

1 Convolutional neural networks can decode eye movement data: A black box approach to
2 predicting task from eye movements

³ Zachary J. Cole¹, Karl M. Kuntzman¹, Michael D. Dodd¹, & Matthew R. Johnson¹

⁴ ¹ University of Nebraska-Lincoln

5 Author Note

The data used for the exploratory and confirmatory analyses in the present manuscript are derived from experiments funded by NIH/NEI Grant 1R01EY022974 to MDD. Additionally, work done to develop the analysis approach was supported by NSF/EPSCoR grant #1632849 and NIH grant GM130461 awarded to MRJ and colleagues.

¹⁰ Correspondence concerning this article should be addressed to Zachary J. Cole, 238
¹¹ Burnett Hall, Lincoln, NE 68588-0308. E-mail: z@neurophysicole.com

12

Abstract

13 Previous attempts to classify task from eye movement data relied on model architectures
14 designed to emulate theoretically defined cognitive processes, and/or data that has been
15 processed into aggregate (e.g., fixations, saccades) or statistical (e.g., fixation density)
16 features. *Black box* convolutional neural networks (CNNs) are capable of identifying relevant
17 features in raw and minimally processed data and images, but difficulty interpreting the
18 mechanisms underlying these model architectures have contributed to challenges in
19 generalizing lab-trained CNNs to applied contexts. In the current study, a CNN classifier
20 was used to classify task from two eye movement datasets (Exploratory and Confirmatory)
21 in which participants searched, memorized, or rated indoor and outdoor scene images. The
22 Exploratory dataset was used to tune the hyperparameters of the model, and the resulting
23 model architecture was re-trained, validated, and tested on the Confirmatory dataset. The
24 data were formatted into raw timeline data (i.e., x-coordinate, y-coordinate, pupil size) and
25 minimally processed images. To further understand the relative informational value of the
26 raw components of the eye movement data, the timeline and image datasets were broken
27 down into subsets with one or more of the components of the data systematically removed.
28 Average classification accuracies were compared between datasets and subsets. Classification
29 of the timeline data consistently outperformed the image data. The Memorize condition was
30 most often confused with the Search and Rate conditions. Pupil size was the least uniquely
31 informative eye movement component when compared with the x- and y-coordinates. The
32 general pattern of results for the Exploratory dataset was replicated in the Confirmatory
33 dataset. Overall, the present study provides a practical and reliable black box solution to the
34 inverse Yarbus problem.

35 *Keywords:* deep learning, eye tracking, convolutional neural network, cognitive state,
36 endogenous attention

37 Word count: 7260

38 Convolutional neural networks can decode eye movement data: A black box approach to
39 predicting task from eye movements

40 **Background**

41 The association between eye movements and mental activity is a fundamental topic of
42 interest in attention research that has provided a foundation for developing a wide range of
43 human assistive technologies. Early work by Yarbus (1967) showed that eye movement
44 patterns appear to differ qualitatively depending on the task-at-hand (for a review of this
45 work, see Tatler, Wade, Kwan, Findlay, & Velichkovsky, 2010). A replication of this work by
46 DeAngelus and Pelz (2009) shows that the differences in eye movements between tasks can
47 be quantified, and appear to be somewhat generalizable. Technological advances and
48 improvements in computing power have allowed researchers to make inferences regarding the
49 mental state underlying eye movement data, also known as the “inverse Yarbus process”
50 (Haji-Abolhassani & Clark, 2014). Current state-of-the-art machine learning and neural
51 network algorithms are capable of identifying diagnostic patterns for the purpose of decoding
52 a variety of data types, but the inner workings of the resulting model solutions are difficult
53 or impossible to interpret. Algorithms that provide such solutions are referred to as *black box*
54 models. Dissections of black box models have been largely uninformative (Zhou, Bau, Oliva,
55 & Torralba, 2019), limiting the potential for researchers to apply the mechanisms underlying
56 successful classification of the data. Still, black box models provide a powerful solution for
57 technological applications such as human-computer interfaces (HCI; for a review, see
58 Lukander, Toivanen, & Puolamäki, 2017). While the internal operations of the model
59 solutions used for HCI applications do not necessarily need to be interpretable to serve their
60 purpose, Lukander et al. (2017) pointed out that the inability to interpret the mechanisms
61 underlying the function of black box solutions impedes the generalizability of these methods,
62 and increases the difficulty of expanding these findings to real life applications. To ground
63 these solutions, researchers guide decoding efforts by using eye movement data and/or
64 models with built-in theoretical assumptions. For instance, eye movement data is processed

65 into meaningful aggregate properties such as fixations or saccades, or statistical features such
66 as fixation density, and the models used to decode these data are structured based on the
67 current understanding of relevant cognitive or neurobiological processes (e.g., MacInnes,
68 Hunt, Clarke, & Dodd, 2018). Despite the proposed disadvantages of black box approaches
69 to classifying eye movement data, there is no clear evidence to support the notion that the
70 grounded solutions described above are actually more valid or definitive than a black box
71 solution.

72 The scope of theoretically informed solutions to decoding eye movement data are
73 limited to the extent of the current theoretical knowledge linking eye movements to cognitive
74 and neurobiological processes. As our theoretical understanding of these processes develops,
75 older theoretically informed models become outdated. Furthermore, these solutions are
76 susceptible to any inaccurate preconceptions that are built into the theory. Consider the case
77 of Greene, Liu, and Wolfe (2012), who were not able to classify the task from commonly
78 used aggregate eye movement features (i.e., number of fixations, mean fixation duration,
79 mean saccade amplitude, percent of image covered by fixations) using correlations, a linear
80 discriminant model, and a support vector machine (see Table 1). This led Greene and
81 colleagues to question the robustness of Yarbus's (1967) findings, inspiring a slew of
82 responses that successfully decoded the same dataset by aggregating the eye movements into
83 different feature sets or implementing different model architectures (see Table 1; i.e.,
84 Haji-Abolhassani & Clark, 2014; Borji & Itti, 2014; Kanan, Ray, Bseiso, Hsiao, & Cottrell,
85 2014). The subsequent re-analyses of these data support Yarbus (1967) and the notion that
86 mental state can be decoded from eye movement data using a variety of combinations of
87 data features and model architectures. Altogether, these re-analyses did not point to an
88 obvious global solution capable of clarifying future approaches to the inverse Yarbus problem
89 beyond what could be inferred from black box model solutions, but did provide a rigorous
90 test of a variety of methodological features which can be applied to theoretical or black box
91 approaches to the inverse Yarbus problem.

92 Eye movements can only delineate tasks to the extent that the cognitive processes
93 underlying the tasks can be differentiated (Król & Król, 2018). Every task is associated with
94 a unique set of cognitive processes (Coco & Keller, 2014; Król & Król, 2018), but in some
95 cases, the cognitive processes for different tasks may produce indistinguishable eye movement
96 patterns. To differentiate the cognitive processes underlying task-evoked eye movements,
97 some studies have chosen to classify tasks that rely on stimuli that prompt easily
98 distinguishable eye movements, such as reading text and searching pictures (e.g., Henderson,
99 Shinkareva, Wang, Luke, & Olejarczyk, 2013). The eye movements elicited by salient
100 stimulus features facilitate task classifications, however, because these eye movements are the
101 consequence of a feature (or features) inherent to the stimulus rather than the task, it is
102 unclear if these classifications are attributable to the stimulus or a complex mental state
103 (e.g., Henderson et al., 2013; Boisvert & Bruce, 2016). Additionally, the distinct nature of
104 exogenously elicited eye movements prompts decoding algorithms to prioritize these
105 bottom-up patterns in the data over higher-level top-down effects (Borji & Itti, 2014). This
106 means that these models are identifying the type of information that is being processed, but
107 are not necessarily reflecting the mental state of the individual observing the stimulus. Eye
108 movements that are the product of bottom-up processes have been reliably decoded, which is
109 relevant for some HCI applications, but does not fit the nature of the inverse Yarbus
110 problem, which is concerned with decoding high-level abstract mental operations that are
111 not dependent on particular stimuli.

112 Currently, an upper limit to how well cognitive task can be classified from eye
113 movement data has not been clearly established. Prior evidence has shown that the
114 task-at-hand is capable of producing distinguishable eye movement features such as the total
115 scan path length, total number of fixations, and the amount of time to the first saccade
116 (Castelhano, Mack, & Henderson, 2009; DeAngelus & Pelz, 2009). Decoding accuracies
117 within the context of determining task from eye movements typically range from chance
118 performance (between 14.29% and 33%) to 59.64% (see Table 1). In one case, Coco and

119 Keller (2014) categorized the same eye movement features used by Greene et al. (2012) with
120 respect to the relative contribution of latent visual or linguistic components of three tasks
121 (visual search, name the picture, name objects in the picture) with 84% accuracy. While this
122 manipulation is reminiscent of other experiments relying on the bottom-up influence of
123 words and pictures (e.g., Henderson et al., 2013; Boisvert & Bruce, 2016) the eye movements
124 in the Coco and Keller (2014) tasks can be attributed to the occurrence of top-down
125 attentional processes. A conceptually similar follow-up to this study classified tasks along
126 two spatial and semantic dimensions, resulting in 51% classification accuracy (chance = 25%;
127 Król & Król, 2018). A closer look at these results showed that the categories within the
128 semantic dimension were consistently misclassified, suggesting that this level of distinction
129 may require a richer dataset, or a more powerful decoding algorithm. Altogether, there is no
130 measurable index of relative top-down or bottom-up influence, but this body of literature
131 suggests that the relative influence of top-down and bottom-up attentional processes may
132 have a role in determining the decodability of the eye movement data.

