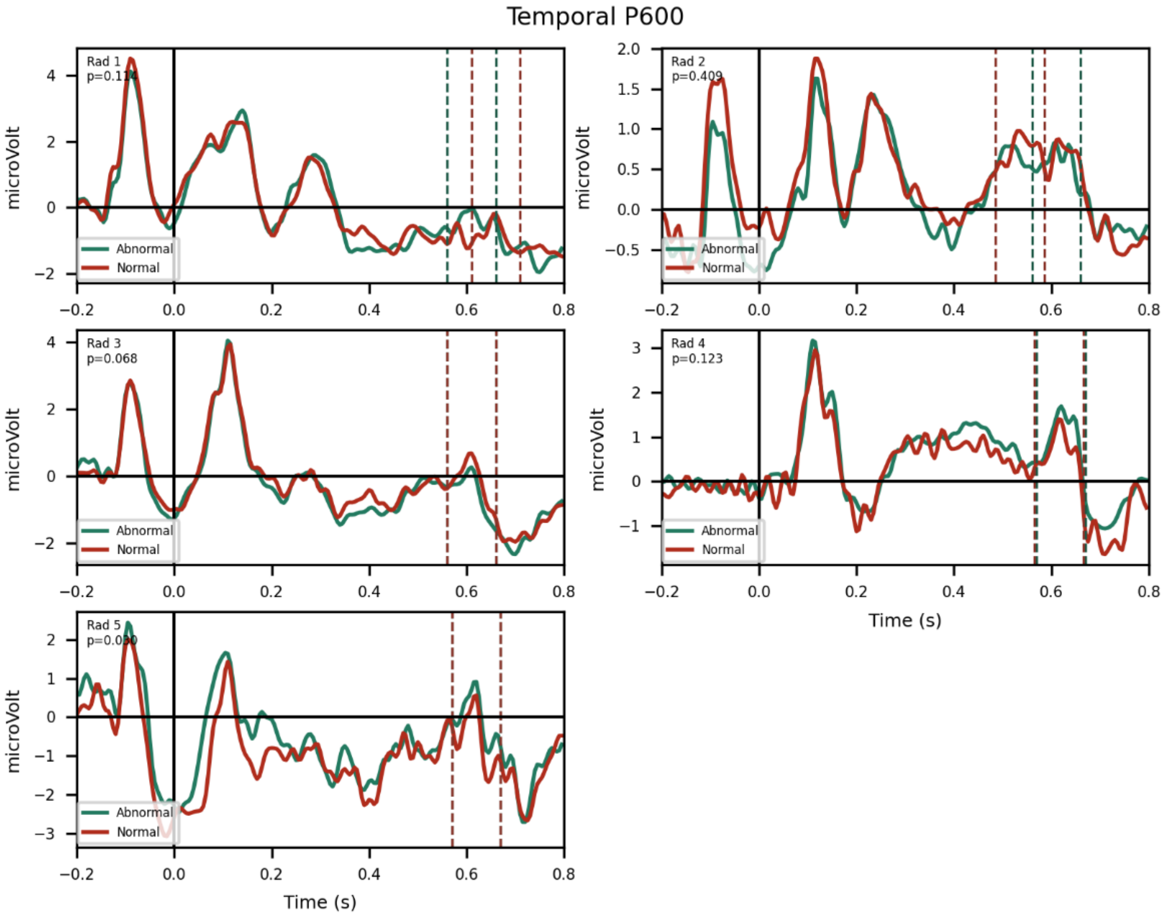


Figure 5.9: *Single subject bootstrapping results showing the Temporal P600 target time windows for abnormal (green dashes) and normal (red dashes), averaged activity, and p-values per radiologist.*

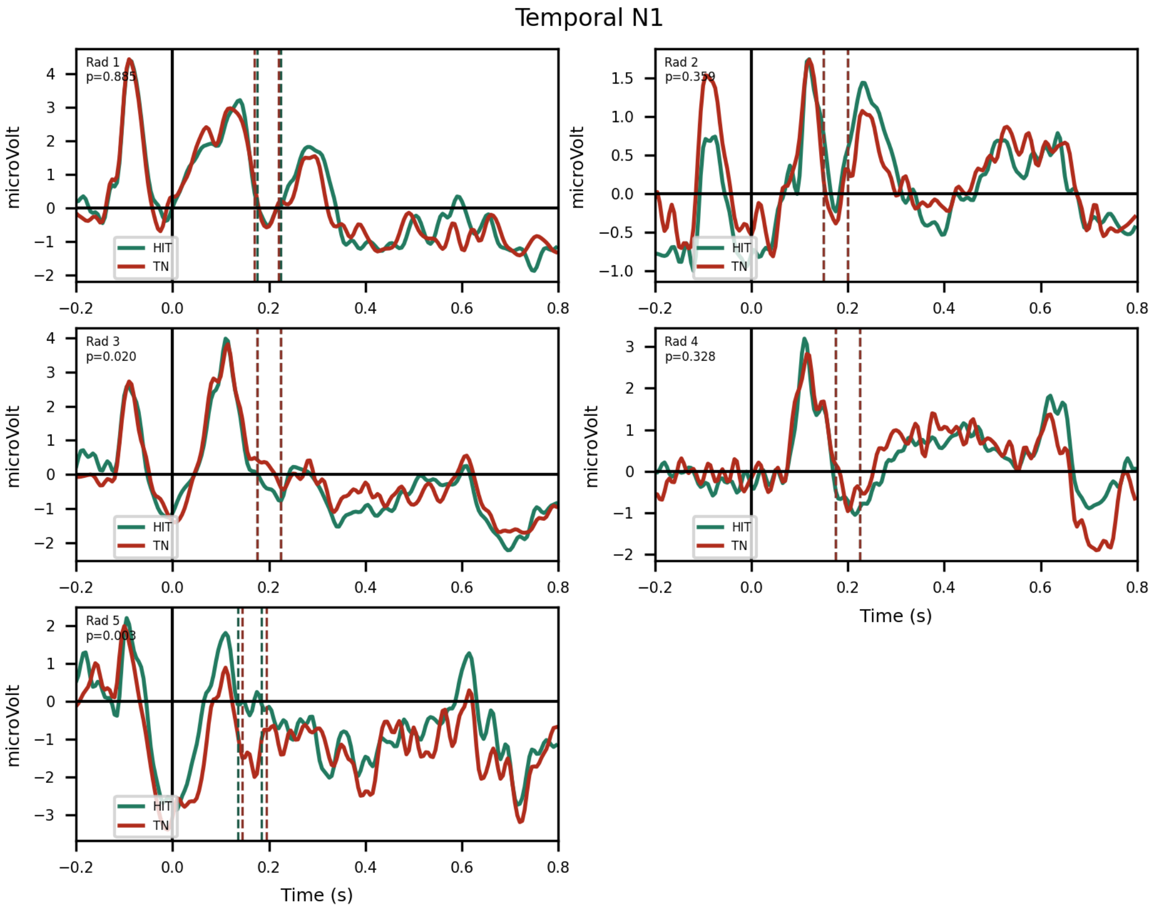


***Figure 5.10:** Single subject bootstrapping results showing the Temporal N1 target time windows for hits (green dashes) and true negatives (red dashes), averaged activity (hits: green, true negatives: red), and p-values per radiologist.*

Central cluster

For the central cluster, there was evidence for differential activity between abnormal and normal mammograms for three out of six ERPs, across three different radiologists (Table 9). The significant findings were characterized by higher amplitudes for abnormal mammograms. Additionally, for the P2, activity trended towards significance in two radiologists, radiologist 3 (higher amplitude for abnormal) and radiologist 5 (higher amplitude for normal) (Fig 11). Similarly, P600 amplitude was significantly higher for abnormal mammograms for radiologist 3 but trended towards significance in the opposite direction for radiologist 4 (Fig 12), again indicating potential individual differences. For hits versus true negatives, there were 4 significant findings, with evidence for differential activity in three out of six ERPs across three radiologists (Table 10). While amplitude was higher for hits in the N2, for the P2, amplitude was higher for true negatives. For the P600, opposite directionality of effects was observed for radiologist 3 (higher for hits) and radiologist 4 (higher for true negatives) (Fig 13). The opposite directionality of effects observed in the P2 and P600 for the central cluster suggest that individual differences may play a role in the neural signature of medical abnormality gist in these ERPs.

Table 5.9: *Results of single subject bootstrapping for the Central cluster, showing the p-value and the average difference between abnormal and normal trials in μV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined.*

	N1		**P2**		**N2**		**P3a**		**N400**		**P600**	
	p	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**
1	0.778	0.063	0.307	0.225	0.521	0.146	0.298	0.230	0.314	0.230	0.342	0.183
2	0.475	0.147	0.832	-0.048	0.753	-0.058	**0.040**	0.367	0.178	0.220	0.258	0.164
3	0.904	-0.028	0.069	0.383	0.953	0.009	0.157	0.299	0.081	0.365	**0.000**	1.270
4	0.730	0.077	0.246	-0.262	0.660	-0.109	0.750	0.076	0.197	-0.308	0.060	-0.383
5	0.100	-0.638	0.060	-0.675	**0.002**	-1.111	0.449	-0.218	0.341	-0.259	0.710	-0.096

Table 5.10: *Results of single subject bootstrapping for the Central cluster, showing the p-value and the average difference between hits and true negatives in μV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined.*

	N1		**P2**		**N2**		**P3a**		**N400**		**P600**	
	p	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**
1	0.323	0.274	0.265	0.323	0.530	0.181	0.954	-0.021	0.098	0.464	0.840	-0.051
2	0.397	0.205	0.458	-0.197	0.178	-0.317	0.181	0.307	0.693	0.084	0.306	0.181
3	0.876	-0.039	0.150	0.404	0.965	-0.013	0.312	0.295	0.155	0.383	**0.000**	1.303
4	0.931	0.022	0.173	-0.401	0.774	-0.094	0.415	0.268	0.788	-0.081	**0.000**	-1.343
5	0.085	-0.845	**0.006**	-1.245	**0.030**	-1.029	0.605	0.185	0.287	-0.372	0.903	0.042

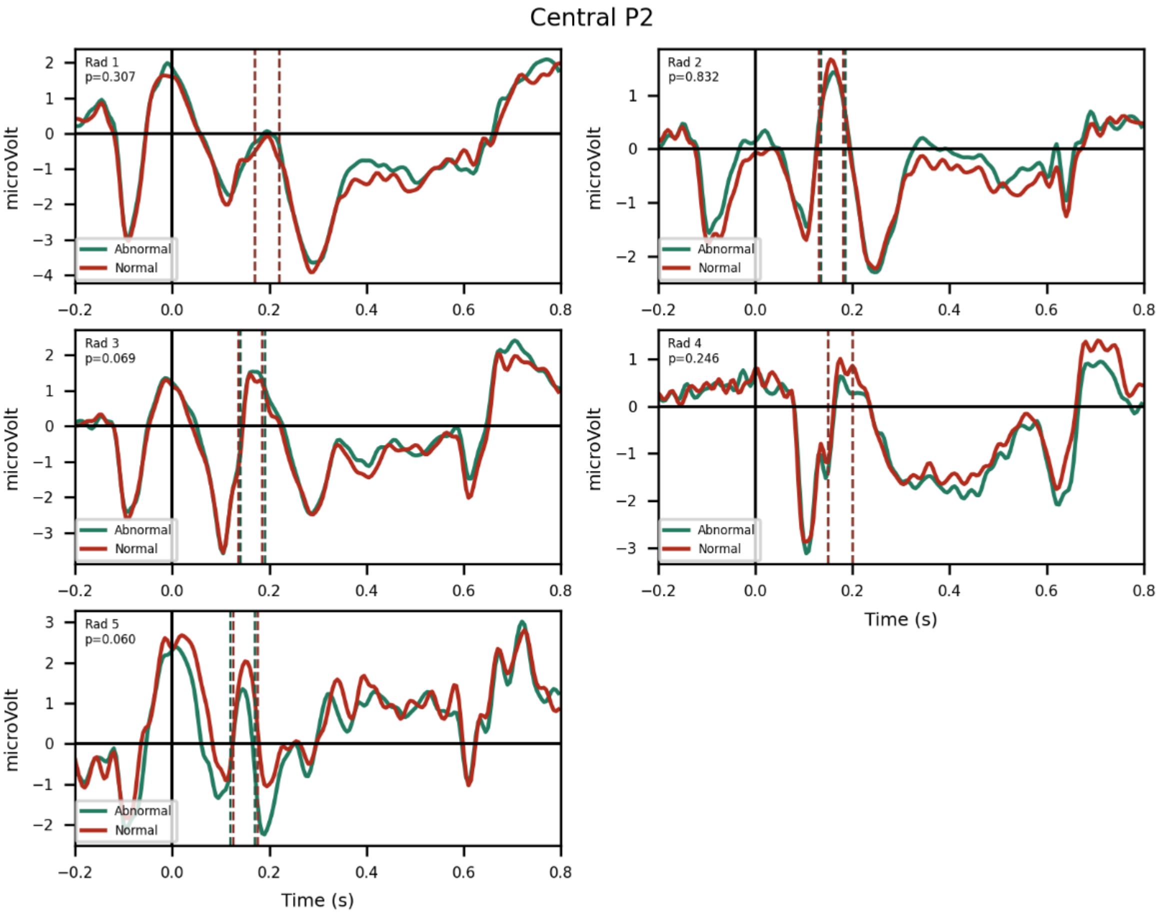


Figure 5.11: *Single subject bootstrapping results showing the Central P2 target time windows for abnormal (green dashes) and normal (red dashes), averaged activity, and p-values per radiologist.*

