

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. Kuntzelman¹, 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. Work
done to develop the analysis approach was supported by NSF/EPSCoR grant #1632849
(MRJ and MDD). Additionally, this work was supported by the National Institute of General
Medical Sciences of the National Institutes of Health [grant number P20 GM130461 awarded
to MRJ and colleagues] and the Rural Drug Addiction Research Center at the University of
Nebraska-Lincoln. The content is solely the responsibility of the authors and does not
necessarily represent the official views of the National Institutes of Health or the University
of Nebraska.

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

17

Abstract

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

38 *Keywords:* deep learning, eye tracking, convolutional neural network, cognitive state,
39 endogenous attention

40

Introduction

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) showed 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 **task using** eye movement data, also known as the “inverse Yarbus process” (Haji-Abolhassani
50 & Clark, 2014).

51 Current state-of-the-art machine learning and neural network algorithms are capable of
52 identifying diagnostic patterns for the purpose of decoding a variety of data types, but the
53 inner workings of the resulting model solutions are difficult or impossible to interpret.

54 Algorithms that provide such solutions are referred to as *black box* models. Dissections of
55 black box models have been largely uninformative (Zhou, Bau, Oliva, & Torralba, 2019),
56 limiting the potential for researchers to apply the mechanisms underlying successful
57 classification of the data. Still, black box models provide a powerful solution for
58 technological applications such as human-computer interfaces (HCI; for a review, see
59 Lukander, Toivanen, & Puolamäki, 2017). While the internal operations of the model
60 solutions used for HCI applications do not necessarily need to be interpretable to serve their
61 purpose, Lukander, Toivanen, and Puolamäki (2017) pointed out that the inability to
62 interpret the mechanisms underlying the function of black box solutions impedes the
63 generalizability of these methods, and increases the difficulty of expanding these findings to
64 real life applications. To ground these solutions, researchers guide decoding efforts by using
65 eye movement data and/or models with built-in theoretical assumptions. For instance, eye

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

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

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. Others may define these terms differently, but for present purposes, our working
97 definitions are that cognitive “processes” are theoretical constructs that could be difficult to
98 isolate in practice, whereas a “task” is a more concrete/explicit set of goals and behaviors
99 imposed by the experimenter in an effort to operationalize one or more cognitive processes.
100 A “mental state,” in contrast, is also a more theoretical term that is a bit more general and
101 could include goals and cognitive processes, but could also presumably encompass other
102 elements like mood or distraction.

103 To differentiate the cognitive processes underlying task-evoked eye movements, some
104 studies have chosen to classify tasks that rely on stimuli that prompt easily distinguishable
105 eye movements, such as reading text (e.g., Henderson, Shinkareva, Wang, Luke, &
106 Olejarczyk, 2013). The eye movements elicited by salient stimulus features facilitate task
107 classifications; however, because these eye movements are the consequence of a feature (or
108 features) inherent to the stimulus rather than the task, it is unclear if these classifications
109 are attributable to the stimulus or a complex mental state (Boisvert & Bruce, 2016; e.g.,
110 Henderson, Shinkareva, Wang, Luke, & Olejarczyk, 2013). Additionally, the distinct nature
111 of exogenously elicited eye movements prompts decoding algorithms to prioritize these
112 bottom-up patterns in the data over higher-level top-down effects (Borji & Itti, 2014). This
113 means that these models are identifying the type of information that is being processed, but
114 are not necessarily reflecting the mental state of the individual observing the stimulus. Eye
115 movements that are the product of bottom-up processes have been reliably decoded, which is
116 relevant for some HCI applications; however, in our view such efforts do not fit the spirit of
117 the inverse Yarbus problem, as most groups seem to construe it. Namely, most attempts at
118 addressing the inverse Yarbus problem are concerned with decoding higher-level abstract

119 mental operations that can be applied to virtually any naturalistic image and are not
120 /textcolor{red}{necessarily} dependent on specific structural elements of the stimuli (e.g.,
121 the highly regular, linear patterns of written text).