133 As shown in Table 1, when eye movement data are prepared for classification, fixation
134 and saccade statistics are typically aggregated along spatial or temporal dimensions,
135 resulting in variables such as fixation density or saccade amplitude (Castelhano et al., 2009;
136 MacInnes et al., 2018; Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011). The
137 implementation of these statistical methods is meant to explicitly provide the decoding
138 algorithm with characteristics of the eye movement data that are representative of
139 theoretically relevant cognitive processes. For example, MacInnes et al. (2018) attempted to
140 provide an algorithm with data designed to be representative of inputs to the frontal eye
141 fields. In some instances, such as the case of Król and Król (2018), grounding the data using
142 theoretically driven aggregation methods may require sacrificing granularity in the dataset.
143 This means that aggregating the data has the potential to wash out certain fine-grained
144 distinctions that could otherwise be detected. Data structures of any kind can only be
145 decoded to the extent at which the data are capable of representing differences between

Table 1

Previous Attempts to Classify Cognitive Task Using Eye Movement Data

Study	Tasks	Features	Model Architecture	Accuracy (Chance)
Greene et al. (2012)	memorize, decade, people, wealth	number of fixations, mean fixation duration, mean saccade amplitude, percent of image covered by fixations, dwell times	linear discriminant, correlation, SVM	25.9% (25%)
Haji-Abolhassani & James (2014)	Greene et al. tasks	fixation clusters	Hidden Markov Models	59.64% (25%)
Kanan et al. (2014)	Greene et al. tasks	mean fixation durations, number of fixations	multi-fixation pattern analysis	37.9% (25%)
Borji & Itti (2014)	Greene et al. tasks	number of fixations, mean fixation duration, mean saccade amplitude, percent of image covered by fixations, first five fixations, fixation density	kNN, RUSBoost	34.34% (25%)
Borji & Itti (2014)	Yarbus tasks (i.e., view, wealth, age, prior activity, clothes, location, time away)	number of fixations, mean fixation duration, mean saccade amplitude, percent of image covered by fixations, first five fixations, fixation density	kNN, RUSBoost	24.21% (14.29%)
Coco & Keller (2014)	search, name picture, name object	Greene et al. features, latency of first fixation, first fixation duration, mean fixation duration, total gaze duration, initiation time, mean saliency at fixation, entropy of attentional landscape	MM, LASSO, SVM	84% (33%)
MacInnes et al. (2018)	view, memorize, search, rate	saccade latency, saccade duration, saccade amplitude, peak saccade velocity, absolute saccade angle, pupil size	augmented Naive Bayes Network	53.9% (25%)
Król & Król (2018)	people, indoors/outdoors, white/black, search	eccentricity, screen coverage	feed forward neural network	51.4% (25%)

¹⁴⁶ categories. Given that the cognitive processes underlying distinct tasks are often overlapping¹⁴⁷ (Coco & Keller, 2014), decreasing the granularity of the data may actually limit the potential

¹⁴⁸ of the algorithm to make fine-grained distinctions between diagnostic components underlying
¹⁴⁹ the target task and the other tasks.

¹⁵⁰ The current state of the literature does not provide any firm guidelines for determining
¹⁵¹ what eye movement features are most meaningful, or what model architectures are most
¹⁵² suited for determining mental state from eye movements. The examples provided in Table 1
¹⁵³ used a variety of eye movement features and model architectures, most of which were
¹⁵⁴ effective to some extent. A proper comparison of these outcomes is difficult because these
¹⁵⁵ datasets vary in levels of chance and data quality. Datasets with more tasks to be classified
¹⁵⁶ have lower levels of chance, lowering the threshold for successful classification. Additionally,
¹⁵⁷ datasets with a lower signal-to-noise ratio will have a lower achievable classification accuracy.
¹⁵⁸ For these reasons, outside of re-analyzing the same datasets, there is no consensus on how to
¹⁵⁹ establish direct comparisons of these model architectures. Given the inability to directly
¹⁶⁰ compare the relative effectiveness of the various theoretical approaches present in the
¹⁶¹ literature, the current study addressed the inverse Yarbus problem by allowing a black box
¹⁶² model to self-determine the most informative features from minimally processed eye
¹⁶³ movement data.

¹⁶⁴ The current study explored pragmatic solutions to the problem of classifying task from
¹⁶⁵ eye movement data by submitting unprocessed x-coordinate, y-coordinate, and pupil size
¹⁶⁶ data to a convolutional neural network (CNN) model. Instead of transforming the data into
¹⁶⁷ theoretically defined units, we allowed the network to learn meaningful patterns in the data
¹⁶⁸ on its own. CNNs have a natural propensity to develop low-level feature detectors similar to
¹⁶⁹ primary visual cortex (e.g., Seeliger et al., 2018); for this reason, they are commonly
¹⁷⁰ implemented for image classification. To test the possibility that the image data are better
¹⁷¹ suited to the CNN classifier, the data were also transformed into raw timeline and simple
¹⁷² image representations. To our knowledge, no study has attempted to address the inverse
¹⁷³ Yarbus problem using any combination of the following methods: (1) Non-aggregated data,

174 (2) image data format, and (3) a black-box CNN architecture. Given that CNN architectures
175 are capable of learning features represented in raw data formats, and are well-suited to
176 decoding multidimensional data that have a distinct spatial or temporal structure, we
177 expected that a non-theoretically-constrained CNN architecture could be capable of decoding
178 data at levels consistent with the current state of the art. Furthermore, despite evidence that
179 black box approaches to the inverse Yarbus problem can impede generalizability (Lukander
180 et al., 2017), we expected that when testing the approach on an entirely separate dataset,
181 providing the model with minimally processed data and the flexibility to identify the unique
182 features within each dataset would result in the replication of our initial findings.

183

Methods

184 **Participants**

185 Two separate datasets were used to develop and test the deep CNN architecture. The
186 two datasets were collected from two separate experiments, which we refer to as Exploratory
187 and Confirmatory. The participants for both datasets consisted of college students
188 (Exploratory $N = 124$; Confirmatory $N = 77$) from the University of Nebraska-Lincoln who
189 participated in exchange for class credit. Participants who took part in the Exploratory
190 experiment did not participate in the Confirmatory experiment. All procedures and
191 materials were approved by the University of Nebraska-Lincoln Institutional Review Board
192 prior to data collection.

193 **Materials and Procedures**

194 Each participant viewed a series of indoor and outdoor scene images while carrying out
195 a search, memorization, or rating task. For the search task, participants were instructed to
196 find a small “Z” or “N” embedded in the image. If the letter was found, the participants
197 were instructed to press a button, which terminated the trial. For the memorization task,
198 participants were instructed to memorize the image for a test that would take place when

199 the task was completed. For the rating task, participants were asked to think about how
200 they would rate the image on a scale from 1 (very unpleasant) to 7 (very pleasant). The
201 participants were prompted for their rating immediately after viewing the image. The same
202 materials were used in both experiments with a minor variation in the procedures. In the
203 Confirmatory experiment, participants were directed as to where search targets might appear
204 in the image (e.g., on flat surfaces). No such instructions were provided in the Exploratory
205 experiment. In actuality, none of the images in either experiment actually contained any
206 search targets.

207 In both experiments, trials were presented in one mixed block, and three separate task
208 blocks. For the mixed block, the trial types were randomly intermixed within the block. For
209 the three separate task blocks, each block was 35 trials consisting entirely of one of the three
210 conditions (Search, Memorize, Rate). Each trial was presented for 10 seconds. The pictures
211 were presented in color, with a size of 1024 x 768 pixels, subtending a visual angle of 23.84 x
212 17.99 degrees.

213 Datasets

214 Eye movements were recorded using an SR Research EyeLink II eye tracker with a
215 sampling rate of 1000Hz. On some of the search trials, a probe was presented on the screen
216 six seconds following the onset of the trial. To avoid confounds resulting from the probe,
217 only the first six seconds of the data in all three conditions were analyzed. Trials that
218 contained fewer than 6000 samples were excluded before analysis. For both datasets, the
219 trials were pooled across participants. After excluding trials, the Exploratory dataset
220 consisted of 12,177 trials and the Confirmatory dataset consisted of 9,301 trials.