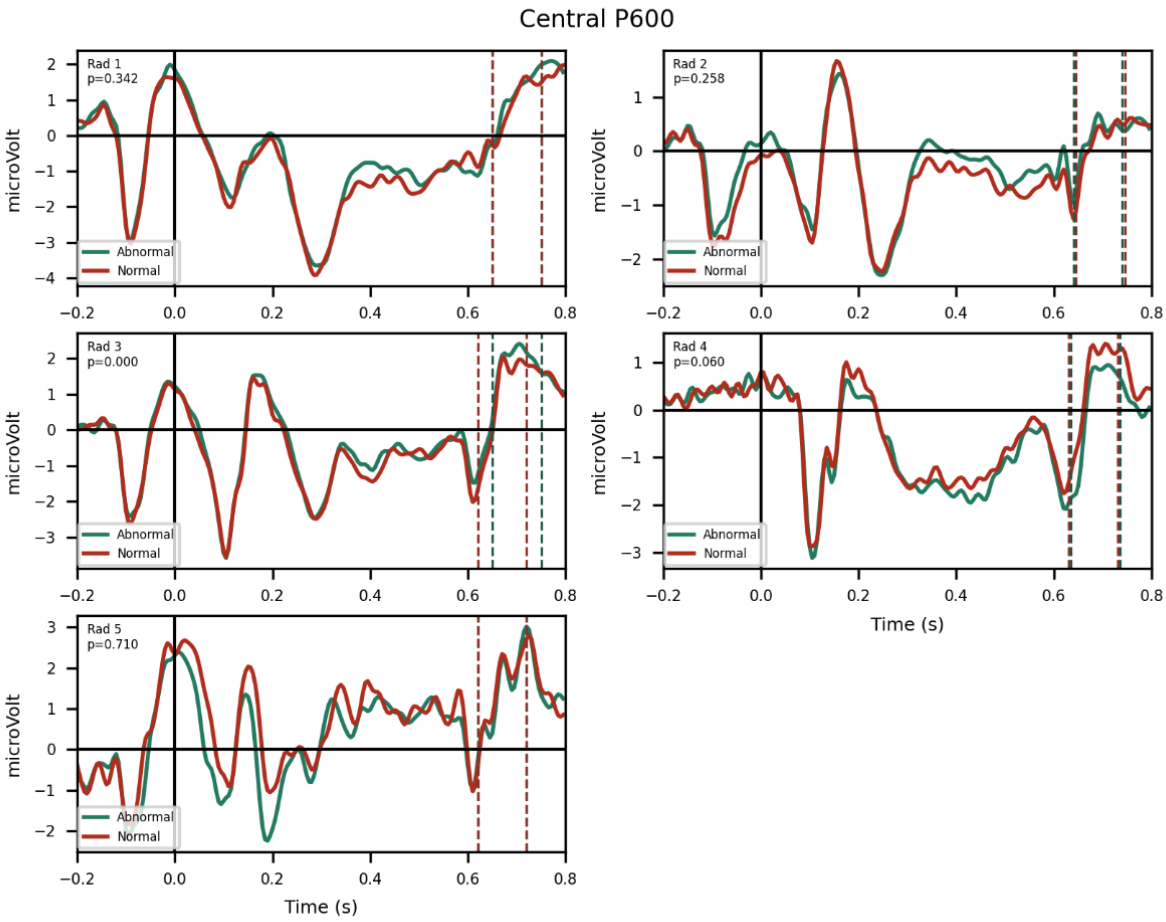


__Figure 5.12:__ Single subject bootstrapping results showing the Central P600 target time windows for abnormal (green dashes) and normal (red dashes), averaged activity, and p-values per radiologist.

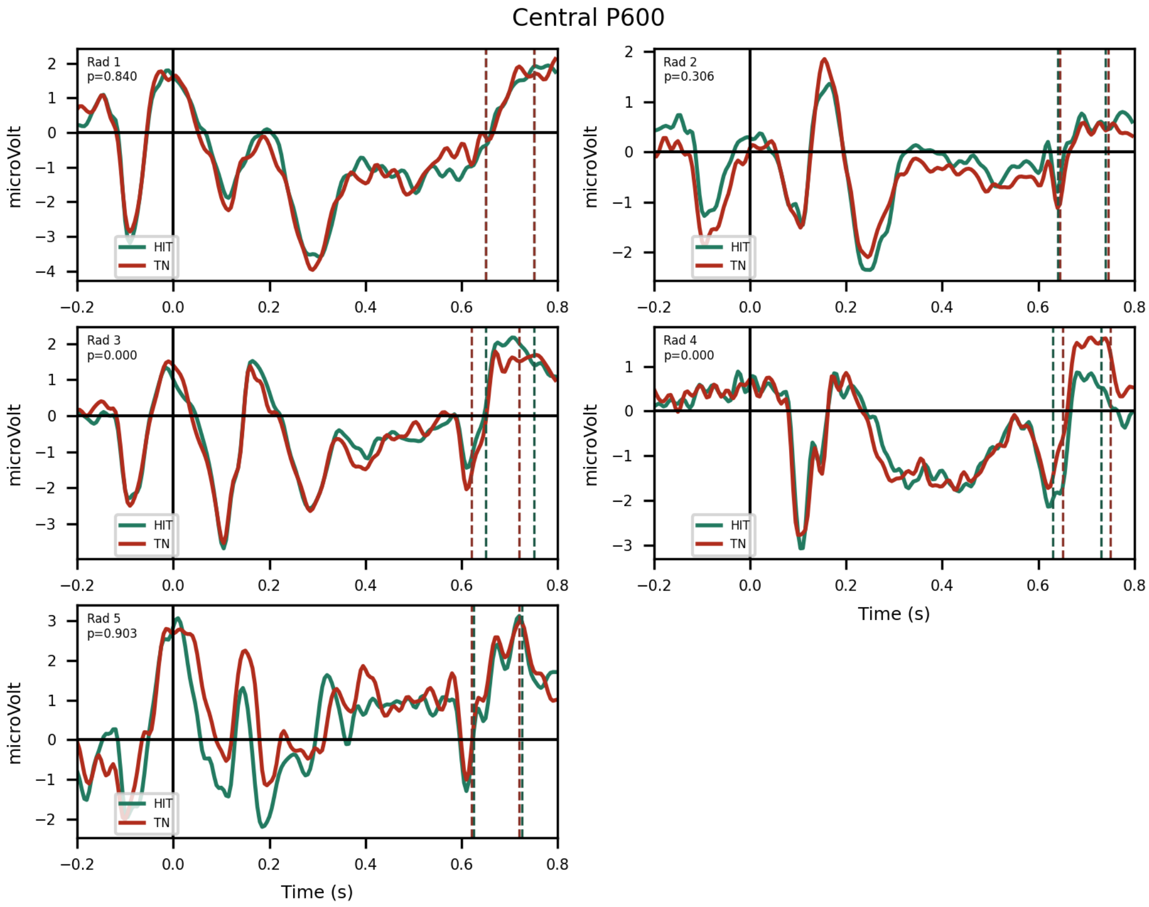


Figure 5.13: *Single subject bootstrapping results showing the Central P600 target time windows for hits (green dashes) and true negatives (red dashes), averaged activity (hits: green, true negatives: red), and p-values per radiologist.*

Frontal cluster

Lastly, for the frontal cluster, there was significant evidence for differential activity between abnormal and normal mammograms for one out of six ERPs, in one radiologist (Table 11). The P600 showed higher activity for normal mammograms. For hits versus true negatives, there was more evidence for differential activity, in three of the six ERPs, in three instances across two radiologists (Table 12). Overall, differential activity in the frontal cluster varied in directionality: the P2, N400, and P600 showed higher amplitude for normals/TNs, whereas the P3a was higher for hits.

Table 5.11: *Results of single subject bootstrapping for the Frontal cluster, showing the p-value and the average difference between abnormal and normal trials in µV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined.*

	N1	**P2**	**N2**	**P3a**	**N400**	**P600**

	p	Δ (μV)	p	Δ (μV)	p	Δ (μV)	p	Δ (μV)	p	Δ (μV)	p	Δ (μV)
1	0.751	0.086	0.968	-0.010	0.144	0.369	0.279	0.273	0.066	0.454	0.204	0.274
2	0.873	0.059	0.795	-0.093	0.910	0.031	0.580	0.159	0.767	0.081	0.644	0.111
3	0.977	0.007	0.965	0.009	0.421	0.228	0.051	0.526	0.749	-0.080	**0.002**	-0.677
4	0.982	-0.001	0.777	-0.083	0.349	-0.292	0.394	-0.253	0.351	-0.262	0.680	0.092
5	0.877	-0.068	0.603	-0.284	0.843	-0.086	0.296	-0.469	0.259	-0.423	0.103	-0.443

Table 5.12: *Results of single subject bootstrapping for the Frontal cluster, showing the p-value and the average difference between hits and true negatives in μV for each of the ERPs per radiologist. P-values under 0.05 are bold and underlined.*

	N1		**P2**		**N2**		**P3a**		**N400**		**P600**	
	p	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**	**p**	**Δ (μV)**
1	0.641	0.157	0.778	-0.096	0.202	0.385	0.852	-0.059	0.813	0.071	0.773	0.082
2	0.493	0.310	0.776	0.129	0.360	0.385	0.855	-0.070	0.696	0.138	0.200	0.381
3	0.106	-0.525	0.498	-0.245	0.231	0.473	**0.019**	0.889	0.583	-0.195	0.547	0.180
4	0.626	-0.208	**0.018**	-0.995	0.068	-0.765	0.353	-0.399	**0.008**	-0.946	0.085	-0.487
5	0.880	0.096	0.610	-0.331	0.357	-0.465	0.382	-0.486	0.320	-0.422	0.765	-0.087

Summary of differential activity

For (ab)normality comparisons, a total of nine instances of significant differential activity were found (6.9% of all comparisons), across four of the five radiologists. For hits-TN, this increased to 18 instances (13.8%), but evidence of differential activity was only observed in three radiologists. Clearly, differential activity was more prominent between hits and true negatives, suggesting that this comparison can tell us more about how/where effective gist extraction takes place. Here, patterns in the observed differential activity will be explored with a focus on hits versus true negatives.

Differential activity between hits and true negatives was observed in at least one instance for each ERP but was most prominent in the P600 (table 13), where evidence was found for 3 radiologists. A similar pattern was observed for (ab)normality comparisons (appendix A), although less clearly so due to the lower number of significant effects. This pattern suggests that differential activity becomes more apparent in later stages of medical gist processing.

Table 5.13: *Total sum and percentage (in brackets) of significant/trending effects per ERP per radiologist for hits vs true negatives. Percentage is calculated based on the number of clusters it was tested at (from Occipital, Parietal, Temporal, Central, and Frontal).*

Rad	P1	N1	P2	N2	P3a	P3b	N400	P600	*Total*
1	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
2	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
3	1 (33.3%)	1 (20%)	0 (0%)	0 (0%)	1 (33.3%)	0 (0%)	0 (0%)	2 (50%)	5 (19.2%)
4	0 (0%)	0 (0%)	3 (75%)	0 (0%)	0 (0%)	2 (100%)	1 (33.3%)	2 (50%)	8 (30.8%)
5	1 (33.3%)	1 (20%)	1 (25%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	1 (25%)	5 (19.2%)
Total	2 (13.3%)	2 (8%)	4 (20%)	1 (10%)	1 (6.7%)	2 (20%)	1 (6.7%)	5 (25%)	18 (13.8%)

Aggregating results across clusters for the hits versus true negatives shows that differential activity was observed in each cluster (table 14). Observations were distributed quite evenly across each of the clusters, suggesting a whole-brain representation of the gist of medical abnormality. Again, a similar pattern was observed for (ab)normality comparisons (appendix B), but without observed effects in occipital cluster and only one effect in the frontal cluster. Together, these results suggest that gist processing takes place across a network of regions, suggesting a distributed representation of the gist of medical abnormality.

Table 5.14: *Total sum and percentage (in brackets) of significant/trending effects per cluster per radiologist for hits vs true negatives. Percentage is calculated based on the number of ERPs that were tested for that cluster.*

Rad	Occipital	Parietal	Temporal	Central	Frontal	*Total*
1	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
2	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
3	1 (25%)	1 (20%)	1 (20%)	1 (16.7%)	1 (16.7%)	5 (19.2%)
4	2 (50%)	3 (60%)	0 (0%)	1 (16.7%)	2 (33.3%)	8 (30.8%)
5	0 (0%)	0 (0%)	3 (60%)	2 (33.3%)	0 (0%)	5 (19.2%)

Total	3 (15%)	4 (16%)	4 (16%)	4 (13.3%)	3 (10%)	18 (13.8%)

Next, individual differences in number of significant findings were observed in this study. While this of course could be caused by differences in data quality, this might also reflect differences in individual neural correlates. Most intriguing, the observed differential activity for radiologist 4 increased drastically from none (0%) for (ab)normality to 10 (38.5%) for hits versus true negatives, which was the highest out of the five radiologists. Interestingly, radiologist 4 had the lowest overall performance (d' 0.224). One might speculate that their gist extraction was relatively weak, resulting in low differential activity especially when looking at all abnormal and normal trials. Potentially, filtering for the limited amount of hits and true negatives allowed the weaker gist signal to be observed. However, it is important to note that this enhancing effect did not occur for radiologist 2, who interestingly had the highest overall performance (d' 0.622).

Lastly, the results section described inverted directionality of observed effects between individuals in the temporal (N1, P600) and central cluster (P2, P600). This prompts the idea that there might be a difference in how individuals represent the gist of medical abnormality, which might affect their neural activity patterns. On the one hand, one might learn the textural and structural elements of normal breast tissue, looking for breaks in this normal structure to identify abnormalities. On the other hand, one might construct representations of abnormal global textures and structural elements that indicate abnormality. These differences in representation might explain the opposite directions of effects when comparing the neural signature of processing the gist of an (ab)normal mammogram, where either normal or abnormal mammograms cause increased EEG activity. Further research is needed to explore the extent of individual differences and potential effects on performance.

Correlations between neural activity and behavioural performance

Investigating the behavioural relevance of observed neural correlates could shed further light on their role in gist extraction. Correlations were calculated between observed differential activity and behavioural performance (d'), or perceptual experience (cases viewed in the past year, years of experience). This section will focus on the later ERPs (P3a, N400, and P600), for the Parietal, Temporal, and Central cluster. Absolute differences in peak activity were used to account for the observed individual differences in direction of effects, as discussed earlier. Thus, a positive correlation means larger differential activity correlated with higher performance or perceptual experience measures. Due to the small sample size, these correlations are heavily exploratory, but it might still show some potentially interesting patterns.

For hits versus true negatives (table 15), the central P3a trended towards a large, positive correlation with d’ (τ=.800, p=.083). The parietal P600 trended towards a large, positive correlation with years of experience (τ=.738, p=.077). Lastly, parietal P600 trended towards a large, positive correlation with years of experience (τ=.738, p=.077). Lastly, the temporal P3a trended towards a large, positive correlation with cases viewed (τ=.738, p=.077). Taken together, while none of the correlations were significant, each of the trends had a positive Kendall’s tau, indicating that the difference in neural activity between hits and true negatives was higher the behavioural measure increased. Most interestingly, the positive correlation between P3a and d’ suggests having a more distinctive central P3a signals for abnormal and normal mammograms was associated with better performance. Future research could further investigate the role of the P3a in medical gist extraction. Effects were found across the parietal, temporal, and central cluster, although only the central cluster correlated with d’, suggesting this region might be especially important for behaviourally relevant elements of medical gist extraction. A similar pattern of results was found for (ab)normality comparisons (appendix C).