122 Currently, there is not a clearly established upper limit to how well cognitive task can
123 be classified from eye movement data. Prior evidence has shown that the task-at-hand is
124 capable of producing distinguishable eye movement features such as the total scan path
125 length, total number of fixations, and the amount of time to the first saccade (Castelhano,
126 Mack, & Henderson, 2009; DeAngelus & Pelz, 2009). Decoding accuracies within the context
127 of determining task from eye movements typically range from chance performance to
128 relatively robust classification (see Table 1). In one case, Coco and Keller (2014) categorized
129 the same eye movement features used by Greene, Liu, and Wolfe (2012) with respect to the
130 relative contribution of latent visual or linguistic components of three tasks (visual search,
131 name the picture, name objects in the picture) with 84% accuracy (chance = 33%). While
132 this manipulation is reminiscent of other experiments relying on the bottom-up influence of
133 words and pictures (Boisvert & Bruce, 2016; e.g., Henderson, Shinkareva, Wang, Luke, &
134 Olejarczyk, 2013) the eye movements in the Coco and Keller (2014) tasks can be attributed
135 to the occurrence of top-down attentional processes. A conceptually related follow-up to this
136 study classified tasks along two spatial and semantic dimensions, resulting in 51%
137 classification accuracy [chance = 25%; Król and Król (2018)]. A closer look at these results
138 showed that the categories within the semantic dimension were consistently misclassified,
139 suggesting that this level of distinction may require a richer dataset, or a more powerful
140 decoding algorithm. Altogether, there is no measurable index of relative top-down or
141 bottom-up influence, but this body of literature suggests that the relative influence of
142 top-down and bottom-up attentional processes may have a role in determining the
143 decodability of the eye movement data.

144 As shown in Table 1, when eye movement data are prepared for classification, fixation

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%)

¹⁴⁵ and saccade statistics are typically aggregated along spatial or temporal dimensions,¹⁴⁶ resulting in variables such as fixation density or saccade amplitude (Castelhano, Mack, &

147 Henderson, 2009; MacInnes, Hunt, Clarke, & Dodd, 2018; Mills, Hollingworth, Van der
148 Stigchel, Hoffman, & Dodd, 2011). The implementation of these statistical methods is meant
149 to explicitly provide the decoding algorithm with characteristics of the eye movement data
150 that are representative of theoretically relevant cognitive processes. For example, MacInnes,
151 Hunt, Clarke, and Dodd (2018) attempted to provide an algorithm with data designed to be
152 representative of inputs to the frontal eye fields. In some instances, such as the case of Król
153 and Król (2018), grounding the data using theoretically driven aggregation methods may
154 require sacrificing granularity in the dataset. This means that aggregating the data has the
155 potential to wash out certain fine-grained distinctions that could otherwise be detected.
156 Data structures of any kind can only be decoded to the extent to which the data are capable
157 of representing differences between categories. Given that the cognitive processes underlying
158 distinct tasks are often overlapping (Coco & Keller, 2014), decreasing the granularity of the
159 data may actually limit the potential of the algorithm to make fine-grained distinctions
160 between diagnostic components underlying the tasks to be decoded.

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

174 movement data.

175 The current study explored pragmatic solutions to the problem of classifying task from
176 eye movement data by submitting minimally processed x-coordinate, y-coordinate, and pupil
177 size data to a convolutional neural network (CNN) model. Instead of transforming the data
178 into theoretically defined units, we allowed the network to learn meaningful patterns in the
179 data on its own. CNNs have a natural propensity to develop low-level feature detectors
180 similar to the primary visual cortex (e.g., Seeliger et al., 2018); for this reason, they are
181 commonly implemented for image classification. In some cases, researchers have found
182 success classifying data that natively exist in a timeline format by first transforming the data
183 to an image-based format and then passing it to a deep neural network classifier (e.g.,
184 Bashivan, Rish, Yeasin, & Codella, 2016); however, it is not always obvious a priori which
185 representation of a particular type of data is best-suited for neural network classifiers to be
186 able to detect informative features, and the ideal representational format must be
187 determined empirically. Thus, to test the possibility that image data might be better suited
188 to the CNN classifier in our eye movement data as well, we also transformed our dataset
189 from raw timelines into simple image representations and compared CNN-based classification
190 of timeline data to that of image data. The image representations we generated also matched
191 the eye movement trace images classically associated with the work of Yarbus (1967) and
192 others, which were the original forays into this line of inquiry.

193 To our knowledge, no study has attempted to address the inverse Yarbus problem
194 using any combination of the following methods: (1) Non-aggregated data, (2) image data
195 format, and (3) a black-box CNN architecture. Given that CNN architectures are capable of
196 learning features represented in raw data formats, and are well-suited to decoding
197 multidimensional data that have a distinct spatial or temporal structure, we expected that a
198 non-theoretically-constrained CNN architecture could be capable of decoding data at levels
199 consistent with the current state of the art. Furthermore, despite evidence that black box

200 approaches to the inverse Yarbus problem can impede generalizability (Lukander, Toivanen,
201 & Puolamäki, 2017), we expected that when testing the approach on an entirely separate
202 dataset, providing the model with minimally processed data and the flexibility to identify
203 the unique features within each dataset would result in the replication of our initial findings.