221 The raw x-coordinate, y-coordinate, and pupil size data collected at every sampling
222 time point in the trial were used as inputs to the deep learning classifier. These data were
223 also used to develop plot image datasets that were classified separately from the raw timeline

datasets. For the plot image datasets, the timeline data for each trial were converted into scatterplot diagrams. The x- and y- coordinates and pupil size were used to plot each data point onto a scatterplot (e.g., see Figure 1). The coordinates were used to plot the location of the dot, pupil size was used to determine the relative size of the dot, and shading of the dot was used to indicate the time-course of the eye movements throughout the trial. The background of the plot images and first data point were white. Each subsequent data point was one shade darker than the previous data point until the final data point was reached. The final data point was black. For standardization, pupil size was divided by 10, and one unit was added. The plots were sized to match the dimensions of the data collection monitor (1024 x 768 pixels) and then shrunk to (240 x 180 pixels) in an effort to reduce the dimensionality of the data.

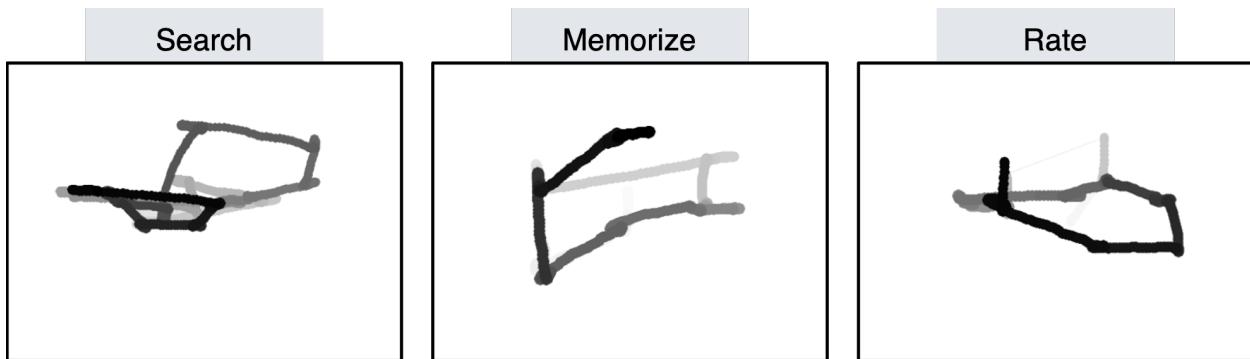


Figure 1. Each trial was represented as an image. Each sample collected within the trial was plotted as a dot in the image. Pupil size was represented by the size of the dot. The time course of the eye movements was represented by the gradual darkening of the dot over time.

Data Subsets. The full timeline dataset was structured into three columns representing the x- and y- coordinates, and pupil size for each data point collected in the first six seconds of each trial. To systematically assess the predictive value of each XY \emptyset (i.e., x-coordinates, y-coordinates, pupil size) component of the data, the timeline and image datasets were batched into subsets that excluded one of the components (i.e., XY \emptyset , X \emptyset P, \emptyset YP), or contained only one of the components (i.e., X \emptyset \emptyset , \emptyset Y \emptyset , \emptyset \emptyset P). For the timeline datasets, this means that the columns to be excluded in each data subset were replaced with zeros. The data were replaced with zeros because removing the columns would change the structure of the data. The same systematic batching process was carried out for the image

²⁴⁴ dataset. See Figure 2 for an example of each of these image data subsets.

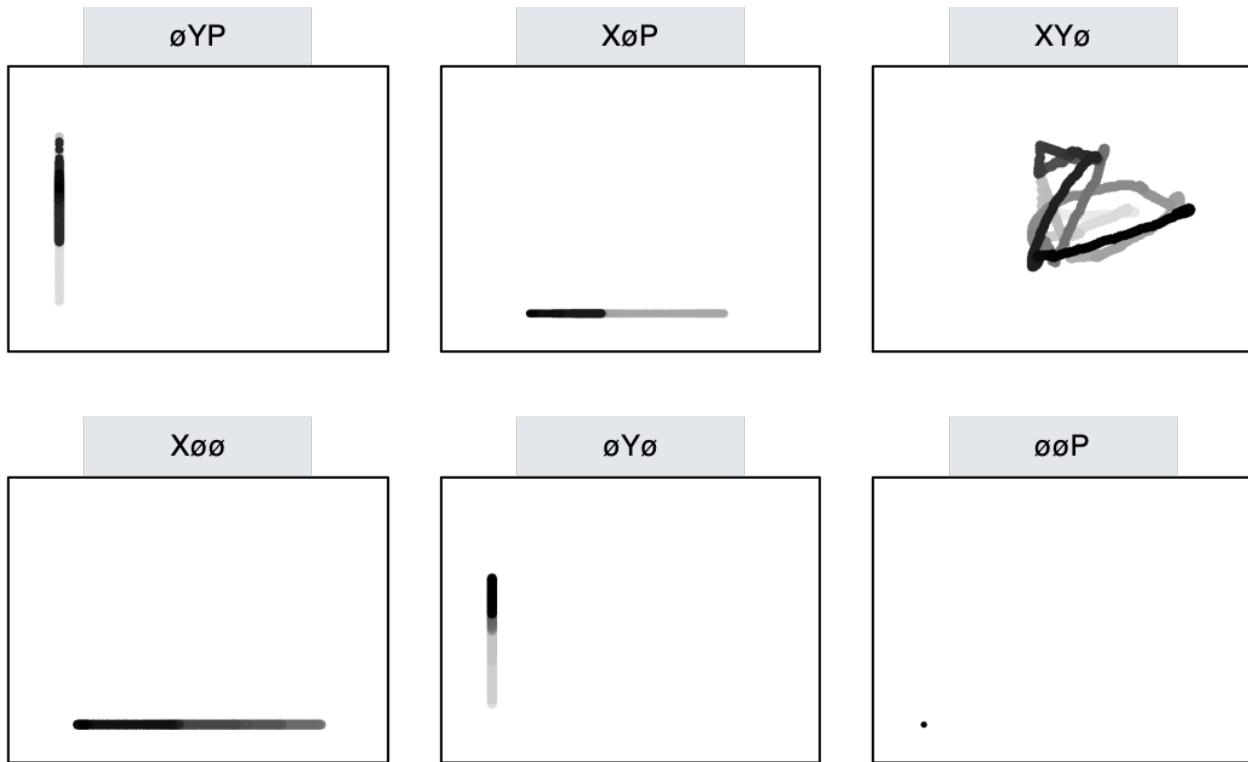


Figure 2. Plot images were used to represent each type of data subset. As with the trials in the full XYP dataset, the time course of the eye movements was represented by the shading of the dot. The first sample of each trial was white, and the last sample was black.

²⁴⁵ Classification

²⁴⁶ Deep CNN model architectures were implemented to classify the trials into Search,
²⁴⁷ Memorize, or Rate categories. Because CNNs act as a digital filter sensitive to the number of
²⁴⁸ features in the data, the differences in the structure of the timeline and image data formats
²⁴⁹ necessitated separate CNN model architectures. The model architectures were developed
²⁵⁰ with the intent of establishing a generalizable approach to classifying cognitive processes
²⁵¹ from eye movement data.

²⁵² The development of these models was not guided by any formal theoretical
²⁵³ assumptions regarding the patterns or features likely to be extracted by the classifier. Like
²⁵⁴ many HCI models, the development of these models followed general intuitions concerned
²⁵⁵ with building a model architecture capable of transforming the data inputs into an

256 interpretable feature set that would not overfit the dataset. The models were developed
257 using version 0.3b of the DeLINEATE toolbox, which operates over a Keras backend
258 (<http://delineate.it>). Each training/test iteration randomly split the data so that 70% of the
259 trial data were allocated to training, 15% of the trial data were allocated to validation, and
260 15% of the trial data were allocated to testing. Training of the model was stopped when
261 validation accuracy did not improve over the span of 100 epochs. Once the early stopping
262 threshold was reached, the resulting model was tested on the held-out test data. This
263 process was repeated 10 times for each model, resulting in 10 classification accuracy scores
264 for each model. The average of the resulting accuracy scores were the subject of comparisons
265 against chance and other datasets or data subsets.

266 The models were developed and tested on the Exploratory dataset. Model
267 hyperparameters were adjusted until the classification accuracies appeared to peak. The
268 model architecture with the highest classification accuracy on the Exploratory dataset was
269 trained, validated, and tested independently on the Confirmatory dataset. This means that
270 the model that was used to analyze the Confirmatory dataset was not trained on the
271 Exploratory dataset. The model architectures used for the timeline and plot image datasets
272 are shown in Figure 3.

273 Analysis

274 Results for the CNN architecture that resulted in the highest accuracy on the
275 Exploratory dataset are reported below. For every dataset tested, a one-sample two-tailed
276 *t*-test was used to compare the CNN accuracies against chance (33%). The Shapiro-Wilk test
277 was used to assess the normality for each dataset. When normality was assumed, the mean
278 accuracy for that dataset was compared against chance using Student's one-sample
279 two-tailed *t*-test. When normality could not be assumed, the median accuracy for that
280 dataset was compared against chance using Wilcoxon's Signed Rank test.