Table 5.15: *Kendall rank correlation for correlations between the absolute difference between hits and true negatives and d’, cases viewed, and years of experience. The table shows the Kendall's τ coefficient and p-value for each correlation, with trending values in bold font.*

		D'		**YEARS OF EXPERIENCE**		**CASES VIEWED**	
Cluster	**ERP**	**τ**	**p**	**τ**	**p**	**τ**	**p**
Parietal	**P600**	-0.600	0.233	0.738	**0.077**	-0.105	0.801
Temporal	**P3a**	0.200	0.817	-0.527	0.207	0.738	**0.077**
	N400	0.000	1.000	-0.527	0.207	0.316	0.448
	P600	0.000	1.000	0.527	0.207	0.105	0.801
Central	**P3a**	0.800	**0.083**	-0.316	0.448	0.105	0.801
	N400	0.200	0.817	-0.316	0.448	0.105	0.801
	P600	-0.600	0.233	-0.105	0.801	0.105	0.801

5.5. Discussion

As expected, the behavioural results of this study replicated the previous literature (Patrick C. Brennan et al., 2018; Evans, Georgian-Smith, et al., 2013; Evans et al., 2016), showing that radiologists are indeed able to extract the gist of abnormality in unilateral mammograms containing obvious or subtle abnormalities, as well as mammograms with only global signals of abnormality (the breast contralateral to the abnormality, or taken before the woman went on to develop cancer).

More importantly, evidence for differential activity between abnormal and normal mammograms was found in the EEG signals, showing that the neural signature of extracting the gist of medical abnormality can be detected. Comparing neural activity in trials with abnormal to normal mammograms showed evidence for some differential activity in all tested regions except for the occipital cluster. However, this evidence was sporadic and varied widely across individuals. Comparing hits and true negatives provided more evidence for differential activity, across all five clusters, although this still differed across individuals. Differential activity was spread across the occipital, parietal, central, temporal, and frontal cluster, suggesting a whole-brain representation of the gist of medical abnormality. This suggests that medical gist extraction takes place across a network of regions across different brain areas, rather than being restricted to mainly the occipital area, with a distributed neural signature for performing medical gist categorisation.

In contrast to the current study, differential activity between manmade and natural scenes has mainly been reported in occipital and parietal regions (Groen et al., 2016). This discrepancy might be due to differences in the properties of the gist signal to be extracted. Neural activity in scene gist extraction has been shown to correlate with contrast energy and spatial coherence, the latter of which also correlated with perceived naturalness of the scene (Groen et al., 2013), both low-level spatially pooled summary statistics. For example, spatial coherence indexes scene fragmentation and can be computed by spatial pooling of early visual areas such as the LGN and V1 (Groen et al., 2013). More intermediate level visual properties of scenes such as size and clutter are also represented in neural activity, as shown by whole-brain MEG decoding (Cichy et al., 2017). Photographs of small and large scenes had differential activity patterns from 141 ms after exposure, while clutter-evoked activity differentiated earlier, around 75 ms, and both remained discriminable up to 600 ms after stimulus onset. Sensor-wise decoding illustrated that both size and clutter could be differentiated in occipital, parietal, central, and temporal areas, although it remained the strongest in the occipital and parietal regions. While the scenes and their size/clutter status were not task-relevant, the results by Cichy, Pantazis, and Oliva (2014) do show that these spatial layout properties are represented more widely across the brain. In contrast to manmade and natural scenes, the gist category of a mammogram is unlikely to be easily characterized using the low-level summary statistics described by Groen et al. (2013), especially in the mammograms without any local abnormalities (contralateral, prior). This complexity can also be recognised in machine learning approaches attempting to accurately categorise mammograms, often combining various features to capture aspects of texture features, edge detection, pixel maps, and more, while still not reaching satisfactory performance

(Jalalian et al., 2013; Kurek, Świderski, Osowski, Kruk, & Barhoumi, 2018). However, it is important to note that just because evidence for differential activity in the occipital region was limited, this does not mean that the occipital region is not important for the gist process – but it does demonstrate a contrast with the scene gist findings, and it establishes that the neural signature of the gist of abnormality is characterized by widespread activity. Indeed, it is likely that differentiating abnormal from normal mammograms requires integration of various extracted image properties to perform this complicated categorisation, recruiting a network of areas across the brain.

Evidence for differential activity was found from early (70-150 ms) to late (500-700 ms) time windows, and across traditional visual (P1, N1, P2, N2) and decision making (P3a, P3b) ERPs as well as later ERPs (N400, P600) that were included as exploratory options. This shows that the gist of medical abnormality was discriminable in neural activity early on and remained detectable throughout processing. Previous research suggested the occipital P2 as a scene-selective ERP, that was shown to be sensitive to scene category and image properties (Harel et al., 2016). In the current study, for (ab)normality comparisons, there was no evidence for differential activity in the P2. However, there was some evidence for differential activity between hits and true negatives in the occipital, parietal, central, and frontal cluster. While these effects were only observed in two radiologists, it suggests that the P2 might be sensitive to gist categorisation in general, rather than solely the scene-selective role that was previously suggested.

Surprisingly, the P600 was the ERP with the most evidence for differential activity in the current study. This suggests that the P600 might be a marker for extracting and perceiving the gist of medical abnormality. The P600 has previously been reported to show differential activity across the scalp, with increased P600 amplitude when viewing a face with an incongruent eye-region (Jemel et al., 1999). Such changes to the composition of a face are known to influence the holistic, or global, impression of the face identity (Richler & Gauthier, 2014; Richler, Mack, Gauthier, & Palmeri, 2009). Based on results from Jemel et al. (1999), it could be that the differential activity in the P600 was caused by similarly “incongruent” distortions in abnormal mammograms. On the other hand, the P600 has commonly been associated with grammar and language research. One such study showed that P600 amplitude can be modulated by the saliency and probability of a stimulus, analogue to the P3b (Coulson, King, & Kutas, 1998). Abnormal mammograms with local abnormalities might be more salient. Additionally, the expected probability (prevalence) of abnormal mammograms is very low, for example 0.7% in a large American screening trial (Pisano et al., 2005). So, while abnormality prevalence was higher in the current study, perceptual experience might still influence the P600 response for the low-probability stimulus of an abnormal mammogram. Thus, the differential activity in the P600 might be explained by the detection of a visual incongruency or might be related to the saliency and probability of abnormal mammograms. Future research should further explore the roles of the P2 and P600 in medical gist extraction and could additionally utilize techniques such as multi-voxel pattern analysis or machine learning decoding to gain further insight into overall patterns of activity in more detail than allowed by an ERP-based approach.

A general pattern emerged in the direction of differential effects in this study, which were characterized by higher amplitudes for abnormal or hit trials than the normal or true negative trials. This was the case for the occipital, parietal, and temporal cluster, for both positive and negative ERPs. One exception was the frontal cluster, where the direction of difference varied. It is difficult to compare these findings to manmade/natural scene research, as these scene categories do not directly map to normal or abnormal category equivalents. Animal present/absent studies have reported higher amplitudes in occipito-temporal electrodes for distractors lacking the animal stimulus (Bacon-Macé et al., 2005). VanRullen and Thorpe (2001) isolated task-related differential activity in both animal and vehicle go/no-go tasks, and showed higher occipital and parietal amplitudes for non-targets, and higher frontal and central amplitude for targets from 150 to ~300 ms. These reports of higher occipital and parietal activity for distractors in object-in-scene detection contrast with the current findings of higher activity for abnormal mammograms, if one assumes the abnormal mammogram would be regarded the ‘target’. However, animal/vehicle presence detection is more akin to an object-in-scene detection task. Additionally, the current study used a 2-AFC task, meaning both the “target” (abnormal) and “distractor” (normal) trials required a response, in contrast to the go/no-go design in the previously discussed studies. Thus, direct comparison with these previous studies remains difficult.

Notably, while the general pattern of differential activity was towards higher amplitudes for abnormal/hit trials, the current study also observed instances in the temporal and central cluster with opposite directionality of differential activity between two individuals. This was observed for the temporal (N1, P600) and central cluster (P2, P600). These differences in directionality of effects suggest that there might be individual differences in the way neural correlates represent the different categories of the gist of medical abnormality. Individual ERP variability has been observed in multiple previous studies, although the exact cause is not always known. In general, background EEG activity is thought to influence latency and amplitude of ERPs. For example, variation in amplitude and latency of P300 components can be influenced by ultradian rhythms in background alpha, delta, and theta EEG activity (Anokhin et al., 2001; Polich, 1997). However, at the time of writing, no known studies have reported reversals in directionality of differential activity between individuals like those observed in this study – although this might also be caused by a lack of single-subject analysis in favour of group-level analyses, which would hide individual differences.

Lastly, the current study observed some trending correlations between differential activity and markers of performance, overall medical experience, and recent perceptual experience, in the temporal, central, and parietal cluster. Each of these correlations was positive, with the differential activity and the behavioural measure increasing together. One previous study also reported a positive link between neural activity and performance, as higher differential activity in an occipito-temporal electrode correlated with better performance on an animal go/no-go task (Bacon-Macé et al., 2005), in agreement with the positive correlation observed between d’ and central P3a in this study. Perceptual and medical experience have also both been shown to influence neural activity, although these studies only reported on overall, rather than differential activity. As an example of perceptual experience, training with symbols of a novel script increased overall occipito-temporal N1 amplitude evoked by seeing these symbols in a one-back task (Brem et al., 2018). Medical expertise with EKGs or chest radiographs increased the amplitude of the N170 evoked by evaluation of that medical imaging modality (Rourke, Cruikshank, Shapke, & Singhal, 2016). Medical expertise also correlated positively with activity in the right FFA, but negatively with lateral occipital cortex activity, when rapidly judging local abnormalities in chest radiographs (Harley et al., 2009). Furthermore, studies have shown that resting state fMRI activity differs between radiology interns and age-matched controls (Y. Wang et al., 2021; Zhang et al., 2022), showing the effect of medical and perceptual experience. What’s more, some of these differences in resting state fMRI activity correlated with performance on a separate nodule detection task. However, as mentioned above, it is important to note that these studies did not look at functional differences (differential activity) but instead reported overall ERP amplitude across all cases or resting state activity. Thus, while there is a clear consensus that there are correlations between perceptual experience/behavioural measures and neural correlates, more research is needed into their correlation with differential activity. Any future research should include differential activity measures as was done by Bacon-Macé et al. (2005) and the current study, as this can shed more light on which neural patterns might be functionally related to experience and/or performance.

An obvious shortcoming of this exploratory study is the small sample size, as a consequence of the extremely niche population and necessity of an in-person visit for the EEG measurement. Future research should aim to increase sample size, as this would allow a more traditional group-level analysis to be performed alongside the detailed single-subject bootstrapping used in this analysis. The most obvious approach would be to increase the number of radiologists, for example via collaboration within a consortium of universities and hospitals, or through a portable

EEG device. Alternatively, the training protocol described in Chapter 4 could be used to induce learning of the gist of medical abnormality in naïve participants, which would allow for larger sample sizes, albeit with less strongly established gist signals. What's more, this second approach could provide insight into the neural stages of development of a new gist category if EEG measurements were performed at multiple time points. By combining both approaches, different levels of perceptual expertise could be explored by comparing medical experts to (un)trained naïve participants. This would bring further insight into the different stages of gist processing, distinguishing low-level visual clues from more advanced, emergent properties resulting from expert gist processing, and decisional and semantic markers.

5.6. Conclusion

Differential activity was observed across multiple brain regions, with evidence to suggest that distributed activity across the whole brain is involved in differentiating medical gist categories. The involvement of areas beyond the occipital and parietal regions suggested that the neural signature of medical gist is characterized by distributed activity, in contrast to the occipito-parietal activity reported in previous scene research. This suggest that the gist of medical abnormality requires more complicated integration of different textural and structural regularities, and complex summary statistics that cannot be easily captured in neural activity during early stages of visual processing. Instead, the gist of medical abnormality takes shape across distributed activity in a network of brain regions. Importantly, individual differences were observed, in strength and numerosity of observed effects and even in the direction of differential activity. This suggests there might be differences in the neural signature of medical abnormality categories between individuals. Additionally, differential activity amplitude in some ERPs correlated with performance, overall medical experience, and recent perceptual experience. This suggests that the neural signature of medical gist categories in an individual might be influenced by their medical and perceptual experience, and it might be associated with performance. Overall, the results of this exploratory study suggest that gist extraction of medical abnormality takes place across a network of brain regions, that integrate visual information to construct representations of medical (ab)normality, with individual differences potentially influencing the exact way that medical gist categories are represented in neural activity. Future research should further investigate possible individual differences in differential neural activity as well as the functional role of different ERPs in the network of regions associated with the extraction of the gist of medical abnormality.

5.7. Appendices

Appendix A

Summarising the differential activity for (ab)normality comparisons across ERPs (Table 17) shows that differential activity was present in early (P1, N1), middle (N2, P3a), and late components (P600). This is in line with the observations for the hits vs true negatives comparison (see main text).

Table 5.16: *Total sum and percentage (in brackets) of significant/trending effects per ERP per radiologist for abnormal vs normal. Percentage is calculated based on the number of clusters it was tested at (from Occipital, Parietal, Temporal, Central, and Frontal).*

Rad	**P1**	**N1**	**P2**	**N2**	**P3a**	**P3b**	**N400**	**P600**	***Total***
1	1 (33.3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (3.8%)
2	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (33.3%)	0 (0%)	0 (0%)	0 (0%)	1 (3.8%)
3	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (75%)	3 (11.5%)
4	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
5	0 (0%)	1 (20%)	0 (0%)	1 (50%)	1 (33.3%)	0 (0%)	0 (0%)	1 (25%)	4 (15.4%)
Total	1 (6.7%)	1 (4%)	0 (0%)	1 (10%)	2 (13.3%)	0 (0%)	0 (0%)	4 (20%)	9 (6.9%)

Appendix B

Summarising the differential activity for (ab)normality comparisons across clusters (Table 17) shows that there was no evidence for differential activity in the occipital region. Most evidence was observed for the temporal, central, and parietal cluster, with one instance of differential activity for the frontal cluster. This pattern broadly matches that observed for hits-TN (see main text), in that the parietal, temporal, and central clusters seem to be the main drivers of differential activity, showing a network of activity for the gist of medical abnormality.

Table 5.17: *Total sum and percentage (in brackets) of significant/trending effects per cluster per radiologist for abnormal vs normal. Percentage is calculated based on the number of ERPs that were tested for that cluster.*

Rad	**Occipital**	**Parietal**	**Temporal**	**Central**	**Frontal**	**Total**
1	0 (0%)	1 (20%)	0 (0%)	0 (0%)	0 (0%)	1 (3.8%)
2	0 (0%)	0 (0%)	0 (0%)	1 (16.7%)	0 (0%)	1 (3.8%)
3	0 (0%)	1 (20%)	0 (0%)	1 (16.7%)	1 (16.7%)	3 (11.5%)
4	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
5	0 (0%)	0 (0%)	3 (60%)	1 (16.7%)	0 (0%)	4 (15.4%)

Total	0 (0%)	2 (8%)	3 (12%)	3 (10%)	1 (3.3%)	9 (6.9%)

Appendix C

For the (ab)normality comparison (table 18), differential activity in the central P3a trended towards a large, positive correlation with d’ (τ=.800, p=.083), as was also observed in the hits-TN correlations (see main text). The central N400 correlated significantly with years of experience (τ=.949, p=.023). Lastly, the temporal N400 correlated significantly with cases viewed (τ=.949, p=.023). While these differ from the ERPs observed for hits-TN, the pattern is the same, with positive correlations between differential activity amplitude and the behavioural measures.