204

Method

205 **Participants**

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

214 **Materials and Procedures**

215 Each participant viewed a series of indoor and outdoor scene images while carrying out
216 a search, memorization, or rating task. For the memorization task, participants were
217 instructed to memorize the image in anticipation of a forced choice recognition test. At the
218 end of each Memorize trial, the participants were prompted to indicate which of two images
219 was just presented. The two images were identical outside of a small change in the display
220 (e.g. object removed or added to the scene). For the rating task, participants were asked to
221 think about how they would rate the image on a scale from 1 (very unpleasant) to 7 (very
222 pleasant). The participants were prompted to provide a rating immediately after viewing the
223 image. For the search task, participants were instructed to find a small 'Z' or 'N' embedded
224 in the image. In reality, targets were not present in the images outside of a small subset of

225 images ($n = 5$) that were not analyzed but were included in the experiment design so
226 participants believed a target was always present. Trials containing the target were excluded
227 because search behavior was likely to stop if the target was found, adding considerable noise
228 to the eye movement data. For consistency between trial types, participants were prompted
229 to indicate if they found a ‘Z’ or ‘N’ at the end of each Search trial.

230 The same materials were used in both experiments with a minor variation in the
231 procedures. In the Confirmatory experiment, participants were directed as to where search
232 targets might appear in the image (e.g., on flat surfaces). No such instructions were provided
233 in the Exploratory experiment.

234 In both experiments, participants completed one mixed block of 120 trials (task cued
235 prior to each trial), or three uniform blocks of 40 trials (task cued prior to each block for a
236 total of 120 trials). Block type was assigned in counterbalanced order. When the blocks were
237 mixed, the trial types were randomly intermixed within the block. For uniform blocks, each
238 block consisted entirely of one of the three conditions (Search, Memorize, Rate), with block
239 types presented in random order. Each stimulus image was presented for 8 seconds. The
240 pictures were presented in color, with a size of 1024 x 768 pixels, subtending a visual angle of
241 23.8° x 18.0°.

242 Eye movements were recorded using an SR Research EyeLink 1000 eye tracker with a
243 sampling rate of 1000Hz. Only the right eye was recorded. The system was calibrated using
244 a nine-point accuracy and validity test. Errors greater than 1° or averaging greater than 0.5°
245 in total were re-calibrated.

246 **Datasets**

247 On some trials, a probe was presented on the screen six seconds after the onset of the
248 trial, which required participants to fixate the probe once detected. To avoid confounds
249 resulting from the probe, only the first six seconds of the data for each trial was analyzed.

250 Trials that contained fewer than 6000 samples within the first six seconds of the trial were
251 excluded before analysis. For both datasets, the trials were pooled across participants. After
252 excluding trials, the Exploratory dataset consisted of 12,177 of the 16,740 total trials, and
253 the Confirmatory dataset consisted of 9,301 of the 10,395 total trials.

254 The raw x-coordinate, y-coordinate, and pupil size data collected at every sampling
255 time point in the trial were used as inputs to the deep learning classifier. These data were
256 also used to develop plot image datasets that were classified separately from the raw timeline
257 datasets. For the plot image datasets, the timeline data for each trial were converted into
258 scatterplot diagrams. The x- and y- coordinates and pupil size were used to plot each data
259 point onto a scatterplot (e.g., see Figure 1). The coordinates were used to plot the location
260 of the dot, pupil size was used to determine the relative size of the dot, and shading of the
261 dot was used to indicate the time-course of the eye movements throughout the trial. The
262 background of the plot images and first data point were white. Each subsequent data point
263 was one shade darker than the previous data point until the final data point was reached.
264 The final data point was black. For standardization, pupil size was divided by 10, and one
265 unit was added. The plots were sized to match the dimensions of the data collection monitor
266 (1024 x 768 pixels) and then shrunk to (240 x 180 pixels) in an effort to reduce the
267 dimensionality of the data.

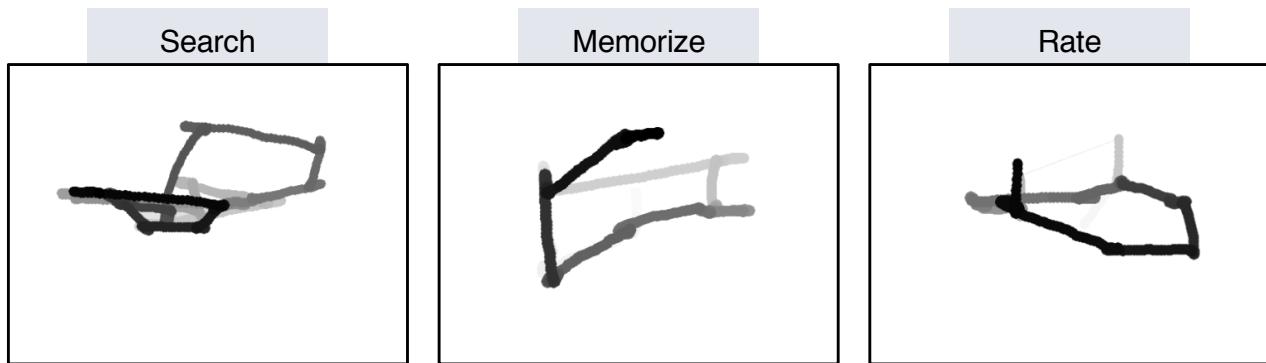


Figure 1. Each trial was represented as an image. Each sample collected within the trial was plotted as a dot in the image. Pupil size was represented by the size of the dot. The time course of the eye movements was represented by the gradual darkening of the dot over time.