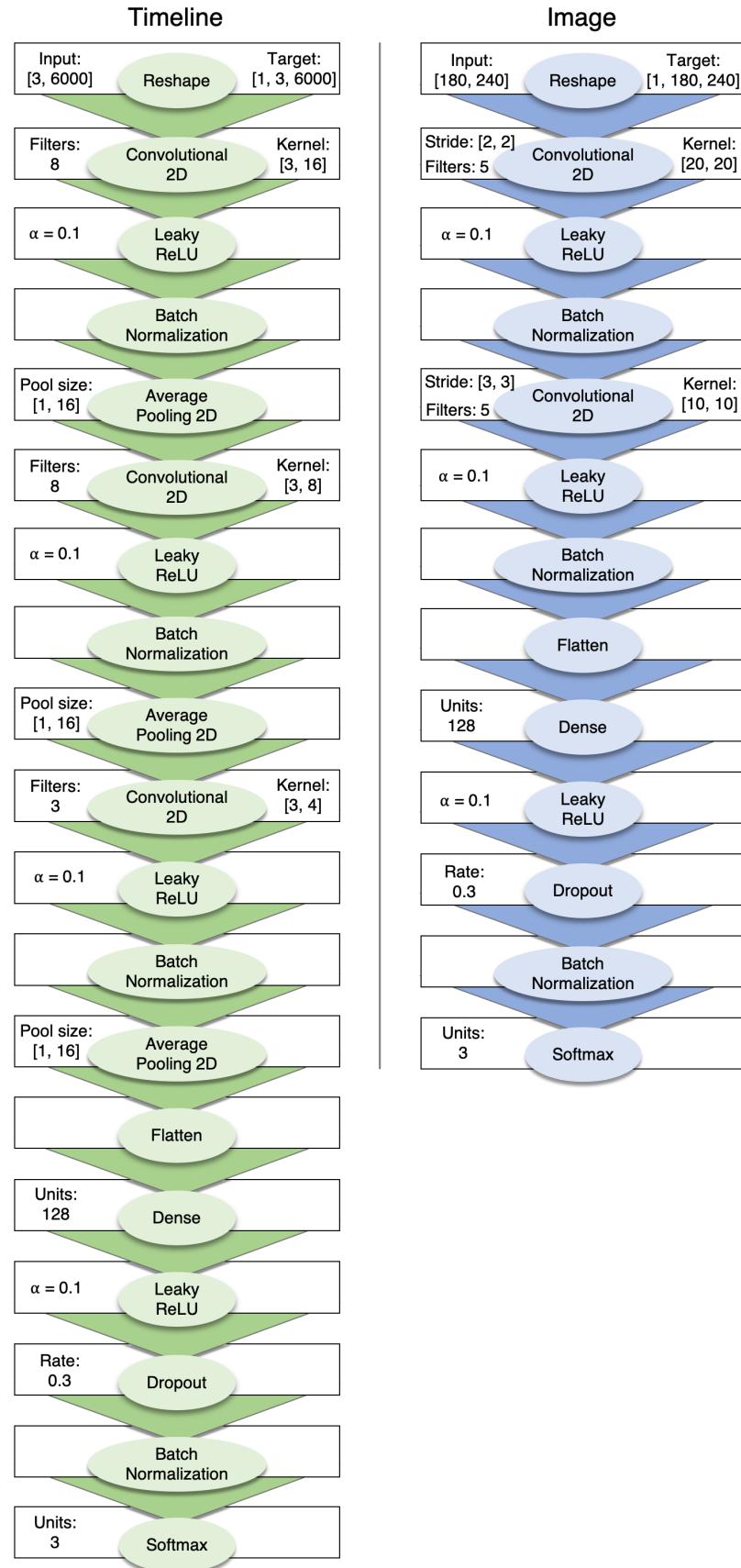


Figure 3. Two different model architectures were used to classify the timeline and image data. Both models were compiled using a categorical crossentropy loss function, and optimized with the Adam algorithm.

281 To determine the relative value of the three components of the eye movement data, the
282 data subsets were compared within the timeline and plot image data types. If classification
283 accuracies were lower when the data were batched into subsets, the component that was
284 removed was assumed to have some unique contribution that the model was using to inform
285 classification decisions. To determine the relative value of the contribution from each
286 component, the accuracies from each subset with one component of the data removed were
287 compared to the accuracies for the full dataset (XYP) using a one-way between-subjects
288 Analysis of Variance (ANOVA). To further evaluate the decodability of each component
289 independently, the accuracies from each subset containing only one component of the eye
290 movement data were compared within a separate one-way between-subjects ANOVA. All
291 post-hoc comparisons were corrected using Tukey's HSD.

292 **Results**

293 **Timeline Data Classification**

294 **Exploratory.** Classification accuracies for the XYP timeline dataset were well above
295 chance (chance = 33%; $M = .526$, $SD = .018$; $t_{(9)} = 34.565$, $p < .001$). Accuracies for
296 classifications of the batched data subsets were all better than chance (see Figure 4). As
297 shown in the confusion matrices displayed in Figure 5, the data subsets with lower overall
298 classification accuracies almost always classified the Memorize condition at or below chance
299 levels of accuracy. Misclassifications of the Memorize condition were split relatively evenly
300 between the Search and Rate conditions.

301 There was a difference in classification accuracy for the XYP dataset and the subsets
302 that had the pupil size, x-coordinate, and y-coordinate data systematically removed ($F_{(3,36)}$
303 $= 47.471$, $p < .001$, $\eta^2 = 0.798$). Post-hoc comparisons against the XYP dataset showed that
304 classification accuracies were not affected by the removal of pupil size or y-coordinate data
305 (see Table 2). The null effect present when pupil size was removed suggests that the pupil
306 size data were not contributing unique information that was not otherwise provided by the x-

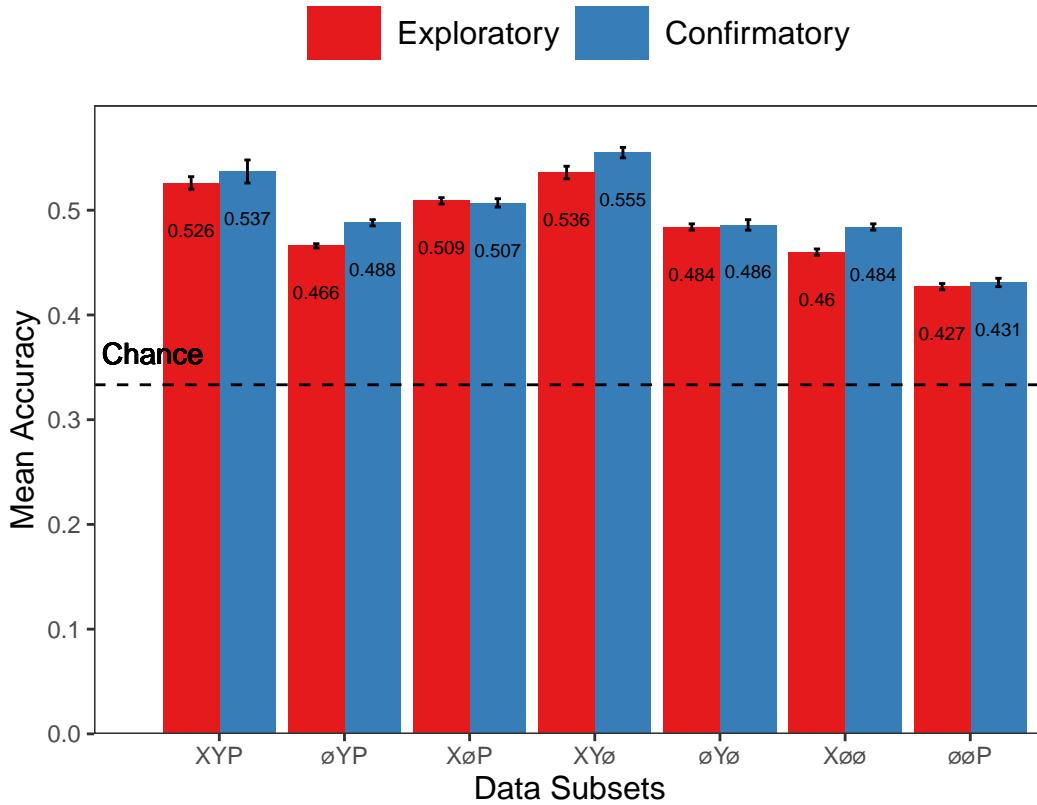


Figure 4. The graph represents the average accuracy reported for each subset of the timeline data. All of the data subsets were decoded at levels better than chance (33%). Each subset is labeled with the mean accuracy. The error bars represent standard errors.

307 and y-coordinates. A strict significance threshold of $\alpha = .05$ implies the same conclusion for
 308 the y-coordinate data, but the relatively low degrees of freedom ($df = 18$) and the borderline
 309 observed p -value ($p = .056$) afford the possibility that there exists a small effect. However,
 310 classification for the $\emptyset Y P$ subset was significantly lower than the $X Y P$ dataset, showing that
 311 the x-coordinate data were uniquely informative to the classification.