Table 5.18: *Kendall rank correlation for correlations between the absolute difference between abnormal and normal trials and d’, cases viewed, and years of experience. The table shows the Kendall's τ coefficient and p-value for each correlation.*

		D'		**YEARS OF EXPERIENCE**		**CASES VIEWED**	
Cluster	**ERP**	**τ**	**p**	**τ**	**p**	**τ**	**p**
Parietal	**P600**	-0.400	0.483	0.316	0.448	-0.527	0.207
Temporal	**P3a**	-0.200	0.817	0.105	0.801	0.105	0.801
	N400	-0.200	0.817	-0.527	0.207	0.949	**0.023**
	P600	-0.400	0.483	-0.105	0.801	0.316	0.448
Central	**P3a**	0.800	**0.083**	-0.527	0.207	0.105	0.801
	N400	-0.200	0.817	0.949	**0.023**	-0.527	0.207
	P600	-0.600	0.233	-0.105	0.801	0.105	0.801

Chapter 6: Discussion

This thesis set out to investigate the processing and acquisition of gist categorisations, primarily focusing on the gist of medical abnormality in mammograms. The introduction of this thesis defined gist extraction as rapid, global, and non-selective, occurring through a set of processes that extract spatial structural regularities and summary statistics as well as basic and intermediate disjunctive features. This thesis aimed to further our understanding of parameters influencing gist extraction through the investigation of the effects of increased exposure time (Chapter 2) and the effects of high spatial frequencies (Chapter 3). Additionally, it was hypothesized that recognising the gist of a category requires the formation of (neuronal) perceptual expectations of the gist properties of that category, as no innate “forest”, “man-made”, or “abnormal mammogram” selective populations of neurons are expected to exist. It was predicted that these perceptual expectations could develop through statistical learning evoked by perceptual exposure alone, without localized feedback. In chapter 4, this thesis investigated whether the gist of medical abnormality could indeed be learned by naïve participants through perceptual exposure with global feedback. Lastly, in chapter 5, the neural signature of gist extraction in medical expertise was investigated in an exploratory study with five expert radiologists. The key findings of each chapter will be summarised below, followed by a discussion of their implications for the medical imaging field.

6.1. Effects of viewing time

The results from chapter 2 suggest that information derived from gist extraction does not accumulate, nor does it decrease over longer periods of viewing time, as performance of expert radiologists rating mammograms did not change significantly when exposure time was unlimited compared to a brief 500 ms. In other words, unlimited exposure time did not result in additional global categorisation information being available to improve performance, nor did the local processing remove the ability of participants to access this global categorisation information. These findings match the results from Evans, Georgian-Smith, et al. (2013) who showed no significant difference in performance on gist of medical abnormality in mammograms with subtle masses or architectural distortions with exposure times of 250, 500, 750, 1000, and 2000 ms. Chapter 2 extends these findings to unlimited exposure time, and broadens them to mammograms with obvious abnormalities, and to priors, mammograms without any localizable abnormalities that went on to develop cancer in the near future.

The lack of improvement with longer exposure time suggests that gist extraction is not a cumulative process that increases in accuracy with longer exposure time. This might be because all available global information is already extracted by the gist processes within the first 500 ms of viewing a mammogram, and further exposure cannot further increase the gist signal. This would fit well with the findings from Evans, Georgian-Smith, et al. (2013) described above. On the other hand, more global information might be available, but is either not accessed or not consciously available to influence perception – maybe because the shift towards selective, local processing decreases the weight given to gist perception.

Importantly, however, performance of radiologists did not diminish with unlimited exposure time either, even when no visible abnormalities were present (in priors). For these priors, a shift away from global gist towards local processing would be expected to decrease performance due to the lack of local abnormalities. This suggests that any further local processing taking place under focused attention during unlimited exposure time did not reduce the ability of radiologists to access their gist-based first impressions. Thus, gist information remained available to the observer even when further processing for local, detailed information took place.

Under the reverse hierarchy model, an initial feedforward sweep creates an automatic or implicit first impression of a more general categorization (the gist), as neurons in these higher visual areas are tuned to higher level categories invariant to scale or position (Hochstein & Ahissar, 2002), allowing recognition of categories from variable viewpoints (Evans & Chong, 2012). Indeed, a feed-forward machine learning model based on the anatomical visual pathways to the prefrontal cortex performed similarly to humans on rapid animal/no-animal scene categorization (Serre et al., 2007), showing that an initial feed-forward sweep can access gist categorisations. Subsequent local processing takes place via re-entrant processing, in which feedback from higher cortical areas returns to lower cortical areas to further refine information through feedback loops with lower cortical regions, allowing for local object binding and focal attention (Hochstein & Ahissar, 2002). The lack of improvement *or* impairment with longer viewing times suggest that the initial feed-forward sweep in which gist is extracted is not impaired by subsequent re-entrant focal processing, as the gist signal remains available with longer exposure.

6.2. Spatial frequency bands

The results from chapter 3 suggest that important information underlying the gist of medical abnormality is contained in specific frequency bands, while the presence of other frequency bands might mask the gist signal with noisy or non-informative content. Filtering out frequencies below 0.5, or below 1.5 cpd both increased the ability of radiologists to recognise the gist of abnormality in priors, mammograms without any localizable actionable lesions that went on to develop cancer in the near future. For the 0.5 cpd filter, this occurred without affecting their gist performance on mammograms with obvious or subtle lesions, while the 1.5 cpd filter caused a slight decrease in performance on subtle abnormalities. On the other hand, 1 or 2 cpd filters provided no such benefits, and instead decreased overall performance. Combined, these findings suggest that 0 – 0.5 cpd contains mainly noise, as removing it increased performance in the 0.5 cpd filter, perhaps through removing LSF breast density signals. While breast density can be a risk factor for breast cancer, previous research found no correlation between BIRAD density and gist of abnormality ratings (Evans, Birdwell, et al., 2013; Evans, Georgian-Smith, et al., 2013; Evans et al., 2016). Next, the 0.5 – 1 cpd contains important spatial structural regularities for the gist signal, as removing it decreased performance in 1 cpd filter. On the other hand, based on these results, 1 – 1.5 cpd seems to contain a mix of noise and gist signal, as removing it in the 1.5 cpd filter increased performance on priors, but decreased performance on subtle lesions. Lastly, the results suggest that the 1.5 – 2 cpd frequency band might contain important information for the gist of abnormality, as the 2 cpd high-pass filter decreased overall performance. Thus, spatial structural regularities that are informative for the gist of medical abnormality seem to be present in 0.5 to 2 cpd, while evidence for noisy signal reducing performance was found for 0 to 0.5 and 1 to 1.5 cpd. These results emphasize the important role of spatial structural regularities in gist extraction, where specific gist categories are likely to be represented by information in specific spatial frequency bands. Our past perceptual experience might form expectations for the spatial structural regularities that represent a category, and this is what then allows observers to rapidly recognise the gist of their visual environment.

Importantly, Chapter 3's results were acquired by a pipeline of spatial frequency filtering, brightness boosting, and contrast normalization. The boosted brightness in the current study was used to ensure that the high-pass filtered parenchyma remained easily visible, as high-pass filters strongly reduce the contrast energy in an image (Perfetto et al., 2020), meaning they have lower overall visibility. By first boosting the overall brightness, this visibility was improved, while the last step of contrast normalization ensured that the high-pass filtered mammograms contained the same contrast levels as the unfiltered mammograms. A potential concern would be that these steps might disturb the 'naturalistic' state of a mammogram. However, mammograms are inherently 'unnatural', as they are 2D visualizations of the x-ray absorption levels of 3D tissue: the visibility of specific tissues in x-rays depends on the specific machine, settings, image processing used (Cole et al., 2005), and even the practitioners' preferential compression force (Mercer et al., 2013). In the clinic, radiologists even use programs to change the contrast or brightness of the cases as they view them. Thus, the brightness increase does not create an unrealistic stimulus, and was conductive to the main aim of the chapter of finding spatial frequency manipulations that could enhance detection rates.

Previous studies support the importance of contrast normalization when investigating the roles of spatial frequency bands in gist extraction. Without contrast normalization, gist detection of obvious and subtle abnormalities dropped slightly for both high-pass and low-pass filtered mammograms (Evans et al., 2016), although the decrease in performance was much more pronounced for low-pass filtered mammograms. Thus, the previous findings by Evans et al. (2016) suggested that more gist information was retained in the HSF than LSF mammograms even before contrast normalization took place. Chapter 3 extends on these findings by showing that high-pass filters combined with contrast normalization retain overall performance as compared to the full spectrum mammograms. The effect of contrast normalization can also be observed in scene gist research. Gist performance was reduced in HSF scenes without contrast normalization, but contrast normalization equalized performance between LSF and HSF scene images (Perfetto et al., 2020). Thus, contrast normalization is essential to reveal the full effects of high-pass filtering in both mammograms and scenes.

More generally, conflicting effects of spatial frequency content have been reported in scene processing. The earliest findings suggested that scene perception of hybrid images relied mostly on LSF with 30 ms viewing time (Schyns & Oliva, 1994), however, this effect swapped to a preference for the HSF scene content at 150 ms viewing time, and what's more, both LSF and HSF content already had a priming effect at 30 ms viewing time. Additionally, these hybrid images were not naturalistic stimuli, as the mixing of spatial frequencies breaks contiguity of contours and edges in the HSF spectrum. Additionally, follow-up research showed that the most recent exposure (either LSF or HSF) influenced the perceived category of a hybrid scene (Oliva & Schyns, 1997), suggesting that our visual system might be able to flexibly tune its sensitivity to expected or relevant spatial frequencies when a full spectrum image is presented. What is more, neuroimaging largely points to the importance of HSF in evoking early visual activity (Kauffmann et al., 2015; Rajimehr et al., 2011) and more interestingly, decoding scene categories from this activity in many scene-selective areas (Berman et al., 2017; Walther et al., 2011). Lastly, a perceptual study reported no significant differences in performance between LSF and HSF scene images (Perfetto et al., 2020). Thus, the more recent findings in wider scene research suggest that the role of spatial frequency bands might depend on the specific task, recent exposure, and expectations, but does suggest that HSF plays an important role in encoding scene category information.

Taken together, both LSF and HSF are available to the visual system early after stimulus onset, but HSF might be especially important for gist extraction processes, especially where only difficult, global signals are present. Interestingly, HSF is able to enhance the gist of abnormality in prior mammograms that go on to develop cancer in the near future, potentially by removing noise from breast density signals in the 0 to 0.5 cpd frequency band.

Future research could use bandpass or bandstop filters to narrow down the effects of the absence or presence of more specific spatial frequency bands, rather than the high-pass filters used in the current study. These bandpass/bandstop filters can selectively isolate *or* filter out a small band of frequencies. This would allow for more controlled adjustment of frequency content and could help identify the exact combination of spatial frequencies that contain the gist of abnormality in mammograms. This could for example be used to filter out F0 – F0.5 and F1-F1.5 frequency bands to investigate whether this combination further enhances the gist signal or could remove/isolate smaller sub-sections of the frequency bands that were identified in this chapter.

Additionally, future research should investigate whether the same or similar effects of high-pass filters can be observed in other medical imaging modalities, such as digital breast tomosynthesis, chest radiographs, or even micrographs or dermatological images. This would provide insight into whether the gist of medical abnormality is present in the same spatial frequency bands across modalities, or whether this is modality specific. Modality specific spatial frequency signatures of the gist of abnormality would be expected, as the ability to extract the gist of abnormality in one domain does not translate to another, for example between mammograms and micrographs of cervical cells (Evans, Georgian-Smith, et al., 2013).

6.3. Statistical learning through perceptual exposure

The results from chapter 4 suggest that the gist of a new category can be learned through perceptual exposure with global feedback, although there might be individual differences in speed and/or ability to learn to recognise the new gist signals. Splitting the group of naïve observers into learners and non-learners based on their training sessions revealed a significant improvement between the pre- and post-training tests for the learners, while there was no significant change for the non-learners. For learners, performance improved across different conspicuities of mammograms, including those without visible abnormalities (contralaterals, priors). The fact that performance improved for these very difficult cases without localizable abnormalities supports the notion that these observers learned to recognise a global textural signal representing the gist of medical abnormality.

While there generally is a lack of research into the learning of a new gist categorisation, the existing results largely match the findings of Chapter 4. Firstly, a study by Hegdé (2020) showed that participants learned to detect the presence/absence of obvious abnormalities in mammograms after untimed perceptual training with global feedback (Hegdé, 2020). This study did not limit viewing time in training nor tests, but its findings support that learning to detect abnormalities in mammograms can occur through perceptual exposure alone. What's more, previous research has shown that observers could learn to detect the presence of camouflaged targets in textures through rapid perceptual training (500 ms viewing time) with global feedback, going from chance-levels to d’s of ~1.2 after approximately 7200 trials of training (X. Chen & Hegdé, 2012). This rapid detection of the presence or absence of a camouflaged target might make use of information extracted by gist processes to rapidly detect whether the global structural regularities of the texture are intact or not. Thus, while not directly a gist extraction task, X. Chen and Hegdé (2012) findings fit with the view that learning of spatial structural regularities can occur through rapid, perceptual training with global feedback.