268 **Data Subsets.** The full timeline dataset was structured into three columns

269 representing the x- and y- coordinates, and pupil size for each data point collected in the
270 first six seconds of each trial. To systematically assess the predictive value of each XYP (i.e.,
271 x-coordinates, y-coordinates, pupil size) component of the data, the timeline and image
272 datasets were batched into subsets that excluded one of the components (i.e., XY \emptyset , X \emptyset P,
273 \emptyset YP), or contained only one of the components (i.e., X \emptyset \emptyset , \emptyset Y \emptyset , \emptyset \emptyset P). For the timeline
274 datasets, this means that the columns to be excluded in each data subset were replaced with
275 zeros. The data were replaced with zeros because removing the columns would change the
276 structure of the data. The same systematic batching process was carried out for the image
277 dataset. See Figure 2 for an example of each of these image data subsets.

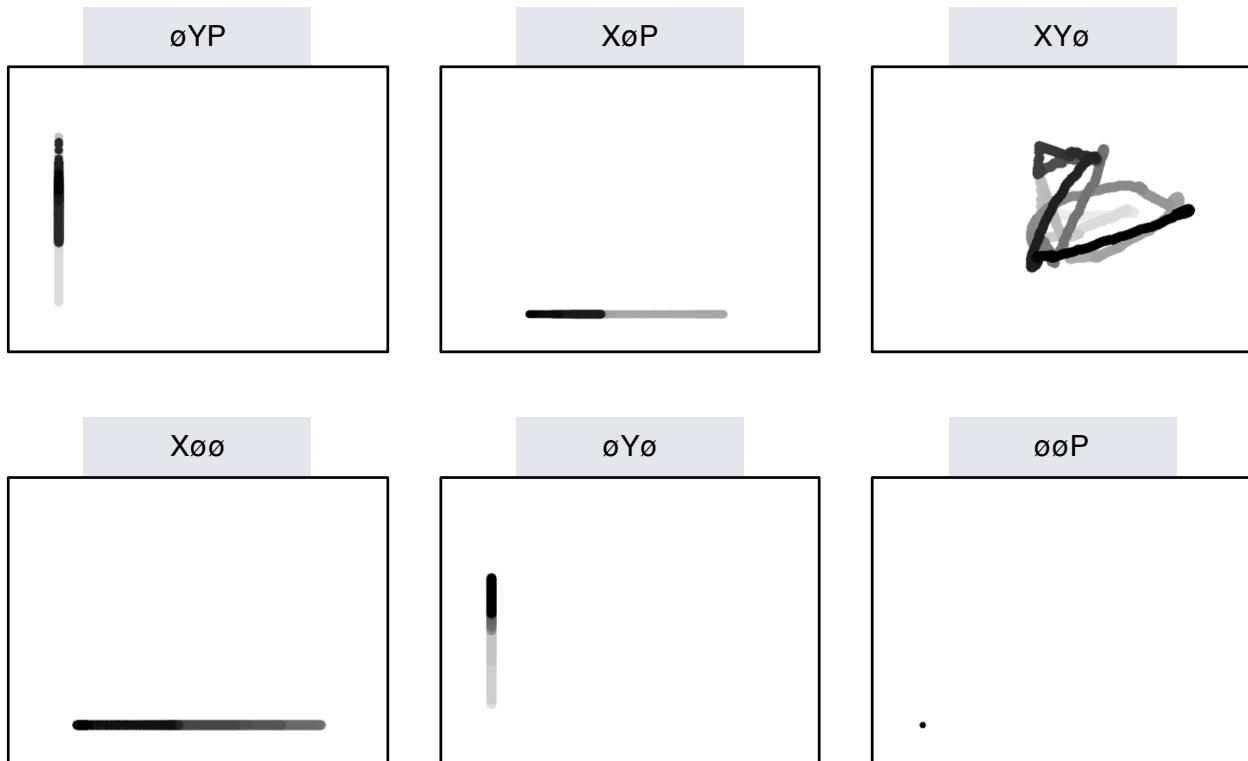


Figure 2. Plot images were used to represent data subsets that excluded one component of the eye movement data (i.e., XY \emptyset , X \emptyset P, \emptyset YP) or contained only one component (i.e., X \emptyset \emptyset , \emptyset Y \emptyset , \emptyset \emptyset P). As with the trials in the full 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.