312 There was also a difference in classification accuracies for the $X \emptyset \emptyset$, $\emptyset Y \emptyset$, and $\emptyset \emptyset P$
 313 subsets ($F_{(2,27)} = 75.145$, $p < .001$, $\eta^2 = 0.848$). Post-hoc comparisons showed that
 314 classification accuracy for the $\emptyset \emptyset P$ subset was lower than the $X \emptyset \emptyset$ and $\emptyset Y \emptyset$ subsets.
 315 Classification accuracy for the $X \emptyset \emptyset$ subset was higher than the $\emptyset Y \emptyset$ subset. Altogether,
 316 these findings suggest that pupil size data was the least uniquely informative to classification
 317 decisions, while the x-coordinate data was the most uniquely informative.

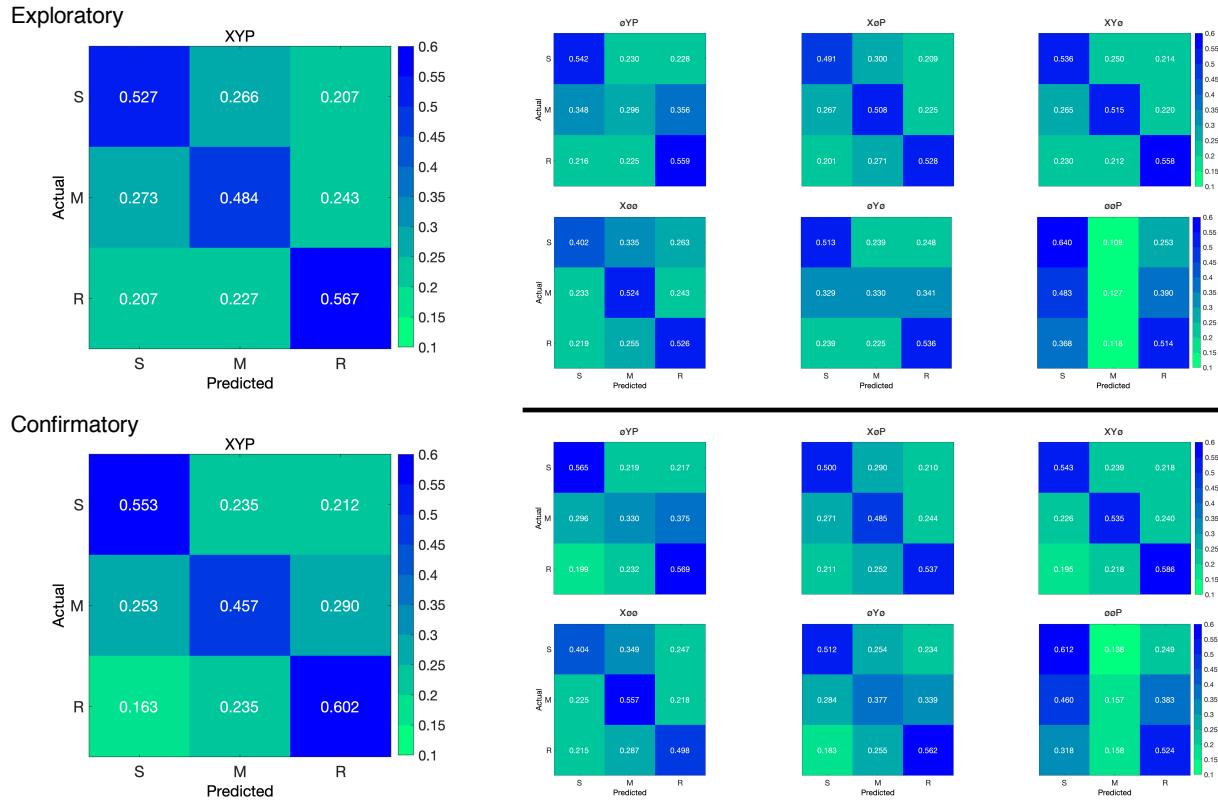


Figure 5. The confusion matrices represent the average classification accuracies for each condition of the timeline data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for trial. The horizontal axis of the confusion matrices represents the condition that was predicted by the model.

Table 2
Timeline Subset Comparisons

Comparison	Exploratory		Confirmatory	
	t	p	t	p
XYP vs. ØYP	9.420	< .001	5.210	< .001
XYP vs. XØP	2.645	.056	3.165	.016
XYP vs. XYø	1.635	.372	1.805	.288
XØØ vs. ØYØ	5.187	< .001	0.495	.874
XØØ vs. ØØP	12.213	< .001	10.178	< .001
ØYØ vs. ØØP	7.026	< .001	9.683	< .001

Confirmatory. Classification accuracies for the Confirmatory XYP timeline dataset

were well above chance ($M = .537$, $SD = 0.036$, $t_{(9)} = 17.849$, $p < .001$). Classification accuracies for the data subsets were also better than chance (see Figure 4). Overall, there was high similarity in the pattern of results for the Exploratory and Confirmatory datasets (see Figure 4). Furthermore, the general trend showing that pupil size was the least

informative eye tracking data component was replicated in the Confirmatory dataset (see Table 2). Also in concordance with the Exploratory timeline dataset, the confusion matrices for these data revealed that the Memorize task was most often confused with the Search and Rate tasks (see Figure 5).

To test the generalizability of the model to other eye tracking data, classification accuracies for the XYP Exploratory and Confirmatory timeline datasets were compared. The Shapiro-Wilk test for normality indicated that the Exploratory ($W = 0.937, p = .524$) and Confirmatory ($W = 0.884, p = .145$) datasets were normally distributed, but Levene's test indicated that the variances were not equal, $F_{(1,18)} = 8.783, p = .008$. Welch's unequal variances *t*-test did not show a difference between the two datasets, $t_{(13.045)} = 0.907, p = .381$, Cohen's *d* = 0.406. These findings indicate that the deep learning model decoded the Exploratory and Confirmatory timeline datasets equally well, but the Confirmatory dataset classifications were less consistent across training/test iterations (as indicated by the increase in standard deviation).

Plot Image Classification

Exploratory. Classification accuracies for the XYP plot image data were better than chance ($M = .436, SD = .020, p < .001$), but were less accurate than the classifications for the XYP Exploratory timeline data ($t_{(18)} = 10.813, p < .001$). Accuracies for the classifications for all subsets of the plot image data except the $\emptyset\emptyset P$ subset were better than chance (see Figure 6). Following the pattern expressed by the timeline dataset, the confusion matrices showed that the Memorize condition was misclassified more often than the other conditions, and appeared to be evenly mis-identified as a Search or Rate condition (see Figure 7).

There was a difference in classification accuracy between the XYP dataset and the data subsets ($F_{(4,45)} = 7.093, p < .001, \eta^2 = .387$). Post-hoc comparisons showed that compared

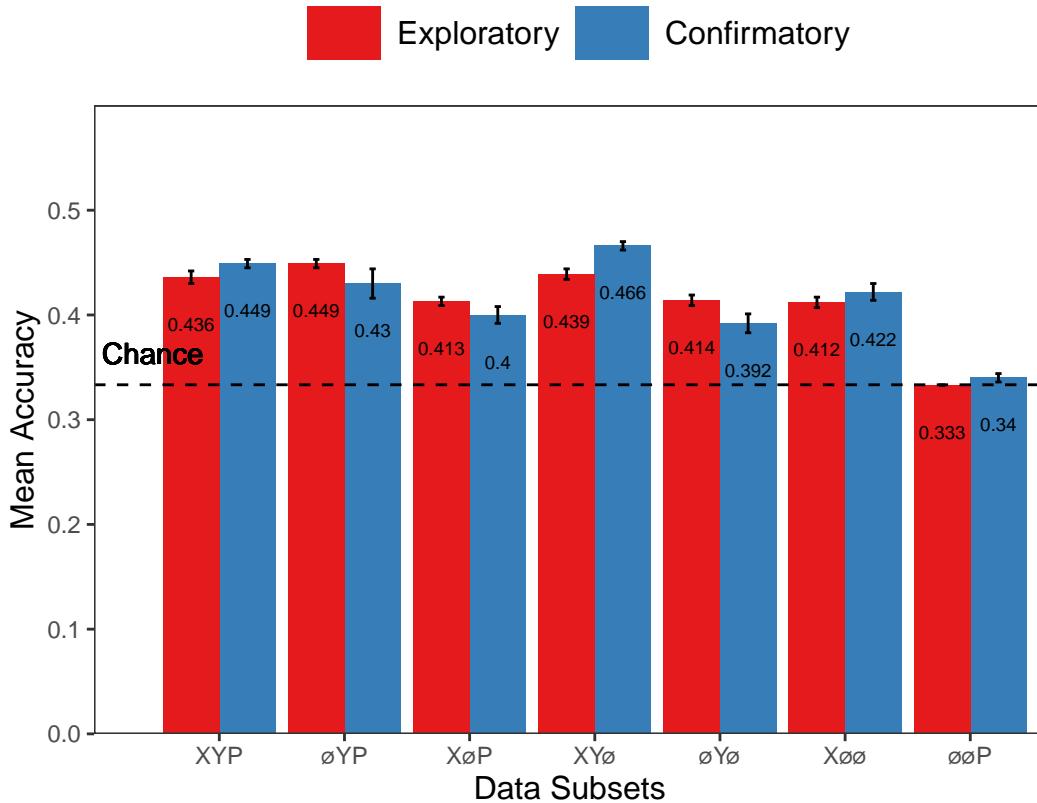


Figure 6. The graph represents the average accuracy reported for each subset of the image data. All of the data subsets except for the Exploratory ØØP dataset were decoded at levels better than chance (33%). Each subset is labeled with the mean accuracy. The error bars represent standard errors.