Furthermore, a recent study suggested that observers were able to learn the gist of medical abnormality in skin histology images after only a brief training (~258 exemplars) (DiGirolamo et al., 2023). But, while participants were tested on a gist task (500 ms), training used long viewing times (up to 24 seconds) and outlined the area of the image containing the skin pathology in the feedback. Interestingly, participants were also able to distinguish the four different skin pathologies within the “abnormal” category with slightly above-chance accuracy after training, suggesting that these pathologies had distinct characteristics that could be recognised within 500 ms. However, the article did not report no baseline performance, nor expert performance on the same task. This makes it difficult to contextualize the observed above-chance performance after training. Furthermore, the histology images all contained localisable abnormalities, which are expected to occur predominantly in the centre of the image, and the experiment did not use an equivalent to the contralaterals/priors used in mammography gist research. Skin histology categorisations might be easier to perform pre-training, or might be more uniform in appearance, allowing simple perceptual rules to be used for their categorisation. This might make the skin histology task more akin to the sexing of chicks. Sexing of chicks is initially difficult (60.5% correct), but observers reached near-expert performance (84% correct) after brief instruction from experts (Biederman & Shiffrar, 1987), suggesting that this task could be learned through instructions on specific exemplars. Skin histology images might similarly contain such key features. However, despite these shortcomings, the methodology and results of DiGirolamo et al. (2023) pose some interesting areas for future research. The ability of radiologists to recognise abnormality subtypes in mammograms in a gist extraction paradigm has not previously been investigated, so it is unknown whether for example histological subtypes of breast cancer (see Makki (2015) for a review) have distinguishable gist characteristics. Thus, it would be interesting to investigate whether radiologists can detect certain subcategories of abnormalities in mammograms – and perhaps even in priors or contralaterals. However, it is very possible that the characteristics of skin pathologies in histology images are more distinctive than histological subtypes of breast carcinoma in mammograms. If abnormality gist subcategories can be distinguished in mammograms, it would be interesting to further investigate the gist properties of these subtypes using image analysis, as well as further investigate the effects of training paradigm on the speed of learning and eventual strength of learned signals, given the many differences in training paradigm used in each of the discussed studies (X. Chen & Hegdé, 2012; DiGirolamo et al., 2023; Hegdé, 2020).

In addition to the potential influence of training paradigm, the subdivision of learners and non-learners in Chapter 4 suggests that there are also individual differences in ability to learn a new category of gist. However, it is not known whether non-learners were fully unable to learn to recognise the gist of abnormality, or if they simply needed a longer training period before learning would be visible. Previous studies show some evidence for individual differences in learning speed. For example Hegdé (2020) reported a wide variability in the length of training needed for participants to reach a performance threshold, varying from 288 to 936 trials. More generally, individual differences have been reported both in speed of perceptual learning (Maniglia & Seitz, 2018; Rotman, Lavie, & Banai, 2020; Waller, 2000), and in perceptual capabilities, such as visual search (Brock, Xu, & Brooks, 2011; Sobel, Gerrie, Poole, & Kane, 2007; M.-J. J. Wang, Lin, & Drury, 1997). Individual differences in learning rates also influence spatial learning in virtual environments (Waller, 2000). Additionally, the influence of individual differences on learning rate was calculated to be 36.8% across a range of visual and auditory perceptual tasks (Yang et al., 2020), comparable to the 38.6% for task-specific factors, supporting the substantial influence of individual differences observed in Chapter 4.

Individual differences might be innate, due to previous experiences, or might even be related to differences in strategy. For example, innate differences in cortical thickness of relevant brain areas correlated with learning rates of a motion discrimination visual search task (Frank et al., 2016) as well as a face view discrimination task (Bi et al., 2014). On the other hand, previous perceptual experience might also influence learning rates, as people with more previous gaming activity were found to have higher general perceptual learning rates (Bavelier et al., 2012; Bejjanki et al., 2014).. Lastly, strategy or general motivation could influence learning rates as well. Learners might have been more motivated, or might have used a more global strategy, while non-learners might have used mal-adaptive strategies, such as focusing on local signals. Indeed, Previous research suggested that learners and non-learner groups utilized different strategies while being trained on a difficult grating orientation task (Dobres & Seitz, 2010). Exploring differences in previous perceptual experiences, neural anatomy, and strategies employed by learners and non-learners in a gist learning task could provide valuable insight into which fixed (innate) and flexible (previous experience/strategy) factors influence the learning of a new gist category.

Individual differences in performance are also present in medical experts. Performance on a gist task correlated with recent perceptual exposure (cases read in the previous year) with an R^2 of 0.2 (Evans et al., 2019). Similarly, specificity on a untimed mammography screening task in laboratory conditions correlated strongly with annual reading volumes, while there was no such correlation with their ability to correctly localize abnormalities (Rawashdeh et al., 2013). Thus, radiologists' performance in both gist and screening tasks were affected by recent previous perceptual exposure, supporting the idea that gist extraction performance is influenced by the amount of recent exposure. However, perceptual exposure did not fully explain individual differences in gist performance, suggesting there are other factors influencing individual performance. Further research should investigate which factors explain individual differences in learning speed and perceptual capabilities. One potential avenue would be to administer a testing battery of general visual processing tasks to see how performance on simple laboratory tasks correlates with training speed and performance on a gist task. Additionally, it would be especially interesting to further investigate explanatory variables for individual differences between radiologists, as this might provide further insight into factors influencing gist extraction and might allow for selection of radiologists with a strong gist signal for risk assessment.

Where learning did take place in Chapter 4, it was unfortunately not strongly retained. The performance of learners did not differ significantly from baseline after seven to ten days of perceptual inactivity. Thus, the gist characteristics of the newly learned category of medical abnormality were not sufficiently encoded to be retained over longer periods of time. It is possible that continuous regular exposure is required to retain the gist of a category, but a gist category might also become more stable after more prolonged periods of exposure. After all, it is unlikely that someone would be unable to categorize a scene image of a mountain, if they had not seen one for 7 days, or even a longer period of time. For context, it is important to keep in mind that the nine sessions of training amounted to only eight hours of exposure across nine to twenty days, followed by at least seven days of no exposure. So, while the training corresponded to viewing and rating almost 6500 instances of a mammogram, this is still relatively minimal compared to our life-long exposure to scene categories or the years of practice that expert radiologists have. Additionally, radiologists' gist performance has been shown to correlate with cases reviewed in the previous year, but not years of experience (Evans et al., 2019), enforcing the idea that continued perceptual experience is important for retaining the gist of medical abnormality.

Other literature on the retention of statistical perceptual learning is relatively scarce. Previous research has shown that statistically learned shape sequences are retained for at least 24 hours, resulting in faster reaction times for the second and third shape of a triplet in an RSVP (Kim, Seitz, Feenstra, & Shams, 2009), as well as above-chance performance on a triplet recognition task (Arciuli & Simpson, 2012). What's more, performance on Arciuli and Simpson (2012)'s triplet recognition task did not differ significantly after 30 minutes, 1, 2, 4, or 24 hours, indicating that the statistical learning was relatively stable and consistent over time. Auditory tone patterns can similarly be retained for 4 and 12 hours (Durrant, Taylor, Cairney, & Lewis, 2011). However, no known research tested the extended retention of perceptual learning paradigms to the level of the 7-day interval used in Chapter 4. Thus, further research of retention of perceptual statistical learning in general, and especially for gist extraction, is needed. In addition to further studies on the retention after perceptual training in naïve participants, future research could investigate the retention of the gist of abnormality in medical experts that are retiring or changing careers to explore longer term retention, or in medical experts going on a brief hiatus of practice, such as a holiday, for shorter term retention. The former carries the risk of being contaminated by effects of cognitive decline with aging, which would require appropriate control subjects, but a combination of both short- and long-term retention in medical experts would provide a fascinating insight into the retention of a consolidated gist category that is hard to get from training naïve participants.

6.4. Neural signature of gist extraction in experts

In Chapter 5 the neural signature of extracting the gist of medical abnormality was explored in five expert radiologists. This study focused on differential activity, as this provides insight into where and when in the brain gist categories are distinctly represented in the neural activity. The results showed evidence for differential activity between normal and abnormal mammograms, across a distributed network of areas and ERPs, suggesting that gist extraction for medical abnormality takes place in a distributed fashion. Differential activity was observed across the occipital, parietal, central, temporal, and frontal cluster, and generally was caused by higher amplitudes for abnormal mammograms than for normal mammograms. Differential activity was observed throughout early and late ERP components, including the P2 that was previously described as a scene-selective ERP (Harel et al., 2016), and most prominently in the P600. Further neuroimaging studies should examine the roles of these ERPs in extracting the gist of medical abnormality. Additionally, more advanced techniques such as computational decoding or multi-voxel pattern analysis of trials should be used in concert with traditional group-level and individual-level ERP analyses.

However, there were also some instances where there were individual differences in the direction of the differential activity, meaning that the same ERP in the same cluster showed higher activity for hits in one radiologist, but higher activity for true negatives in another. Opposing directionality was observed in the temporal (N1, P600) and central cluster (P2, P600).

These findings suggest that the neural signature of the gist of medical abnormality might vary between individuals at some timepoints of neural processing. As there were some correlations between behavioural and expertise measures (d', years of experience, number of cases viewed last year), it is also possible that there is an influence of (recent) perceptual experience on the neural signature of the gist of medical abnormality. As this finding was based on an exploratory study of only five radiologists, a more in-depth investigation of individual differences in gist extraction is needed. Additionally, it is unknown whether there is similar individual variation in the neural signature of scene gist categorisations, such as natural vs man-made, which could be addressed in future research.

Lastly, it would be interesting to combine the training protocol described in Chapter 4 with the EEG measurements of Chapter 5 to investigate the neural activity of a gist task before and after perceptual training. This might also provide additional insights into the individual differences observed in both chapters, by exploring neural differences between learners and non-learners. It would be especially interesting to explore if training brings about distributed representations throughout the cortex, and to look at the P2 and P600 activity that was most prevalent in the radiologists in Chapter 5.

6.5. Implications for medical imaging

Medical images acquired in a screening process, such as mammograms from breast screening, need to be reviewed by a medical expert to assess whether there are any suspected abnormalities. If sufficient evidence of a potential abnormality is found, the patient is referred to follow-up procedures such as a biopsy. As such, screening is a time intensive process, that relies heavily on the first step accurately detecting cases with suspicious abnormalities. If an abnormality is missed, this can have disastrous consequences for the patient, as early detection is vital to increase chances of positive health outcomes (Coleman, 2017), and screening is relatively infrequent (3 years in the UK (NICE, 2017)). However, incorrect referral for follow-up also has a negative impact through both healthcare costs (Chubak et al., 2010) and the mental impact on the patient (Jatoi et al., 2006; Sandin et al., 2002). While the exact impact is incredibly difficult to estimate, a meta-analysis estimated that breast screening in the UK led to a 20% reduction in mortality, with approximately 11-19% of the cancers diagnosed constituting overdiagnosis (Marmot et al., 2013). Thus, assessing screening cases requires the medical expert to strike a fine balance between being scrupulous/selective, and vigilant in each case.

It might seem like a simple solution to simply increase screening frequency, but this has multiple disadvantages. Firstly, it increases the burden on the screening process. Secondly, it increases healthcare costs and the chance of overdiagnosis. Thirdly, it increases the radiation dose each woman receives, which can slightly increase risk of radiation-related cancer (De Gonzalez, 2011; L. M. Warren, Dance, & Young, 2016). Instead, a balance needs to be struck between identifying at-risk women for further scrutiny or screening prioritization.

Here, the gist of medical abnormality comes into play as a potential cost-effective way of identifying mammograms that contain subtle signs of medical abnormality, either because the breast currently contains subtle cancer that was missed or because the woman is at risk of developing cancer in the near future. Mammograms with a high gist of abnormality score could be reassessed for any missed cancers by a different medical expert. This would work similar to the double reading used in for example the UK, which has considerable benefits in detecting additional cancers that would have otherwise been missed (Patrick C Brennan et al., 2019; Ciatto et al., 2005; Dinnes et al., 2001; R. Warren & Duffy, 1995). If no currently localizable abnormalities are found, the woman could be invited for more frequent screening, as a high gist score indicates increased risk of developing cancer in the near future (priors). This would be analogous to the way women with certain genetic markers increasing the risk of developing breast cancer are currently prioritized for more frequent screening (Pruthi et al., 2010), except that it would be applicable to the entire population without requiring costly genotyping. However, if a gist of abnormality risk rating were to be implemented in practice, every effort should be made to optimize the process. The results from this thesis form the basis for several suggestions for such optimizations to be considered, and recommendations for future research into potential further avenues of improvement.

Firstly, the findings from chapter 2 suggest that 500 ms viewing time followed by a gist rating is sufficient for a medical expert to extract the gist of medical abnormality. And even in the unlimited viewing condition, the average time a radiologist spent on a case was only 5.5 ± 1.9 (95% CI) seconds. While further research could investigate any differences in mental effort, fatigue, or other mental effects of rapid versus unlimited viewing time to ensure the burden on medical experts is minimised, these findings support the idea that gist scoring of mammograms could be a time-efficient method. This considerable speed might even allow gist scores to be collected from multiple radiologists and aggregated into one risk factor, that could help prioritize at-risk women for more frequent screening.

Secondly, the results from chapter 3 suggest that the gist signal in mammograms at risk of future cancers (priors) can be enhanced by removing spatial frequencies below 0.5 cpd and increasing the brightness of the remaining high-frequency signals. Since this image enhancement did not affect overall performance, and significantly boosted detection of future risk, it would greatly increase the usefulness of gist risk factor scoring. Of course, future research could further enhance this effect by fine-tuning and optimizing the spatial frequency filters to be used, for example bandstop filters which can be used to remove smaller bands of frequencies between two cut-off points. Additionally, programs used to view mammograms often allow radiologists to modify brightness and contrast (Pisano et al., 2005). Previous research has proposed various algorithms for contrast enhancement (Jenifer, Parasuraman, & Kadirvelu, 2016; Tripathy & Swarnkar, 2020), however, these methods are often only evaluated based on computed properties (e.g. contrast improvement index) rather than on radiologists' performance. What's more, differences in the neural signature of gist extraction suggest that there are differences in the representation of the gist of medical abnormality even after years of perceptual exposure. It is possible that individual differences also influence the effectiveness of image enhancements, especially as for example contrast sensitivity differs is known to differ between individuals (Owsley, Sekuler, & Siemsen, 1983; D. Peterzell, Werner, & Kaplan, 1991; D. H. Peterzell, Werner, & Kaplan, 1995). Thus, future research might want to explore the effects of combining spatial frequency filters with other types of image enhancements, such as contrast enhancement or brightness increases, in order to find the combination that most increases radiologists' performance, while keeping an eye out for potential individual differences.

Thirdly, by showing that non-experts can be trained on the gist of medical abnormality, this thesis opens the door to considerations of developing further training to improve the accuracy of gist extraction in medical experts. Indeed, an interesting question is whether perceptual training can further boost the ability of radiologists to extract the gist of abnormality – although it is possible radiologists are already performing near the peak of their ability. Alternatively, it might not be time efficient to perceptually train radiologists to increase their gist performance and use them for gist screening. With perceptual training it might be possible to outsource the gist scoring of mammograms to perceptually trained, rather than medically trained individuals. Radiographers, also known as radiologic technologists, have some pre-existing perceptual experience with mammograms as they are responsible for taking the x-ray images. This might make them better primed to learn to detect the gist of abnormality than general population observers. Indeed, radiographers are able to accurately detect cancer in screening mammograms or function as a second reader to a consultant radiologist, after a short accredited training course (Van den Biggelaar, Nelemans, & Flobbe, 2008; Wivell, Denton, Eve, Inglis, & Harvey, 2003), illustrating the ability of radiographers to quickly pick up perceptual signals in mammograms. Thus, to increase the feasibility of a gist screening program researchers should consider developing a gist training program for radiographers. In developing training paradigms, researchers should remain aware of potential individual differences. These individual differences might affect the speed of learning or even the level of performance someone is able to reach.