278 Classification

279 Deep CNN model architectures were implemented to classify the trials into Search,
280 Memorize, or Rate categories. Because CNNs act as a digital filter sensitive to the number of
281 features in the data, the differences in the structure of the timeline and image data formats
282 necessitated separate CNN model architectures. The model architectures were developed
283 with the intent of establishing a generalizable approach to classifying task from eye
284 movement data.

285 The development of these models was not guided by any formal theoretical
286 assumptions regarding the patterns or features likely to be extracted by the classifier. Like
287 many HCI models, the development of these models followed general intuitions concerned
288 with building a model architecture capable of transforming the data inputs into an
289 interpretable feature set that would not overfit the dataset. The models were developed
290 using version 0.3b of the DeLINEATE toolbox, which operates over a Keras backend
291 (<http://delineate.it>) (Kuntzelman et al., 2021). Each training/test iteration randomly split
292 the data so that 70% of the trials were allocated to training, 15% to validation, and 15% to
293 testing. (This approach achieves essentially the same benefit of a more traditional k-fold
294 cross-validation approach insofar as it allows all data to be used as both training and test
295 without double-dipping; however, by resampling the data instead of using strict fold
296 divisions, we can sidestep the issue of how to incorporate a validation set into the k-fold
297 approach.) Training of the model was stopped when validation accuracy did not improve
298 over the span of 100 epochs. Once the early stopping threshold was reached, the resulting
299 model was tested on the held-out test data. This process was repeated 10 times for each
300 model, resulting in 10 classification accuracy scores for each model. The resulting accuracy
301 scores were used for the comparisons against chance and other datasets or data subsets.

302 The models were developed and tested on the Exploratory dataset. Model
303 hyperparameters were adjusted until the classification accuracies on the test data appeared

304 to peak, with no obvious evidence of excessive overfitting during the training process. The
305 model architecture with the highest classification accuracy on the Exploratory dataset was
306 trained, validated, and tested independently on the Confirmatory dataset. This means that
307 the model that was used to analyze the Confirmatory dataset was not trained on the
308 Exploratory dataset. For all of the analyses that excluded one or more components of the
309 eye movement data (e.g., XYØ, XØP, ØYP, and so on), new models were trained for each
310 data subset (i.e., data subset analyses did not use the model that had already been trained
311 on the full XYP dataset). The model architectures used for the timeline and plot image
312 datasets are shown in Figure 3, with some additional details on the architecture
313 hyperparameters in the figure caption.

314 **Analysis**

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

322 To determine the independent contributions of the three components of the eye
323 movement data, the data subsets were compared within the timeline and plot image data
324 types. If classification accuracies were lower when the data were batched into subsets, the
325 component that was removed was assumed to have some unique contribution that the model
326 was using to inform classification decisions. To determine the uniqueness of the contribution
327 from each component, the accuracies from each subset with one component of the data
328 removed were compared to the accuracies for the full dataset (XYP) using a one-way
329 between-subjects Analysis of Variance (ANOVA). To further evaluate the decodability of