- ³⁴⁸ to the XYP dataset, there was no effect of removing pupil size or the x-coordinates, but
³⁴⁹ classification accuracy was worse when the y-coordinates were removed (see Table 3).

Table 3
Image Subset Comparisons

Comparison	Exploratory		Confirmatory	
	t	p	t	p
XYP vs. ØYP	1.792	.391	1.623	.491
XYP vs. XØP	2.939	.039	4.375	< .001
XYP vs. XYØ	0.474	.989	1.557	.532
XØØ vs. ØYØ	0.423	.906	2.807	.204
XØØ vs. ØØP	13.569	< .001	5.070	< .001
ØYØ vs. ØØP	13.235	< .001	7.877	< .001

- ³⁵⁰ There was also a difference in classification accuracies between the XØØ, ØYØ, and
³⁵¹ ØØP subsets (Levene's test: $F_{(2,27)} = 3.815$, $p = .035$; Welch correction for lack of

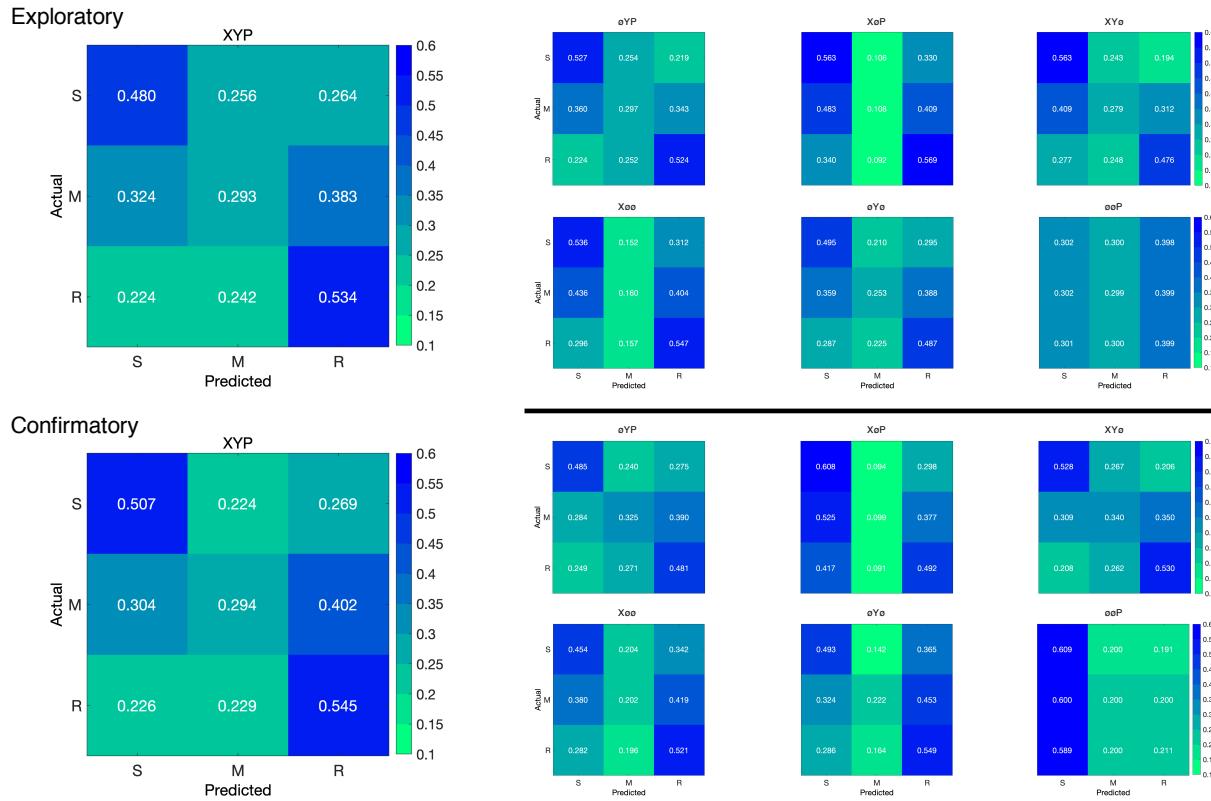


Figure 7. The confusion matrices represent the average classification accuracies for each condition of the image data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents the condition that was predicted by the model.

homogeneity of variances: $F_{(2,17.993)} = 228.137, p < .001, \eta^2 = .899$). Post-hoc comparisons showed that there was no difference in classification accuracies for the Xoo and oYo subsets, but classification for the ooP subset were less accurate than the Xoo and oYo subsets.

Confirmatory. Classification accuracies for the XYP confirmatory image dataset were well above chance ($M = .449, SD = 0.012, t_{(9)} = 31.061, p < .001$), but were less accurate than the classifications of the confirmatory timeline dataset ($t_{(18)} = 11.167 p < .001$). Accuracies for classifications of the data subsets were also all better than chance (see Figure 6). The confusion matrices followed the pattern showing that the Memorize condition was confused most often, and was relatively evenly mis-identified as a Search or Rate trial (see Figure 7). As with the timeline data, the general trend showing that pupil size data was the least informative to the model was replicated in the Confirmatory dataset (see Table 3).

364 To test the generalizability of the model, the classification accuracies for the XYP

365 Exploratory and Confirmatory plot image datasets were compared. The independent samples

366 *t*-test showed that the deep learning model did equally well at classifying the Exploratory

367 and Confirmatory plot image datasets, $t_{(18)} = 1.777$, $p = .092$, Cohen's $d = 0.795$.

368 **Discussion**

369 The present study aimed to produce a practical and reliable example of a black box

370 solution to the inverse Yarbus problem. To implement this solution, we classified raw

371 timeline and minimally processed plot image data using a CNN model architecture. To our

372 knowledge, this study was the first to provide a solution to determining mental state from

373 eye movement data using each of the following: (1) Non-aggregated eye tracking data (i.e.,

374 raw x-coordinates, y-coordinates, pupil size), (2) timeline and image data formats (see

375 Figure 2), and (3) a black box CNN architecture. This study probed the relative predictive

376 value of the x-coordinate, y-coordinate, and pupil size components of the eye movement data

377 using a CNN. The CNN was able to decode the timeline and plot image data better than

378 chance, although only the timeline datasets were decoded with state-of-the-art accuracy.

379 Datasets with lower classification accuracies were not able to differentiate the cognitive

380 processes underlying the Memorize task from the cognitive processes underlying the Search

381 and Rate tasks. Decoding subsets of the data revealed that pupil size was the least uniquely

382 informative component of the eye movement data. This pattern of findings was consistent

383 between the Exploratory and Confirmatory datasets.

384 Although several aggregate eye movement features have been tested as task predictors,

385 to our knowledge, no other study has assessed the predictive value of the data format (viz.,

386 data in the format of a plot image). Our results suggest that although CNNs are robust

387 image classifiers, eye movement data is decoded in the standard timeline format more

388 effectively than in image format. This may be because the image data format contains less

389 decodable information than the timeline format. Over the span of the trial (six seconds), the

390 eye movements occasionally overlapped. When there was an overlap in the image data
391 format, the more recent data points overwrote the older data points. This resulted in some
392 information loss that did not occur when the data were represented in the raw timeline
393 format. Despite this loss of information, the plot image format was still decoded with better
394 than chance accuracy. To further examine the viability of classifying task from eye
395 movement image datasets, future research might consider representing the data in different
396 forms such as 3-dimensional data formats, or more complex color combinations capable of
397 representing overlapping data points.

398 When considering the superior performance of the timeline data (vs., plot image data),
399 we must also consider the differences in the model architectures. Because the structures of
400 the timeline and plot image data formats were different, the models decoding those data
401 structures also needed to be different. Both models were optimized individually on the
402 Exploratory dataset before being tested on the Confirmatory dataset. For both timeline and
403 plot image formats, there was good replicability between the Exploratory and Confirmatory
404 datasets, demonstrating that these architectures performed similarly from experiment to
405 experiment. An appropriately tuned CNN should be capable of learning any arbitrary
406 function, but given that the upper bound for decodability of these datasets is unknown,
407 there is the possibility that a model architecture exists that is capable of classifying the plot
408 image data format more accurately than the model used to classify the timeline data.
409 Despite this possibility, the convergence of these findings with other studies (see Table 1)
410 suggests that the results of this study are approaching a ceiling for the potential to solve the
411 inverse Yarbus problem with eye movement data. Although the true capacity to predict
412 mental state from eye movement data is unknown, standardizing datasets in the future could
413 provide a point for comparison that can more effectively indicate which methods are most
414 effective at solving the inverse Yarbus problem.