Lastly, while this thesis focused on the gist of abnormality in mammograms, it should not be taken to mean that these findings only apply to this image modality. The gist of abnormality has been reported in various other medical imaging modalities, underlining that this is a general perceptual ability, not a special property of mammograms. Gist of abnormality has been reported in digital breast tomosynthesis (C. C. Wu et al., 2019), chest radiographs (Carmody et al., 1981; Kundel & Nodine, 1975), skin pathology (Brunyé et al., 2021; DiGirolamo et al., 2023), and even pap test images (micrographs) of cervical cells (Evans, Georgian-Smith, et al., 2013). While the specifics of for example the spatial frequency filters that enhance the gist signal in each medical modality might vary, the general 'gist' of the findings in this thesis might very well apply to other modalities.

6.6. Conclusion

In conclusion, this thesis adds to the body of work showing that gist extraction takes place through rapid extraction of global visual information, as it occurs even when no local abnormalities are present in the mammogram. It was also shown that, in addition to occurring rapidly, gist information remains available for guiding perception even when further selective local processing becomes available. Furthermore, this thesis highlights the importance of higher spatial frequency information for the gist of medical abnormality. It also provides the first evidence for perceptual learning of the gist of a new category, showing that some observers can learn the gist of abnormality by receiving global feedback, although the learning is poorly retained and influenced by individual differences. The neural signature of extracting the gist of medical abnormality indicates the use of a distributed network of cortical regions, with potential individual differences in how gist categories are represented. The findings of this thesis also have important implications for medical image processing, as they provide ways to boost the gist of medical abnormality in mammograms through spatial filtering, as well as a first indication for the possibility of a perceptual training paradigm, which could be used to boost the extraction of the gist of medical abnormality in residents or radiographers. Together, the findings of this thesis inform ways that the gist of medical abnormality could be utilized as a risk factor in the medical toolbox, as well as providing further insight into gist extraction in general.

Thesis Abbreviations

AIC: Akaike Information Criterion

AUC: Area under the curve

BF: Bayes factor

CC: Craniocaudal

CPD: Cycles per degree

DNN: Deep neural network

EEG: Electroencephalogram

ERP: Event related potential

HSF: High spatial frequencies

LOC: lateral occipital complex

LSF: Low spatial frequencies

MEG: Magnetoencephalography

MLO: Mediolateral oblique

OPA: occipital place area

PPA: parahippocampal place area

RHT: Reverse hierarchy theory

ROC: Receiver operator curve

RSC: retrosplenial cortex

RSVP: Rapid serial visual presentation

SBC: Single breast classifier

SBC+HM: Single breast classifier plus heatmap

References

Aberg, K. C., & Herzog, M. H. (2012). Different types of feedback change decision criterion and sensitivity differently in perceptual learning. *Journal of Vision, 12*(3), 3-3.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control, 19*(6), 716-723.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological science, 19*(4), 392-398.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences, 106*(18), 7345-7350.

Anokhin, A. P., Van Baal, G., Van Beijsterveldt, C., De Geus, E., Grant, J., & Boomsma, D. (2001). Genetic correlation between the P300 event-related brain potential and the EEG power spectrum. *Behavior Genetics, 31*(6), 545-554.

Antal, A., Kéri, S., Kovács, G., Liszli, P., Janka, Z., & Benedek, G. (2001). Event-related potentials from a visual categorization task. *Brain Research Protocols, 7*(2), 131-136.

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is lasting and consistent over time. *Neuroscience Letters, 517*(2), 133-135.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological science, 12*(2), 157-162.

Attarha, M., & Moore, C. M. (2015). The perceptual processing capacity of summary statistics between and within feature dimensions. *Journal of Vision, 15*(4), 9-9.

Bacon-Macé, N., Macé, M. J. M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision research, 45*, 1459-1469. doi:10.1016/j.visres.2005.01.004

Balas, B., & Conlin, C. (2015). Invariant texture perception is harder with synthetic textures: Implications for models of texture processing. *Vision research, 115*, 271-279.

Balas, B., Conlin, C., & Shipman, D. (2016). Summary Statistics and Material Categorization in the Visual Periphery. *ACM Transactions on Applied Perception (TAP), 14*(2), 1-13.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision, 9*(12), 13-13.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*, 617-629. doi:10.1038/nrn1476

Bastin, J., Vidal, J. R., Bouvier, S., Perrone-Bertolotti, M., Bénis, D., Kahane, P., . . . Epstein, R. A. (2013). Temporal components in the parahippocampal place area revealed by human intracerebral recordings. *Journal of Neuroscience, 33*(24), 10123-10131.

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., & Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences, 117*(48), 30071-30078.

Bavelier, D., Green, C. S., Pouget, A., & Schrater, P. (2012). Brain plasticity through the life span: learning to learn and action video games. *Annual review of neuroscience, 35*, 391-416.

Bejjanki, V. R., Zhang, R., Li, R., Pouget, A., Green, C. S., Lu, Z.-L., & Bavelier, D. (2014). Action video game play facilitates the development of better perceptual templates. *Proceedings of the National Academy of Sciences, 111*(47), 16961-16966.

Berman, D., Golomb, J. D., & Walther, D. B. (2017). Scene content is predominantly conveyed by high spatial frequencies in scene-selective visual cortex. *PloS one, 12*, 1-16. doi:10.1371/journal.pone.0189828

Bernardi, D., Ciatto, S., Pellegrini, M., Anesi, V., Burlon, S., Cauli, E., . . . Targa, L. (2012). Application of breast tomosynthesis in screening: incremental effect on mammography acquisition and reading time. *The British journal of radiology, 85*(1020), e1174-e1178.

Berns, E. A., Hendrick, R. E., Solari, M., Barke, L., Reddy, D., Wolfman, J., . . . Willis, L. (2006). Digital and screen-film mammography: comparison of image acquisition and interpretation times. *American Journal of Roentgenology, 187*(1), 38-41.

Bi, T., Chen, J., Zhou, T., He, Y., & Fang, F. (2014). Function and structure of human left fusiform cortex are closely associated with perceptual learning of faces. *Current Biology, 24*(2), 222-227.

Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 640.

Bird, R. E., Wallace, T. W., & Yankaskas, B. C. (1992). Analysis of cancers missed at screening mammography. *Radiology, 184*(3), 613-617.

Boucart, M., Moroni, C., Thibaut, M., Szaffarczyk, S., & Greene, M. (2013). Scene categorization at large visual eccentricities. *Vision research, 86*, 35-42.

Boyd, N. F., Martin, L. J., Bronskill, M., Yaffe, M. J., Duric, N., & Minkin, S. (2010). Breast tissue composition and susceptibility to breast cancer. *Journal of the National Cancer Institute, 102*(16), 1224-1237.

Brady, T. F., & Oliva, A. (2007). Statistical learning of temporal predictability in scene gist. *Journal of Vision, 7*(9), 1050-1050.

Brainard, D. H. (1997). The psychophysics toolbox.

Brem, S., Hunkeler, E., Mächler, M., Kronschnabel, J., Karipidis, I. I., Pleisch, G., & Brandeis, D. (2018). Increasing expertise to a novel script modulates the visual N1 ERP in healthy adults. *International Journal of Behavioral Development, 42*(3), 333-341.

Brennan, P. C., Gandomkar, Z., Ekpo, E. U., Tapia, K., Trieu, P. D., Lewis, S. J., . . . Evans, K. K. (2018). Radiologists can detect the ‘gist’of breast cancer before any overt signs of cancer appear. *Scientific reports, 8*(1), 1-12.

Brennan, P. C., Ganesan, A., Eckstein, M. P., Ekpo, E. U., Tapia, K., Mello-Thoms, C., . . . Juni, M. Z. (2019). Benefits of independent double reading in digital mammography: a theoretical evaluation of all possible pairing methodologies. *Academic Radiology, 26*(6), 717-723.

Brock, J., Xu, J. Y., & Brooks, K. R. (2011). Individual differences in visual search: Relationship to autistic traits, discrimination thresholds, and speed of processing. *Perception, 40*(6), 739-742.

Brunyé, T. T., Drew, T., Saikia, M. J., Kerr, K. F., Eguchi, M. M., Lee, A. C., . . . Elmore, J. G. (2021). Melanoma in the blink of an eye: Pathologists’ rapid detection, classification, and localization of skin abnormalities. *Visual cognition, 29*(6), 386-400.

Cant, J. S., & Xu, Y. (2017). The contribution of object shape and surface properties to object ensemble representation in anterior-medial ventral visual cortex. *Journal of cognitive neuroscience, 29*(2), 398-412.

Carmody, D. P., Nodine, C. F., & Kundel, H. L. (1981). Finding lung nodules with and without comparative visual scanning. . *Perception & psychophysics, 29*, 594-598.

Carrigan, A. J., Wardle, S. G., & Rich, A. N. (2018). Finding cancer in mammograms: if you know it’s there, do you know where? *Cognitive research: principles and implications, 3*(1), 10.

Chen, W., & Samuelson, F. W. (2014). The average receiver operating characteristic curve in multireader multicase imaging studies. *The British journal of radiology, 87*(1040), 20140016.

Chen, X., & Hegdé, J. (2012). Learning to break camouflage by learning the background. *Psychological science, 23*(11), 1395-1403.

Chin, M. D., Evans, K. K., Wolfe, J. M., Bowen, J., & Tanaka, J. W. (2018). Inversion effects in the expert classification of mammograms and faces. *Cognitive research: principles and implications, 3*(1), 31.

Chong, S. C., & Treisman, A. M. (2003). Representation of statistical properties. *Vision research, 43*(4), 393-404.

Chubak, J., Boudreau, D. M., Fishman, P. A., & Elmore, J. G. (2010). Cost of breast-related care in the year following false positive screening mammograms. *Medical care, 48*(9), 815.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in cognitive sciences, 4*(5), 170-178.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive psychology, 36*(1), 28-71.

Ciatto, S., Ambrogetti, D., Bonardi, R., Catarzi, S., Risso, G., Rosselli Del Turco, M., & Mantellini, P. (2005). Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening programme. *Journal of medical screening, 12*(2), 103-106.

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage, 153*, 346-358.

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature neuroscience, 17*(3), 455-462.

Cohen, M. A., Alvarez, G. A., & Nakayama, K. (2011). Natural-scene perception requires attention. *Psychological science, 22*(9), 1165-1172.

Cole, E. B., Pisano, E. D., Zeng, D., Muller, K., Aylward, S. R., Park, S., . . . Freimanis, R. (2005). The Effects of Gray Scale Image Processing on Digital Mammography Interpretation Performance. *Academic Radiology, 12*(5), 585-595. doi:https://doi.org/10.1016/j.acra.2005.01.017

Coleman, C. (2017). *Early detection and screening for breast cancer.* Paper presented at the Seminars in oncology nursing.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and cognitive processes, 13*(1), 21-58.

D'Orsi, C., Bassett, L., & Feig, S. (2018). Breast imaging reporting and data system (BI-RADS). *Breast imaging atlas, 4th edn. American College of Radiology, Reston*.

Darrien, J. H., Herd, K., Starling, L. J., Rosenberg, J. R., & Morrison, J. D. (2001). An analysis of the dependence of saccadic latency on target position and target characteristics in human subjects. *BMC neuroscience, 2*. doi:10.1186/1471-2202-2-13

De Gonzalez, A. B. (2011). Estimates of the potential risk of radiation-related cancer from screening in the UK. In (Vol. 18, pp. 163-164): SAGE Publications Sage UK: London, England.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2010). Key visual features for rapid categorization of animals in natural scenes. *Frontiers in psychology, 1*, 21.

Delorme, A., Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research, 19*(2), 103-113.

Deubel, H. (2008). The time course of presaccadic attention shifts. *Psychological research, 72*(6), 630-640.

Devillez, H., Mollison, M. V., Hagen, S., Tanaka, J. W., Scott, L. S., & Curran, T. (2019). Color and spatial frequency differentially impact early stages of perceptual expertise training. *Neuropsychologia, 122*, 62-75.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology, 5*, 781.

DiGirolamo, G. J., DiDominica, M., Qadri, M. A., Kellman, P. J., Krasne, S., Massey, C., & Rosen, M. P. (2023). Multiple expressions of "expert" abnormality gist in novices following perceptual learning. *Cognitive research: principles and implications, 8*(1), 1-14.

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The occipital place area is causally and selectively involved in scene perception. *Journal of Neuroscience, 33*(4), 1331-1336.

Dillon, M. R., Persichetti, A. S., Spelke, E. S., & Dilks, D. D. (2018). Places in the brain: bridging layout and object geometry in scene-selective cortex. *Cerebral cortex, 28*(7), 2365-2374.

Dinnes, J., Moss, S., Melia, J., Blanks, R., Song, F., & Kleijnen, J. (2001). Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. *The Breast, 10*(6), 455-463.

Dobres, J., & Seitz, A. R. (2010). Perceptual learning of oriented gratings as revealed by classification images. *Journal of Vision, 10*(13), 8-8.

Downing, P. E., Chan, A.-Y., Peelen, M., Dodds, C., & Kanwisher, N. (2006). Domain specificity in visual cortex. *Cerebral cortex, 16*(10), 1453-1461.

Duncan, J., Ward, R., & Shapiro, K. (1994). Direct measurement of attentional dwell time in human vision. *Nature, 369*(6478), 313-315.

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia, 49*(5), 1322-1331.

Emery, K. J., & Webster, M. A. (2019). Individual differences and their implications for color perception. *Current opinion in behavioral sciences, 30*, 28-33.

Epstein, R. A., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: recognition, navigation, or encoding? *Neuron, 23*(1), 115-125.

Epstein, R. A., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598-601.

Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. *PloS one, 8*(5).

Evans, K. K., & Chong, S. C. (2012). Distributed attention and its implication for visual perception. *From Perception to Consciousness: Searching with Anne Treisman*, 288-296.

Evans, K. K., Cohen, M. A., Tambouret, R., Horowitz, T., Kreindel, E., & Wolfe, J. M. (2011). Does visual expertise improve visual recognition memory? *Attention, Perception, & Psychophysics, 73*(1), 30-35.