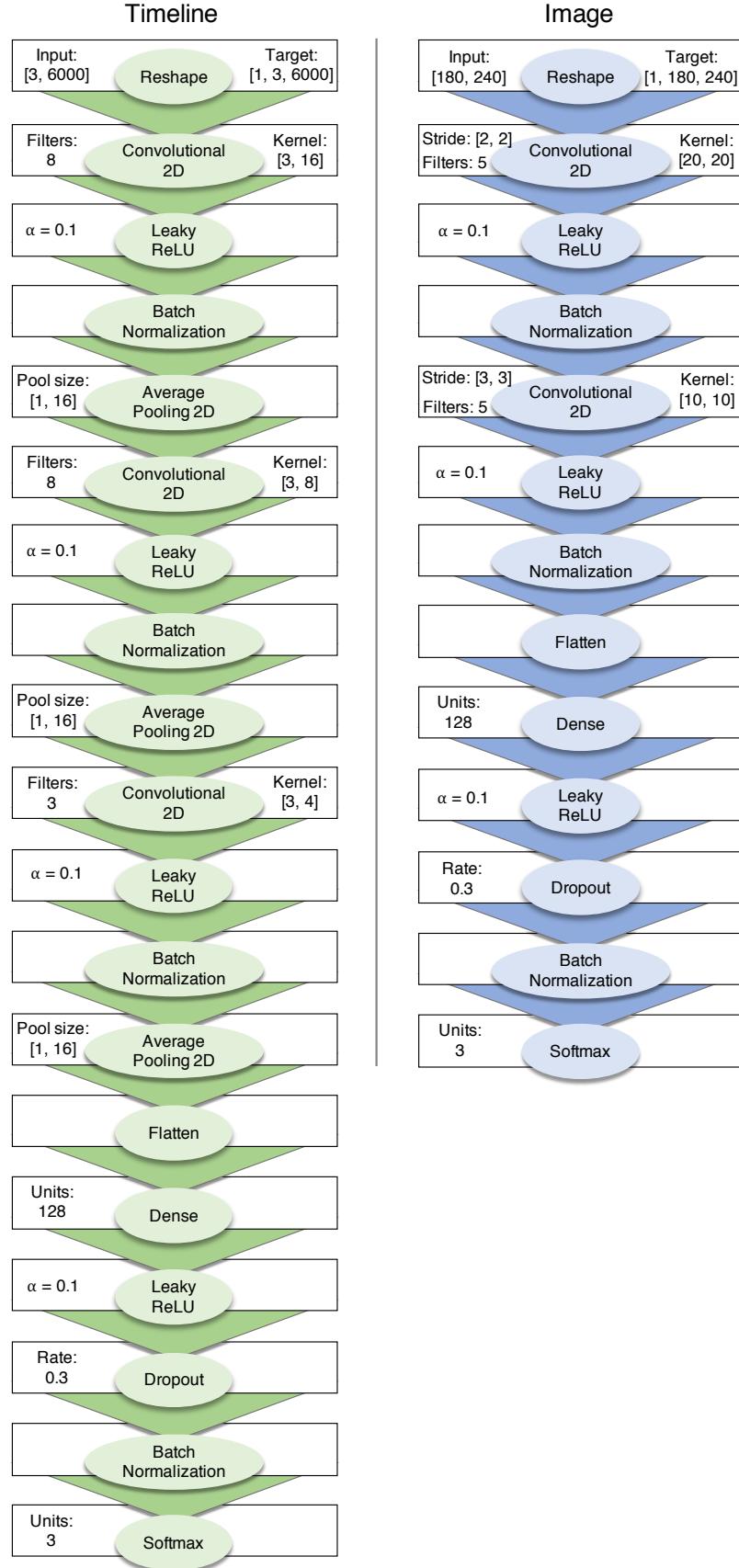


Figure 3. Two different model architectures were used to classify the timeline and image data. Both models were compiled using a categorical crossentropy loss function, and optimized with the Adam algorithm. Optimizer parameters were initial learning rate = 0.005, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 0.1$. The timeline model had 16,946 trainable parameters (29,998 total); the image model had 18,525 trainable parameters (18,827 total).

330 each component independently, the accuracies from each subset containing only one
331 component of the eye movement data were compared within a separate one-way
332 between-subjects ANOVA. All post-hoc comparisons were corrected using Tukey's HSD.

333 **Results**

334 **Timeline Data Classification**

335 **Exploratory.** Classification accuracies for the XYP timeline dataset were well above
336 chance (chance = .33; $M = .526$, $SD = .018$; $t_9 = 34.565$, $p < .001$). Accuracies for
337 classifications of the batched data subsets were all better than chance (see Figure 4). As
338 shown in the confusion matrices displayed in Figure 5, the data subsets with lower overall
339 classification accuracies almost always classified the Memorize condition at or below chance
340 levels of accuracy. Misclassifications of the Memorize condition were split relatively evenly
341 between the Search and Rate conditions.

342 There was a difference in classification accuracy for the XYP dataset and the subsets
343 that had the pupil size, x-coordinate, and y-coordinate data systematically removed ($F_{3,36} =$
344 47.471 , $p < .001$, $\eta^2 = 0.798$). Post-hoc comparisons against the XYP dataset showed that
345 classification accuracies were not affected by the removal of pupil size or y-coordinate data
346 (see Table 2). The null effect present when pupil size was removed suggests that the pupil
347 size data were not contributing unique information that was not otherwise provided by the x-
348 and y-coordinates. A strict significance threshold of $\alpha = .05$ implies the same conclusion for
349 the y-coordinate data, but the relatively low degrees of freedom ($df = 18$) and the borderline
350 observed p -value ($p = .056$) afford the possibility that there exists a small effect. However,
351 classification for the \emptyset YP subset was significantly lower than the XYP dataset, showing that
352 the x-coordinate data were uniquely informative to the classification.