415 In the current study, the Memorize condition was most often confused with the Search

416 and Rate conditions, especially for the datasets with lower overall accuracy. This suggests
417 that the eye movements associated with the Memorize task were potentially lacking unique
418 or informative features to decode. This means that eye movements associated with the
419 Memorize condition were interpreted as noise, or were sharing features of underlying
420 cognitive processes that were represented in the eye movements associated with the Search
421 and Rate tasks. Previous research (e.g., Król & Król, 2018) has attributed the inability to
422 differentiate one condition from the others to the overlapping of sub-features in the eye
423 movements between two tasks that are too subtle to be represented in the eye movement
424 data.

425 To more clearly understand how the different tasks influenced the decodability of the
426 eye movement data, additional analyses were conducted on the Exploratory and
427 Confirmatory timeline datasets (see Appendix). These analyses showed that classification
428 accuracy improved when the Memorize condition was removed. A closer look at these results
429 shows that when the Memorize condition was included in the subset, classification accuracies
430 of the Search and Rate conditions was lower. Altogether, these results could indicate that
431 the eye movement features underlying the Memorize condition are shared with the Search
432 and Rate conditions, or that the Memorize condition is contributing a substantial amount of
433 noise.

434 When determining the relative contributions of the the eye movement features used in
435 this study (x-coordinates, y-coordinates, pupil size), the pupil size data was consistently the
436 least uniquely informative. When pupil size was removed from the Exploratory and
437 Confirmatory timeline and plot image datasets, classification accuracy remained stable (vs.,
438 XYP dataset). Furthermore, classification of the ØØP subset was the lowest of all of the data
439 subsets, and in one instance, was no better than chance. Although these findings indicate
440 that, in this case, pupil size was a relatively uninformative component of the eye movement
441 data, previous research has associated changes in pupil size as indicators of working memory

442 load (Kahneman & Beatty, 1966; Karatekin, Couperus, & Marcus, 2004), arousal (Wang et
443 al., 2018), and cognitive effort (Porter, Troscianko, & Gilchrist, 2007). The results of the
444 current study indicate that the changes in pupil size associated with these underlying
445 processes are not useful in delineating the tasks being classified (i.e., Search, Memorize,
446 Rate), potentially because these tasks do not evoke a reliable pattern of changes in pupil size.

447 The findings from the current study support the notion that black box CNNs are a
448 viable approach to determining task from eye movement data. In a recent review, Lukander
449 et al. (2017) expressed concern regarding the lack of generalizability of black box approaches
450 when decoding eye movement data. Overall, the current study showed a consistent pattern
451 of results for the XYP timeline and image datasets, but some minor inconsistencies in the
452 pattern of results for the x- and y- coordinate subset comparisons. These inconsistencies may
453 be a product of overlap in the cognitive processes underlying the three tasks. When the data
454 are batched into subsets, at least one dimension (i.e., x-coordinates, y-coordinates, or pupil
455 size) is removed, leading to a potential loss of information. When the data provide fewer
456 meaningful distinctions, finer-grained inferences are necessary for the tasks to be
457 distinguishable. As shown by Coco and Keller (2014), eye movement data can be more
458 effectively decoded when the cognitive processes underlying the tasks are explicitly
459 differentiable. While the cognitive processes distinguishing memorizing, searching, or rating
460 an image are intuitively different, the eye movements elicited from these cognitive processes
461 are not easily differentiated. To correct for potential mismatches between the distinctive
462 task-diagnostic features in the data and the level of distinctiveness required to classify the
463 tasks, future research could more definitively conceptualize the cognitive processes
464 underlying the task-at-hand.

465 Classifying mental state from eye movement data is often carried out in an effort to
466 advance technology to improve educational outcomes, strengthen the independence of
467 physically and mentally handicapped individuals, or improve HCI's (Koochaki &

468 Najafizadeh, 2018). Given the previous questions raised regarding the reliability and
469 generalizability of black-box CNN classification, the current study first tested models on an
470 exploratory dataset, then confirmed the outcome using a second independent dataset.
471 Overall, the findings of this study indicate that this black-box approach is capable of
472 producing a stable and generalizable outcome. Future studies that incorporate stimulus
473 features might have the potential to surpass current state-of-the-art classification. According
474 to Bulling, Weichel, and Gellersen (2013), incorporating stimulus feature information into
475 the dataset may provide improve accuracy relative to decoding gaze location data and pupil
476 size. Alternatively, Borji and Itti (2014) suggested that accounting for salient features in the
477 the stimulus might leave little to no room for theoretically defined classifiers to consider
478 mental state. Future research should examine the potential for the inclusion of stimulus
479 feature information in addition to the eye movement data to boost black-box CNN
480 classification accuracy of image data beyond that of timeline data.

481

References

- 482 Boisvert, J. F. G., & Bruce, N. D. B. (2016). Predicting task from eye movements: On the
483 importance of spatial distribution, dynamics, and image features. *Neurocomputing*,
484 207, 653–668. <https://doi.org/10.1016/j.neucom.2016.05.047>
- 485 Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task.
486 *Journal of Vision*, 14(3), 29–29. <https://doi.org/10.1167/14.3.29>
- 487 Bulling, A., Weichel, C., & Gellersen, H. (2013). EyeContext: Recognition of high-level
488 contextual cues from human visual behaviour. In *Proceedings of the SIGCHI
489 Conference on Human Factors in Computing Systems - CHI '13* (p. 305). Paris,
490 France: ACM Press. <https://doi.org/10.1145/2470654.2470697>
- 491 Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye
492 movement control during active scene perception. *Journal of Vision*, 9(3), 6–6.
493 <https://doi.org/10.1167/9.3.6>
- 494 Coco, M. I., & Keller, F. (2014). Classification of visual and linguistic tasks using
495 eye-movement features. *Journal of Vision*, 14(3), 11–11.
496 <https://doi.org/10.1167/14.3.11>
- 497 DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited.
498 *Visual Cognition*, 17(6-7), 790–811. <https://doi.org/10.1080/13506280902793843>
- 499 Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to predict
500 observers' task from eye movement patterns. *Vision Res*, 62, 1–8.
501 <https://doi.org/10.1016/j.visres.2012.03.019>
- 502 Haji-Abolhassani, A., & Clark, J. J. (2014). An inverse Yarbus process: Predicting observers'
503 task from eye movement patterns. *Vision Research*, 103, 127–142.

504 <https://doi.org/10.1016/j.visres.2014.08.014>

505 Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013).

506 Predicting Cognitive State from Eye Movements. *PLoS ONE*, 8(5), e64937.

507 <https://doi.org/10.1371/journal.pone.0064937>

508 Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*,

509 154(3756), 1583–1585. Retrieved from <https://www.jstor.org/stable/1720478>

510 Kanan, C., Ray, N. A., Bseiso, D. N. F., Hsiao, J. H., & Cottrell, G. W. (2014). Predicting

511 an observer's task using multi-fixation pattern analysis. In *Proceedings of the*

512 *Symposium on Eye Tracking Research and Applications - ETRA '14* (pp. 287–290).

513 Safety Harbor, Florida: ACM Press. <https://doi.org/10.1145/2578153.2578208>

514 Karatekin, C., Couperus, J. W., & Marcus, D. J. (2004). Attention allocation in the

515 dual-task paradigm as measured through behavioral and psychophysiological

516 responses. *Psychophysiology*, 41(2), 175–185.

517 <https://doi.org/10.1111/j.1469-8986.2004.00147.x>

518 Koochaki, F., & Najafizadeh, L. (2018). Predicting Intention Through Eye Gaze Patterns.

519 In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (pp. 1–4).

520 <https://doi.org/10.1109/BIOCAS.2018.8584665>

521 Król, M. E., & Król, M. (2018). The right look for the job: Decoding cognitive processes

522 involved in the task from spatial eye-movement patterns. *Psychological Research*.

523 <https://doi.org/10.1007/s00426-018-0996-5>

524 Lukander, K., Toivanen, M., & Puolamäki, K. (2017). Inferring Intent and Action from Gaze

525 in Naturalistic Behavior: A Review. *International Journal of Mobile Human*

526 *Computer Interaction*, 9(4), 41–57. <https://doi.org/10.4018/IJMHCI.2017100104>

- 527 MacInnes, W., Joseph, Hunt, A. R., Clarke, A. D. F., & Dodd, M. D. (2018). A Generative
528 Model of Cognitive State from Task and Eye Movements. *Cognitive Computation*,
529 10(5), 703–717. <https://doi.org/10.1007/s12559-018-9558-9>
- 530 Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011).
531 Examining the influence of task set on eye movements and fixations. *Journal of*
532 *Vision*, 11(8), 17–17. <https://doi.org/10.1167/11.8.17>
- 533 Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and
534 counting: Insights from pupillometry. *Quarterly Journal of Experimental Psychology*
535 (2006), 60(2), 211–229. <https://doi.org/10.1080/17470210600673818>
- 536 Seeliger, K., Fritzsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., &
537 van Gerven, M. A. J. (2018). Convolutional neural network-based encoding and
538 decoding of visual object recognition in space and time. *NeuroImage*, 180, 253–266.
539 <https://doi.org/10.1016/j.neuroimage.2017.07.018>
- 540 Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus,
541 Eye Movements, and Vision. *I-Perception*, 1(1), 7–27. <https://doi.org/10.1068/i0382>
- 542 Wang, C.-A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C., & Munoz, D. P. (2018).
543 Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional
544 Face Task. *Frontiers in Neurology*, 9. <https://doi.org/10.3389/fneur.2018.01029>
- 545 Yarbus, A. (1967). Eye Movements and Vision. Retrieved January 24, 2019, from
546 [http://wexler.free.fr/library/files/yarbus%20\(1967\)%20eye%20movements%20and%20vision.pdf](http://wexler.free.fr/library/files/yarbus%20(1967)%20eye%20movements%20and%20vision.pdf)
- 548 Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2019). Comparing the Interpretability of Deep
549 Networks via Network Dissection. In W. Samek, G. Montavon, A. Vedaldi, L. K.
550 Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and*