Evans, K. K., Culpan, A. M., & Wolfe, J. M. (2019). Detecting the "GIST" of breast cancer in mammograms three years before localized signs of cancer are visible. *British Journal of Radiology, 92*. doi:10.1259/bjr.20190136

Evans, K. K., Georgian-Smith, D., Tambouret, R., Birdwell, R. L., & Wolfe, J. M. (2013). The gist of the abnormal: Above-chance medical decision making in the blink of an eye. *Psychonomic Bulletin and Review, 20*, 1170-1175. doi:10.3758/s13423-013-0459-3

Evans, K. K., Haygood, T. M., Cooper, J., Culpan, A. M., & Wolfe, J. M. (2016). A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. *Proceedings of the National Academy of Sciences of the United States of America, 113*, 10292-10297. doi:10.1073/pnas.1606187113

Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). When categories collide: Accumulation of information about multiple categories in rapid scene perception. *J Psychological Science, 22*(6), 739-746.

Evans, K. K., & Treisman, A. M. (2005). Perception of objects in natural scenes: is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance, 31*(6), 1476.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of cognitive neuroscience, 13*(2), 171-180.

Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of cognitive neuroscience, 19*(9), 1488-1497.

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. *International Journal of Cancer*.

Finnigan, S., O'Connell, R. G., Cummins, T. D., Broughton, M., & Robertson, I. H. (2011). ERP measures indicate both attention and working memory encoding decrements in aging. *Psychophysiology, 48*(5), 601-611.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological science, 12*(6), 499-504.

Fiser, J., & Aslin, R. N. (2002a). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 458.

Fiser, J., & Aslin, R. N. (2002b). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 99*(24), 15822-15826.

Fitzmaurice, C., Allen, C., Barber, R. M., Barregard, L., Bhutta, Z. A., Brenner, H., . . . Dandona, L. (2017). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. *JAMA oncology, 3*(4), 524-548.

Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology, 45*(1), 152-170.

Frank, S. M., Reavis, E. A., Greenlee, M. W., & Tse, P. U. (2016). Pretraining cortical thickness predicts subsequent perceptual learning rate in a visual search task. *Cerebral cortex, 26*(3), 1211-1220.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature neuroscience, 14*(9), 1195.

Gandomkar, Z., Siviengphanom, S., Ekpo, E. U., Suleiman, M. a., Li, T., Xu, D., . . . Brennan, P. C. (2021). Global processing provides malignancy evidence complementary to the information captured by humans or machines following detailed mammogram inspection. *Scientific reports, 11*(1), 1-12.

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research, 16*(2), 123-144.

Gezeck, S., Fischer, B., & Timmer, J. (1997). Saccadic reaction times: a statistical analysis of multimodal distributions. *Vision research, 37*(15), 2119-2131.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience, 14*(5), 350-363.

Gonzalez, C. M. G., Clark, V. P., Fan, S., Luck, S. J., & Hillyard, S. A. (1994). Sources of attention-sensitive visual event-related potentials. *Brain topography, 7*(1), 41-51.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., . . . Parkkonen, L. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience*, 267.

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive psychology, 58*, 137-176. doi:10.1016/j.cogpsych.2008.06.001.Recognition

Grill-Spector, K. (2003). The neural basis of object perception. *Current opinion in neurobiology, 13*(2), 159-166.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision research, 41*(10-11), 1409-1422.

Groen, I. I. A., Ghebreab, S., Lamme, V. A., & Scholte, H. S. (2016). The time course of natural scene perception with reduced attention. *Journal of neurophysiology, 115*(2), 931-946.

Groen, I. I. A., Ghebreab, S., Prins, H., Lamme, V. A. F., & Steven Scholte, H. (2013). From image statistics to scene gist: Evoked neural activity reveals transition from low-level natural image structure to scene category. *Journal of Neuroscience, 33*, 18814-18824. doi:10.1523/JNEUROSCI.3128-13.2013

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1714), 20160102.

Gross, C. G., Rocha-Miranda, C. d., & Bender, D. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of neurophysiology, 35*(1), 96-111.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751-R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*(3), 718.

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological science, 17*(7), 572-576.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Harel, A., Groen, I. I. A., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The temporal dynamics of scene processing: A multifaceted EEG investigation. *eNeuro, 3*, 1-18. doi:10.1523/ENEURO.0139-16.2016

Harel, A., Kravitz, D. J., & Baker, C. I. (2013). Deconstructing visual scenes in cortex: gradients of object and spatial layout information. *Cerebral cortex, 23*(4), 947-957.

Harley, E. M., Pope, W. B., Villablanca, J. P., Mumford, J., Suh, R., Mazziotta, J. C., . . . Engel, S. A. (2009). Engagement of fusiform cortex and disengagement of lateral occipital cortex in the acquisition of radiological expertise. *Cerebral cortex, 19*, 2746-2754. doi:10.1093/cercor/bhp051

Hegdé, J. (2020). Deep learning can be used to train naïve, nonprofessional observers to detect diagnostic visual patterns of certain cancers in mammograms: a proof-of-principle study. *Journal of Medical Imaging, 7*(2), 022410.

Henderson, J. M., Larson, C. L., & Zhu, D. C. (2008). Full scenes produce more activation than close-up scenes and scene-diagnostic objects in parahippocampal and retrosplenial cortex: an fMRI study. *Brain and cognition, 66*(1), 40-49.

Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid invariant encoding of scene layout in human OPA. *Neuron, 103*(1), 161-171. e163.

Hillyard, S. A., & Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Sciences, 95*(3), 781-787.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36*(5), 791-804.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology, 195*(1), 215-243.

Jackson-Nielsen, M., Cohen, M. A., & Pitts, M. A. (2017). Perception of ensemble statistics requires attention. *Consciousness and cognition, 48*, 149-160.

Jalalian, A., Mashohor, S. B., Mahmud, H. R., Saripan, M. I. B., Ramli, A. R. B., & Karasfi, B. (2013). Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clinical imaging, 37*(3), 420-426.

JASP-Team. (2020). JASP (Version 0.14.1). Retrieved from https://jasp-stats.org/

Jatoi, I., Zhu, K., Shah, M., & Lawrence, W. (2006). Psychological distress in US women who have experienced false-positive mammograms. *Breast cancer research and treatment, 100*(2), 191-200.

Jeffreys, H. (1998). *The theory of probability*: OUP Oxford.

Jemel, B., George, N., Olivares, E., Fiori, N., & Renault, B. (1999). Event-related potentials to structural familiar face incongruity processing. *Psychophysiology, 36*(4), 437-452.

Jenifer, S., Parasuraman, S., & Kadirvelu, A. (2016). Contrast enhancement and brightness preserving of digital mammograms using fuzzy clipped contrast-limited adaptive histogram equalization algorithm. *Applied Soft Computing, 42*, 167-177.

Johannes, S., Münte, T., Heinze, H., & Mangun, G. (1995). Luminance and spatial attention effects on early visual processing. *Cognitive Brain Research, 2*(3), 189-205.

Johnson, S. P. (2011). Development of visual perception. *Wiley interdisciplinary reviews: Cognitive science, 2*(5), 515-528.

Jones, T., Hadley, H., Cataldo, A. M., Arnold, E., Curran, T., Tanaka, J. W., & Scott, L. S. (2020). Neural and behavioral effects of subordinate-level training of novel objects across manipulations of color and spatial frequency. *European Journal of Neuroscience, 52*(11), 4468-4479.

Joubert, O. R., Rousselet, G. A., Fabre-Thorpe, M., & Fize, D. (2009). Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. *Journal of Vision, 9*(1), 1-16.

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision research, 47*(26), 3286-3297.

Kanske, P., Plitschka, J., & Kotz, S. A. (2011). Attentional orienting towards emotion: P2 and N400 ERP effects. *Neuropsychologia, 49*(11), 3121-3129.

Kauffmann, L., Ramanoël, S., Guyader, N., Chauvin, A., & Peyrin, C. (2015). Spatial frequency processing in scene-selective cortical regions. *NeuroImage, 112*, 86-95.

Kauffmann, L., Ramanoël, S., & Peyrin, C. (2014). The neural bases of spatial frequency processing during scene perception. *Frontiers in Integrative Neuroscience, 8*, 1-14. doi:10.3389/fnint.2014.00037

Kihara, K., & Takeda, Y. (2010). Time course of the integration of spatial frequency-based information in natural scenes. *Vision research, 50*(21), 2158-2162.

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: is it long-term and implicit? *Neuroscience Letters, 461*(2), 145-149.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3?

Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *Journal of Neuroscience, 31*(20), 7322-7333.

Kundel, H. L., & Nodine, C. F. (1975). Interpreting chest radiographs without visual search. *Radiology, 116*(3), 527-532.

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology-Radiological Society of North America, 242*(2), 396-402.

Kundel, H. L., Nodine, C. F., Krupinski, E. A., & Mello-Thoms, C. (2008). Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. *Academic Radiology, 15*(7), 881-886.

Kurek, J., Świderski, B., Osowski, S., Kruk, M., & Barhoumi, W. (2018). Deep learning versus classical neural approach to mammogram recognition. *Bulletin of the Polish Academy of Sciences, Technical Sciences, 66*(6).

Kutas, M., & Hillyard, S. A. (1982). The lateral distribution of event-related potentials during sentence processing. *Neuropsychologia, 20*(5), 579-590.

Kuzmiak, C. M., Cole, E., Zeng, D., Kim, E., Koomen, M., Lee, Y., . . . Pisano, E. D. (2010). Comparison of image acquisition and radiologist interpretation times in a diagnostic mammography center. *Academic Radiology, 17*(9), 1168-1174.

Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision, 9*(10), 6-6.

Levenson, R. M., Krupinski, E. A., Navarro, V. M., & Wasserman, E. A. (2015). Pigeons (Columba livia) as trainable observers of pathology and radiology breast cancer images. *PloS one, 10*(11), e0141357.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2003). Natural scene categorization in the near absence of attention: Further explorations. *Journal of Vision, 3*, 331-331. doi:10.1167/3.9.331

Li, Q., Joo, S. J., Yeatman, J. D., & Reinecke, K. (2020). controlling for participants’ Viewing Distance in Large-Scale, psychophysical online experiments Using a Virtual chinrest. *Scientific reports, 10*(1), 1-11.

Li, W., Piëch, V., & Gilbert, C. D. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nature neuroscience, 7*(6), 651-657.

Lloyd, R., Hodgson, M. E., & Stokes, A. (2002). Visual categorization with aerial photographs. *Annals of the Association of American Geographers, 92*(2), 241-266.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5*(5), 552-563.

Loschky, L. C., Boucart, M., Szaffarczyk, S., Beugnet, C., Johnson, A., & Tang, J. L. (2015). The contributions of central and peripheral vision to scene gist recognition with a 180° visual field. *Journal of Vision, 15*(12), 570-570.

Loschky, L. C., Ringer, R. V., Ellis, K., & Hansen, B. C. (2015). Comparing rapid scene categorization of aerial and terrestrial views: A new perspective on scene gist. *Journal of Vision, 15*(6), 11-11.

Lowe, M. X., Rajsic, J., Ferber, S., & Walther, D. B. (2018). Discriminating scene categories from brain activity within 100 milliseconds. *cortex, 106*, 275-287.

Lowe, M. X., Rajsic, J., Gallivan, J. P., Ferber, S., & Cant, J. S. (2017). Neural representation of geometry and surface properties in object and scene perception. *NeuroImage, 157*, 586-597. doi:10.1016/j.neuroimage.2017.06.043

MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature neuroscience, 14*(10), 1323.

Maguire, E. A. (2001). The retrosplenial contribution to human navigation: a review of lesion and neuroimaging findings. *Scandinavian journal of psychology, 42*(3), 225-238.

Maguire, J. F., & Howe, P. D. (2016). Failure to detect meaning in RSVP at 27 ms per picture. *Attention, Perception, & Psychophysics, 78*(5), 1405-1413.

Majid, A. S., de Paredes, E. S., Doherty, R. D., Sharma, N. R., & Salvador, X. (2003). Missed breast carcinoma: pitfalls and pearls. *Radiographics, 23*(4), 881-895.

Makki, J. (2015). Diversity of breast carcinoma: histological subtypes and clinical relevance. *Clinical medicine insights: Pathology, 8*, CPath. S31563.

Mandelblatt, J. S., Cronin, K. A., Bailey, S., Berry, D. A., De Koning, H. J., Draisma, G., . . . Plevritis, S. K. (2009). Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. *Annals of internal medicine, 151*(10), 738-747.

Mangun, G. R. (1995). Neural mechanisms of visual selective attention. *Psychophysiology, 32*(1), 4-18.

Maniglia, M., & Seitz, A. R. (2018). Towards a whole brain model of Perceptual Learning. *Current opinion in behavioral sciences, 20*, 47-55.

Marmot, M. G., Altman, D., Cameron, D., Dewar, J., Thompson, S., & Wilcox, M. (2013). The benefits and harms of breast cancer screening: an independent review. *British journal of cancer, 108*(11), 2205-2240.

Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: Metric and categorical effects on ensemble perception of hue. *JOSA A, 31*(4), A93-A102.

Maurer, D. (2013). chapter 1: Infant Visual Perception: Methods of Study. *Infant perception: From sensation to cognition: Basic visual processes, 1*, 1.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The american statistician, 32*(1), 12-16.

Mercer, C. E., Hogg, P., Szczepura, K., & Denton, E. R. (2013). Practitioner compression force variation in mammography: A 6-year study. *Radiography, 19*(3), 200-206. doi:https://doi.org/10.1016/j.radi.2013.06.001

Näätänen, R., & Picton, T. (1986). N2 and automatic versus controlled processes. *Electroencephalography & Clinical Neurophysiology Supplement, 38*, 169-186.

NICE, N. I. f. H. a. C. E.-. (2017, May 2022). Breast Screening. Retrieved from https://cks.nice.org.uk/topics/breast-screening/

Nodine, C., & Kundel, H. (1987). The cognitive side of visual search in radiology. In *Eye movements from physiology to cognition* (pp. 573-582): Elsevier.

Nodine, C. F., Mello-Thoms, C., Kundel, H. L., & Weinstein, S. P. (2002). Time course of perception and decision making during mammographic interpretation. *American Journal of Roentgenology, 179*(4), 917-923.

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: fully automated statistical thresholding for EEG artifact rejection. *Journal of neuroscience methods, 192*(1), 152-162.

Oliva, A. (2005). Gist of the scene. In *Neurobiology of attention* (pp. 251-256): Elsevier.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive psychology, 34*(1), 72-107.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision, 42*(3), 145-175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research, 155*, 23-36.

Oruç, I., Krigolson, O., Dalrymple, K., Nagamatsu, L. S., Handy, T. C., & Barton, J. J. (2011). Bootstrap analysis of the single subject with event related potentials. *Cognitive neuropsychology, 28*(5), 322-337.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language, 31*(6), 785-806.

Ouyang, W., Zeng, X., Wang, X., Qiu, S., Luo, P., Tian, Y., . . . Li, H. (2016). DeepID-Net: Object detection with deformable part based convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(7), 1320-1334.