353 There was also a difference in classification accuracies for the X $\emptyset\emptyset$, \emptyset Y \emptyset , and $\emptyset\emptyset$ P
354 subsets ($F_{2,27} = 75.145$, $p < .001$, $\eta^2 = 0.848$). Post-hoc comparisons showed that

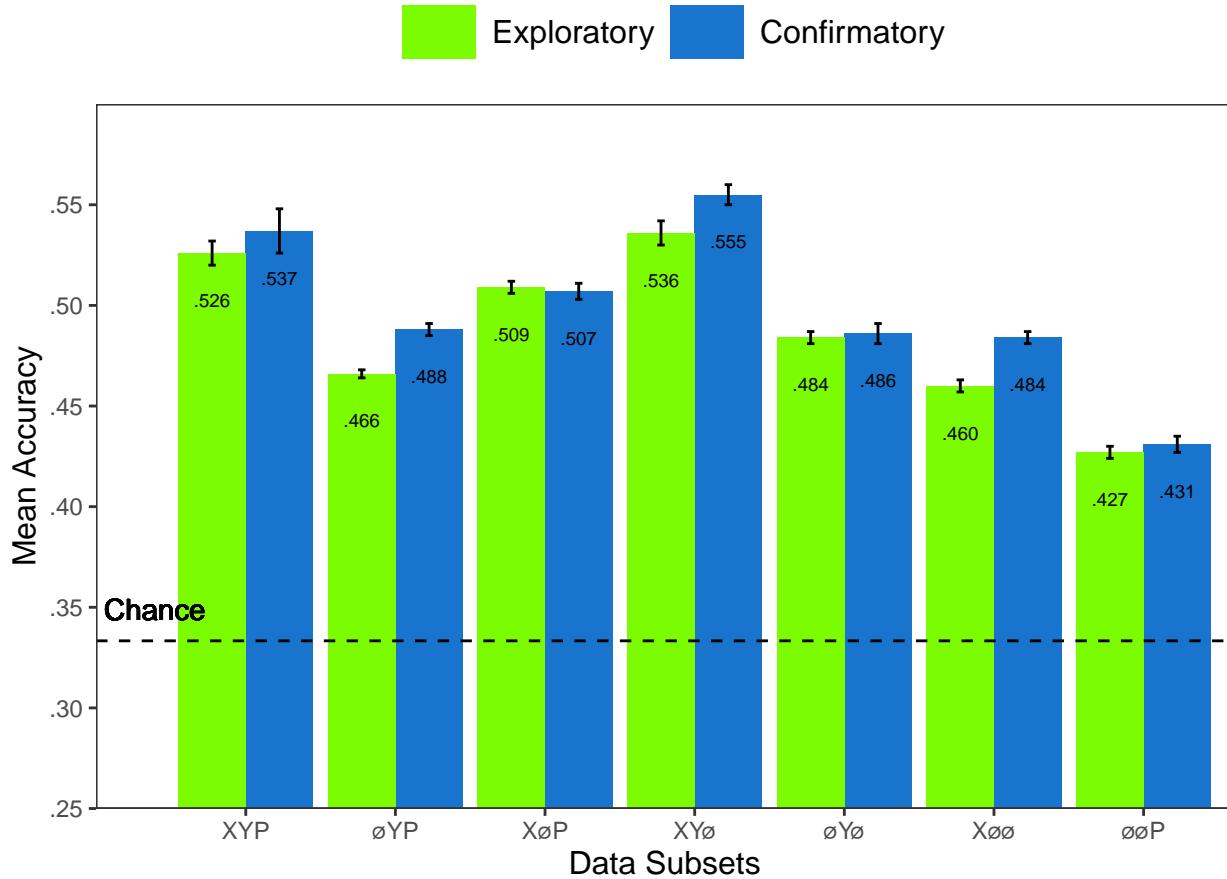


Figure 4. All of the data subsets were decoded at levels better than chance (.33). Each subset is labeled with the mean accuracy. The error bars represent standard errors.

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

355 classification accuracy for the ØØP subset was lower than the XØØ and ØYØ subsets.

356 Classification accuracy for the XØØ subset was higher than the ØYØ subset. Altogether,

357 these findings suggest that pupil size data was the least uniquely informative to classification

358 decisions, while the x-coordinate data was the most uniquely informative.