- 551 *Visualizing Deep Learning* (pp. 243–252). Cham: Springer International Publishing.
- 552 https://doi.org/10.1007/978-3-030-28954-6_12

Appendix

553 Additional analyses were conducted to clarify the effect of task on classification accuracy.
554 These supplementary analyses were not seen as central to the current study, but could prove
555 to be informative to researchers attempting to replicate or extend these findings in the
556 future. The results from the primary analyses showed that classification accuracies were the
557 lowest for the Memorize condition, but these findings did not indicate if the Memorize
558 condition was adding noise to the data, or was providing redundant information to the
559 model. To further understand why classification accuracy was lower for the Memorize
560 condition than it was for the Search or Rate condition, the Exploratory and Confirmatory
561 timeline datasets were systematically batched into subsets with the Search (S), Memorize
562 (M), or Rate (R) condition removed (i.e., \emptyset MR, S \emptyset R, SM \emptyset).

563 Overall, the accuracies for all of the data subsets observed in the supplementary
564 analysis were higher than the accuracies observed in the main analysis (see Figure A1).
565 Chance accuracy levels for the primary analysis was 33%, but because one of the tasks was
566 removed from each element observed in the supplementary analyses, chance accuracy for
567 these analyses was 50%. Given the data analyzed for these supplementary purposes have
568 different thresholds of chance performance, any conclusions drawn from a comparison
569 between the primary and supplementary datasets could be misleading. For this reason, this
570 supplementary analysis is focused only on comparing the data subsets with one task removed.

571 All of the data subsets analyzed in this supplementary analysis were decoded with
572 better than chance accuracy (see Figure A1). The same pattern of results was observed in
573 both the Exploratory and Confirmatory datasets. When the Memorize condition was
574 removed, classification accuracy improved (see Table A1). When the Rate condition was
575 removed, classification was the worst. When the Memorize condition was included, the
576 Memorize condition was more accurately predicted than the Search and Rate conditions (see

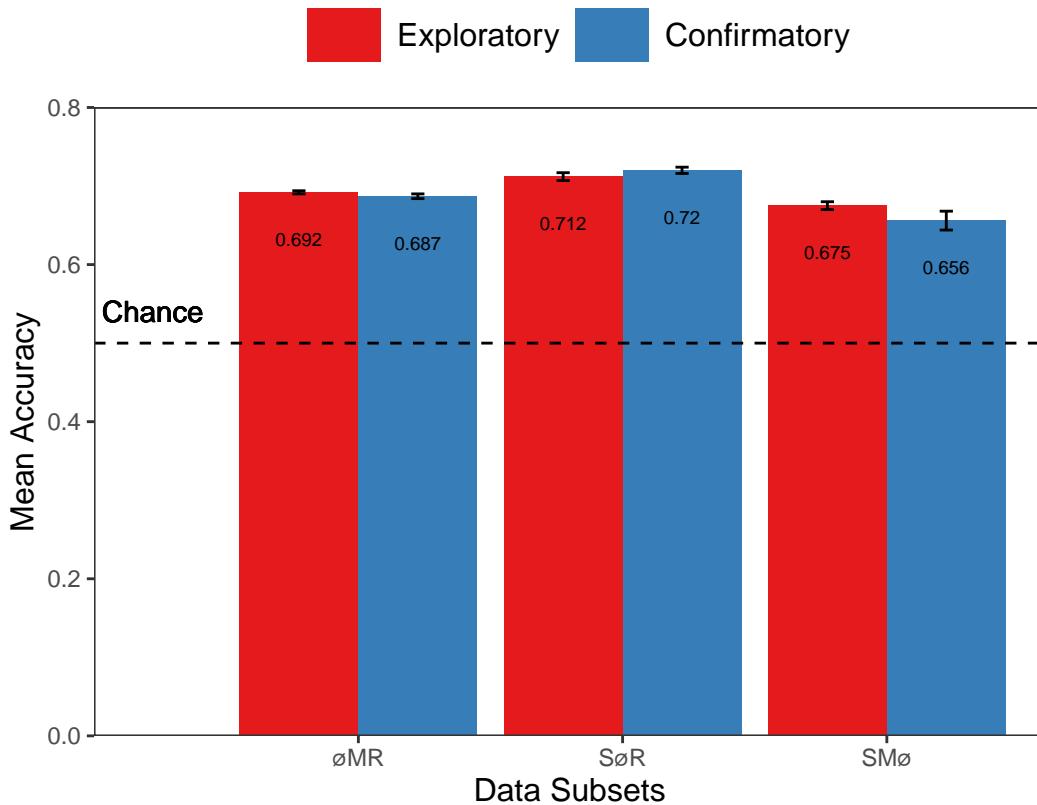


Figure A1. The graph represents the average accuracy reported for each subset of the Exploratory and Confirmatory timeline data. All of the data subsets were decoded at levels better than chance (50%). The error bars represent standard errors.

⁵⁷⁷ Figure A2).

Table A1
Supplementary Subset Comparisons

Comparison	Exploratory		Confirmatory	
	t	p	t	p
øMR vs. SøR	3.248	.008	3.094	.012
øMR vs. SMø	2.875	.021	2.923	.018
SøR vs. SMø	6.123	< .001	6.017	< .001

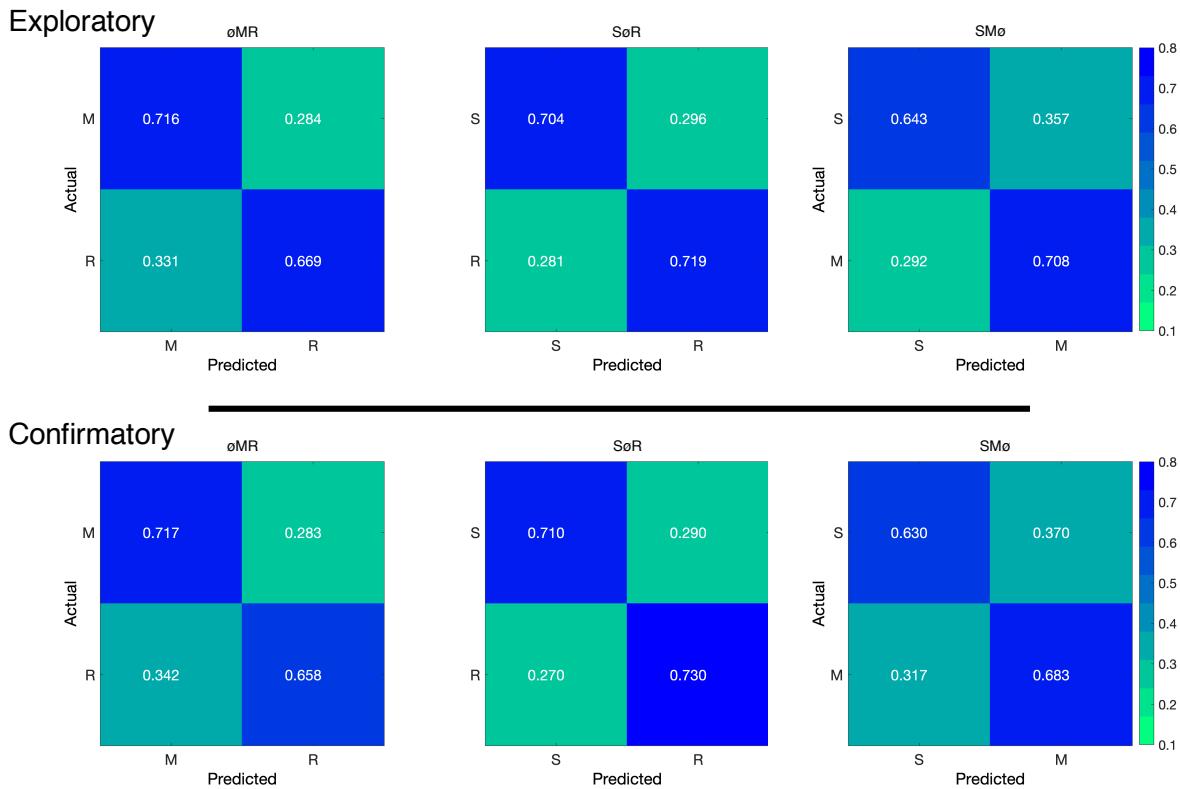


Figure A2. The confusion matrices represent the average classification accuracies for each condition of the timeline data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for trial. The horizontal axis of the confusion matrices represents the condition that was predicted by the model.