Owsley, C., Sekuler, R., & Siemsen, D. (1983). Contrast sensitivity throughout adulthood. *Vision research, 23*(7), 689-699.

Palmeri, T. J., Wong, A. C., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *Trends in cognitive sciences, 8*(8), 378-386.

Park, S., Konkle, T., & Oliva, A. (2015). Parametric coding of the size and clutter of natural scenes in the human brain. *Cerebral cortex, 25*(7), 1792-1805.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature neuroscience, 4*(7), 739-744.

Patel, S. H., & Azzam, P. N. (2005). Characterization of N200 and P300: selected studies of the event-related potential. *International journal of medical sciences, 2*(4), 147.

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature, 460*, 94-97. doi:10.1038/nature08103

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision, 10*, 437-442.

Perfetto, S., Wilder, J., & Walther, D. B. (2020). Effects of spatial frequency filtering choices on the perception of filtered images. *Vision, 4*(2), 29.

Perrett, D. I., & Oram, M. W. (1993). Neurophysiology of shape processing. *Image and Vision Computing, 11*(6), 317-333.

Peterzell, D., Werner, J. S., & Kaplan, P. S. (1991). Individual differences in contrast sensitivity functions of human adults and infants: A brief review. *The changing visual system: Maturation and aging in the central nervous system*, 391-396.

Peterzell, D. H., Werner, J. S., & Kaplan, P. S. (1995). Individual differences in contrast sensitivity functions: Longitudinal study of 4-, 6-and 8-month-old human infants. *Vision research, 35*(7), 961-979.

Pisano, E. D., Gatsonis, C., Hendrick, E., Yaffe, M., Baum, J. K., Acharyya, S., . . . D'Orsi, C. (2005). Diagnostic performance of digital versus film mammography for breast-cancer screening. *New England Journal of Medicine, 353*(17), 1773-1783.

Polich, J. (1997). On the relationship between EEG and P300: individual differences, aging, and ultradian rhythms. *International journal of psychophysiology, 26*(1-3), 299-317.

Polich, J. (2003). Theoretical overview of P3a and P3b. *Detection of change*, 83-98.

Poltoratski, S., & Xu, Y. (2013). The association of color memory and the enumeration of multiple spatially overlapping sets. *Journal of Vision, 13*(8), 6-6.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision, 40*(1), 49-70.

Potter, M. C. (1975). Meaning in visual search. *Science, 187*, 965-966.

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics, 76*(2), 270-279.

Pringle, H. L., Kramer, A. F., & Irwin, D. E. (2004). *Individual differences in the visual representation of scenes*: MIT Press.

Pruthi, S., Gostout, B. S., & Lindor, N. M. (2010). *Identification and management of women with BRCA mutations or hereditary predisposition for breast and ovarian cancer.* Paper presented at the Mayo Clinic Proceedings.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature, 435*(7045), 1102-1107.

Raat, E., Farr, I., Wolfe, J. M., & Evans, K. K. (2021). Comparable prediction of breast cancer risk from a glimpse or a first impression of a mammogram. *Cognitive research: principles and implications, 6*(1), 1-14.

Rajimehr, R., Devaney, K. J., Bilenko, N. Y., Young, J. C., & Tootell, R. B. (2011). The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys. *PLoS biology, 9*(4).

Rawashdeh, M. A., Lee, W. B., Bourne, R. M., Ryan, E. A., Pietrzyk, M. W., Reed, W. M., . . . Brennan, P. C. (2013). Markers of good performance in mammography depend on number of annual readings. *Radiology, 269*(1), 61-67.

Reed, W. M., Lee, W. B., Cawson, J. N., & Brennan, P. C. (2010). Malignancy detection in digital mammograms: important reader characteristics and required case numbers. *Academic Radiology, 17*(11), 1409-1413.

Reeder, R. R., Stein, T., & Peelen, M. V. (2016). Perceptual expertise improves category detection in natural scenes. *Psychonomic bulletin & review, 23*(1), 172-179.

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological bulletin, 140*(5), 1281.

Richler, J. J., Mack, M. L., Gauthier, I., & Palmeri, T. J. (2009). Holistic processing of faces happens at a glance. *Vision research, 49*(23), 2856-2861.

Rivolta, D., Palermo, R., Schmalzl, L., & Williams, M. A. (2012). An early category-specific neural response for the perception of both places and faces. *Cognitive neuroscience, 3*(1), 45-51.

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision, 11*(12), 18-18.

Rotman, T., Lavie, L., & Banai, K. (2020). Rapid perceptual learning: A potential source of individual differences in speech perception under adverse conditions? *Trends in Hearing, 24*, 2331216520930541.

Rourke, L., Cruikshank, L. C., Shapke, L., & Singhal, A. (2016). A neural marker of medical visual expertise: implications for training. *Advances in Health Sciences Education, 21*(5), 953-966.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature neuroscience, 5*, 629-630. doi:10.1038/nn866

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004). Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision research, 44*(9), 877-894.

Sandin, B., Chorot, P., Valiente, R. M., Lostao, L., & Santed, M. A. (2002). Adverse psychological effects in women attending a second-stage breast cancer screening. *Journal of psychosomatic research, 52*(5), 303-309.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological science, 5*(4), 195-200.

Scutt, D., Lancaster, G. A., & Manning, J. T. (2006). Breast asymmetry and predisposition to breast cancer. *Breast cancer research, 8*(2), R14.

Semizer, Y., Michel, M., Evans, K., & Wolfe, J. (2018). Texture as a diagnostic signal in mammograms.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, 104*(15), 6424-6429.

Sharan, L., Rosenholtz, R., & Adelson, E. (2009). Material perception: What can you see in a brief glance? *Journal of Vision, 9*(8), 784-784.

Sharma, P. (2016). Biology and management of patients with triple-negative breast cancer. *The oncologist, 21*(9), 1050-1062.

Sheridan, H., & Reingold, E. M. (2017). The holistic processing account of visual expertise in medical image perception: A review. *Frontiers in psychology, 8*, 1620.

Shin, K. S., Kang, D.-H., Choi, J.-S., Kim, Y. Y., & Kwon, J. S. (2008). Neuropsychological correlates of N400 anomalies in patients with schizophrenia: A preliminary report. *Neuroscience Letters, 448*(2), 226-230.

Shinn-Cunningham, B., Varghese, L., Wang, L., & Bharadwaj, H. (2017). Individual differences in temporal perception and their implications for everyday listening. *The Frequency-Following Response*, 159-192.

Simoncelli, E. P. (2003). Vision and the statistics of the visual environment. *Current opinion in neurobiology, 13*(2), 144-149.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Skottun, B. C. (2015). On the use of spatial frequency to isolate contributions from the magnocellular and parvocellular systems and the dorsal and ventral cortical streams. *Neuroscience & Biobehavioral Reviews, 56*, 266-275.

Smith, A. T., Singh, K. D., Williams, A., & Greenlee, M. W. (2001). Estimating receptive field size from fMRI data in human striate and extrastriate visual cortex. *Cerebral cortex, 11*(12), 1182-1190.

Sobel, K. V., Gerrie, M. P., Poole, B. J., & Kane, M. J. (2007). Individual differences in working memory capacity and visual search: The roles of top-down and bottom-up processing. *Psychonomic bulletin & review, 14*(5), 840-845.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature, 400*(6747), 869-873.

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4–5-year-old children. *Developmental science, 18*(4), 556-568.

Swensson, R. G. (1980). A two-stage detection model applied to skilled visual search by radiologists. *Perception & psychophysics, 27*(1), 11-16.

Tabár, L., Dean, P. B., Chen, T. H.-H., Yen, A. M.-F., Chiu, S. Y.-H., Tot, T., . . . Duffy, S. W. (2014). The impact of mammography screening on the diagnosis and management of early-phase breast cancer. *Breast Cancer*, 31-78.

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). *Deepface: Closing the gap to human-level performance in face verification.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual review of neuroscience, 19*(1), 109-139.

Theeuwes, J., Godijn, R., & Pratt, J. (2004). A new estimation of the duration of attentional dwell time. *Psychonomic bulletin & review, 11*, 60-64.

Thorpe, S. J., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520-522. doi:10.1038/381520a0

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & BuÈlthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience, 14*(5), 869-876.

Treisman, A. M. (2006). How the deployment of attention determines what we see. *Visual cognition, 14*(4-8), 411-443.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology, 12*(1), 97-136.

Treisman, A. M., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *1982, 14*, 107-141.

Treviño, M., Turkbey, B., Wood, B. J., Pinto, P. A., Czarniecki, M., Choyke, P. L., & Horowitz, T. S. (2020). Rapid perceptual processing in two-and three-dimensional prostate images. *Journal of Medical Imaging, 7*(2), 022406.

Tripathy, S., & Swarnkar, T. (2020). Unified preprocessing and enhancement technique for mammogram images. *Procedia Computer Science, 167*, 285-292.

Turk-Browne, N. B. (2012). Statistical learning and its consequences. In *The influence of attention, learning, and motivation on visual search* (pp. 117-146): Springer.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*(4), 552.

Twomey, D. M., Murphy, P. R., Kelly, S. P., & O'Connell, R. G. (2015). The classic P300 encodes a build-to-threshold decision variable. *European Journal of Neuroscience, 42*, 1636-1643. doi:10.1111/ejn.12936

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature neuroscience, 5*(7), 682-687.

Utochkin, I. S., & Vostrikov, K. O. (2017). The numerosity and mean size of multiple objects are perceived independently and in parallel. *PloS one, 12*(9). doi:10.1371/journal.pone.0185452

Vachon, C. M., Brandt, K. R., Ghosh, K., Scott, C. G., Maloney, S. D., Carston, M. J., . . . Sellers, T. A. (2007). Mammographic breast density as a general marker of breast cancer risk. *Cancer Epidemiology and Prevention Biomarkers, 16*(1), 43-49.

Van den Biggelaar, F., Nelemans, P., & Flobbe, K. (2008). Performance of radiographers in mammogram interpretation: a systematic review. *The Breast, 17*(1), 85-90.

Vann, S. D., Aggleton, J. P., & Maguire, E. A. (2009). What does the retrosplenial cortex do? *Nature Reviews Neuroscience, 10*(11), 792-802.

VanRullen, R. (2007). The power of the feed-forward sweep. *Advances in Cognitive Psychology, 3*(1-2), 167.

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of cognitive neuroscience, 13*, 454-461. doi:10.1162/08989290152001880

Vogel, E. K., & Luck, S. J. (2000). The visual N1 component as an index of a discrimination process. *Psychophysiology, 37*(2), 190-203.

Vogels, R., & Orban, G. A. (1996). Coding of stimulus invariances by inferior temporal neurons. In *Progress in brain research* (Vol. 112, pp. 195-211): Elsevier.

Voss, M. W., Kramer, A. F., Basak, C., Prakash, R. S., & Roberts, B. (2010). Are expert athletes 'expert'in the cognitive laboratory? A meta-analytic review of cognition and sport expertise. *Applied Cognitive Psychology, 24*(6), 812-826.

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience, 2018*.

Waller, D. (2000). Individual differences in spatial learning from computer-simulated environments. *Journal of Experimental Psychology: Applied, 6*(4), 307.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience, 29*, 10573-10581. doi:10.1523/JNEUROSCI.0559-09.2009

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. J. P. o. t. N. A. o. S. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *108*(23), 9661-9666.

Wang, M.-J. J., Lin, S.-C., & Drury, C. G. (1997). Training for strategy in visual search. *International Journal of Industrial Ergonomics, 20*(2), 101-108.

Wang, Y., Jin, C., Yin, Z., Wang, H., Ji, M., Dong, M., & Liang, J. (2021). Visual experience modulates whole-brain connectivity dynamics: A resting-state fMRI study using the model of radiologists. *Human brain mapping, 42*(14), 4538-4554.

Warren, L. M., Dance, D. R., & Young, K. C. (2016). Radiation risk of breast screening in England with digital mammography. *The British journal of radiology, 89*(1067), 20150897.

Warren, R., & Duffy, W. (1995). Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. *The British journal of radiology, 68*(813), 958-962.

White, D., & Burton, A. M. (2022). Individual differences and the multidimensional nature of face perception. *Nature Reviews Psychology, 1*(5), 287-300.

Williams, D. W., & Sekuler, R. (1984). Coherent global motion percepts from stochastic local motions. *ACM SIGGRAPH Computer Graphics, 18*(1), 24-24.

Wivell, G., Denton, E. R. E., Eve, C. B., Inglis, J. C., & Harvey, I. (2003). Can radiographers read screening mammograms? *Clinical Radiology, 58*, 63-67.

Wolfe, J. M. (1983). Influence of spatial frequency, luminance, and duration on binocular rivalry and abnormal fusion of briefly presented dichoptic stimuli. *Perception, 12*(4), 447-456.

Wolfe, J. M., Vo, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and non-selective pathways. *Trends in Cognitive Science, 15*, 77-84. doi:10.1016/j.tics.2010.12.001.Visual

Wong, A. C.-N., Palmeri, T. J., & Gauthier, I. (2009). Conditions for facelike expertise with objects: Becoming a Ziggerin expert—but which type? *Psychological science, 20*(9), 1108-1117.

Wu, C. C., D'Ardenne, N. M., Nishikawa, R. M., & Wolfe, J. M. (2019). Gist processing in digital breast tomosynthesis. *Journal of Medical Imaging, 7*(2), 022403.

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., . . . Kim, E. (2019). Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging, 39*(4), 1184-1194.

Wurster, S. W., Sitek, A., Chen, J., Evans, K. K., Kim, G., & Wolfe, J. M. (2019). Human Gist Processing Augments Deep Learning Breast Cancer Risk Assessment. *arXiv preprint arXiv:1912.05470*.

Xu, B., Rourke, L., Robinson, J. K., & Tanaka, J. W. (2016). Training melanoma detection in photographs using the perceptual expertise training approach. *Applied Cognitive Psychology, 30*(5), 750-756.

Yang, J., Yan, F.-F., Chen, L., Xi, J., Fan, S., Zhang, P., . . . Huang, C.-B. (2020). General learning ability in perceptual learning. *Proceedings of the National Academy of Sciences, 117*(32), 19092-19100.

Zhang, T., Dong, M., Wang, H., Jia, R., Li, F., Ni, X., & Jin, C. (2022). Visual expertise modulates baseline brain activity: a preliminary resting-state fMRI study using expertise model of radiologists. *BMC neuroscience, 23*(1), 1-11.

Zheng, B., Sumkin, J. H., Zuley, M. L., Wang, X., Klym, A. H., & Gur, D. (2012). Bilateral mammographic density asymmetry and breast cancer risk: a preliminary assessment. *European journal of radiology, 81*(11), 3222-3228.