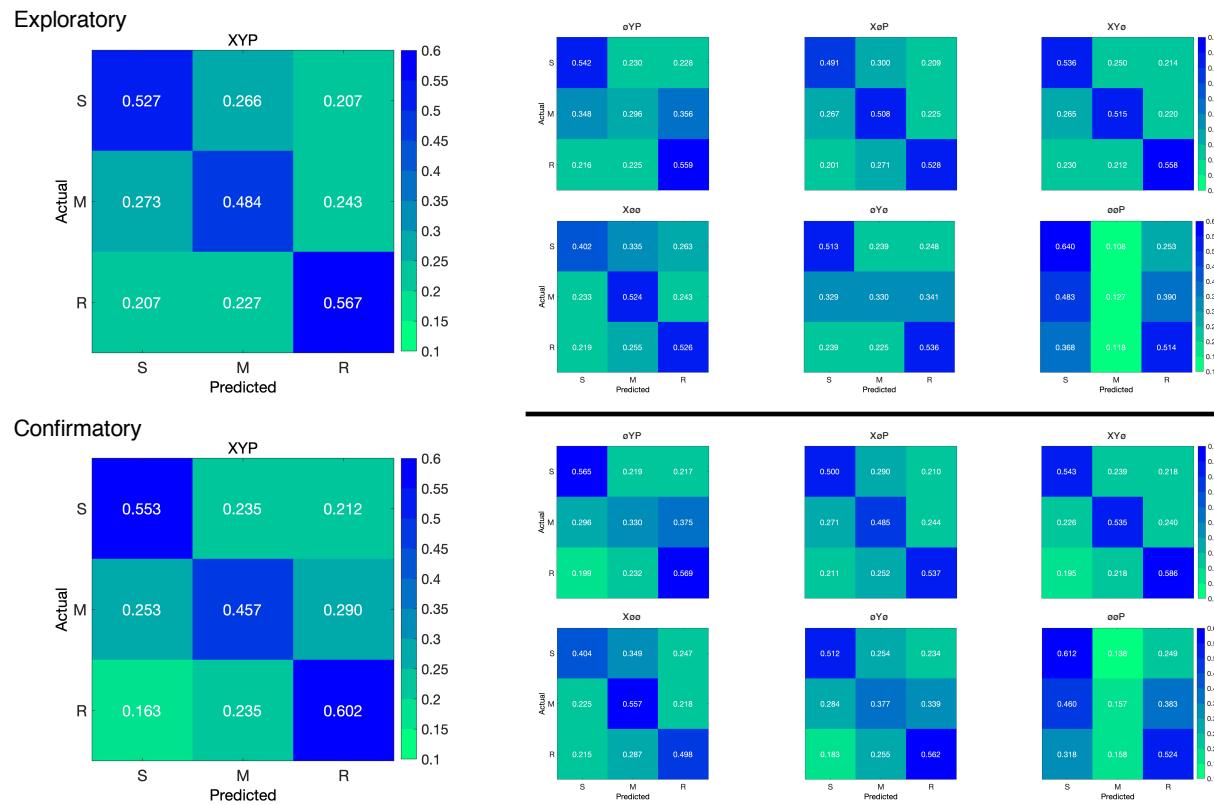


Figure 5. The confusion matrices represent the average classification accuracies for each condition of the timeline data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents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

359 **Confirmatory.** Classification accuracies for the Confirmatory XYP timeline dataset
 360 were well above chance ($M = .537$, $SD = 0.036$, $t_9 = 17.849$, $p < .001$). Classification
 361 accuracies for the data subsets were also better than chance (see Figure 4). Overall, there
 362 was high similarity in the pattern of results for the Exploratory and Confirmatory datasets
 363 (see Figure 4). Furthermore, the general trend showing that pupil size was the least
 364 informative eye tracking data component was replicated in the Confirmatory dataset (see
 365 Table 2). Also in concordance with the Exploratory timeline dataset, the confusion matrices
 366 for these data revealed that the Memorize task was mis-classified more often than the Search
 367 and Rate tasks (see Figure 5).

368 To test the stability of the model architecture, classification accuracies for the XYP
 369 Exploratory and Confirmatory timeline datasets were compared. The Shapiro-Wilk test for
 370 normality indicated that the Exploratory ($W = 0.937$, $p = .524$) and Confirmatory ($W =$

371 0.884, $p = .145$) datasets were normally distributed, but Levene's test indicated that the
372 variances were not equal, $F_{1,18} = 8.783$, $p = .008$. Welch's unequal variances t -test did not
373 show a difference between the two datasets, $t_{13.045} = 0.907$, $p = .381$, Cohen's $d = 0.406$.
374 These findings indicate that the deep learning model decoded the Exploratory and
375 Confirmatory timeline datasets equally well, but the Confirmatory dataset classifications
376 were less consistent across training/test iterations (as indicated by the increase in standard
377 deviation).

378 Plot Image Classification

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

387 There was a difference in classification accuracy between the XYP dataset and the data
388 subsets ($F_{4,45} = 7.093$, $p < .001$, $\eta^2 = .387$). Post-hoc comparisons showed that compared to
389 the XYP dataset, there was no effect of removing pupil size or the x-coordinates, but
390 classification accuracy was worse when the y-coordinates were removed (see Table 3).

391 There was also a difference in classification accuracies between the $X\emptyset\emptyset$, $\emptyset Y\emptyset$, and
392 $\emptyset\emptyset P$ subsets (Levene's test: $F_{2,27} = 3.815$, $p = .035$; Welch correction for lack of
393 homogeneity of variances: $F_{2,17.993} = 228.137$, $p < .001$, $\eta^2 = .899$). Post-hoc comparisons
394 showed that there was no difference in classification accuracies for the $X\emptyset\emptyset$ and $\emptyset Y\emptyset$
395 subsets, but classification for the $\emptyset\emptyset P$ subset were less accurate than the $X\emptyset\emptyset$ and $\emptyset Y\emptyset$

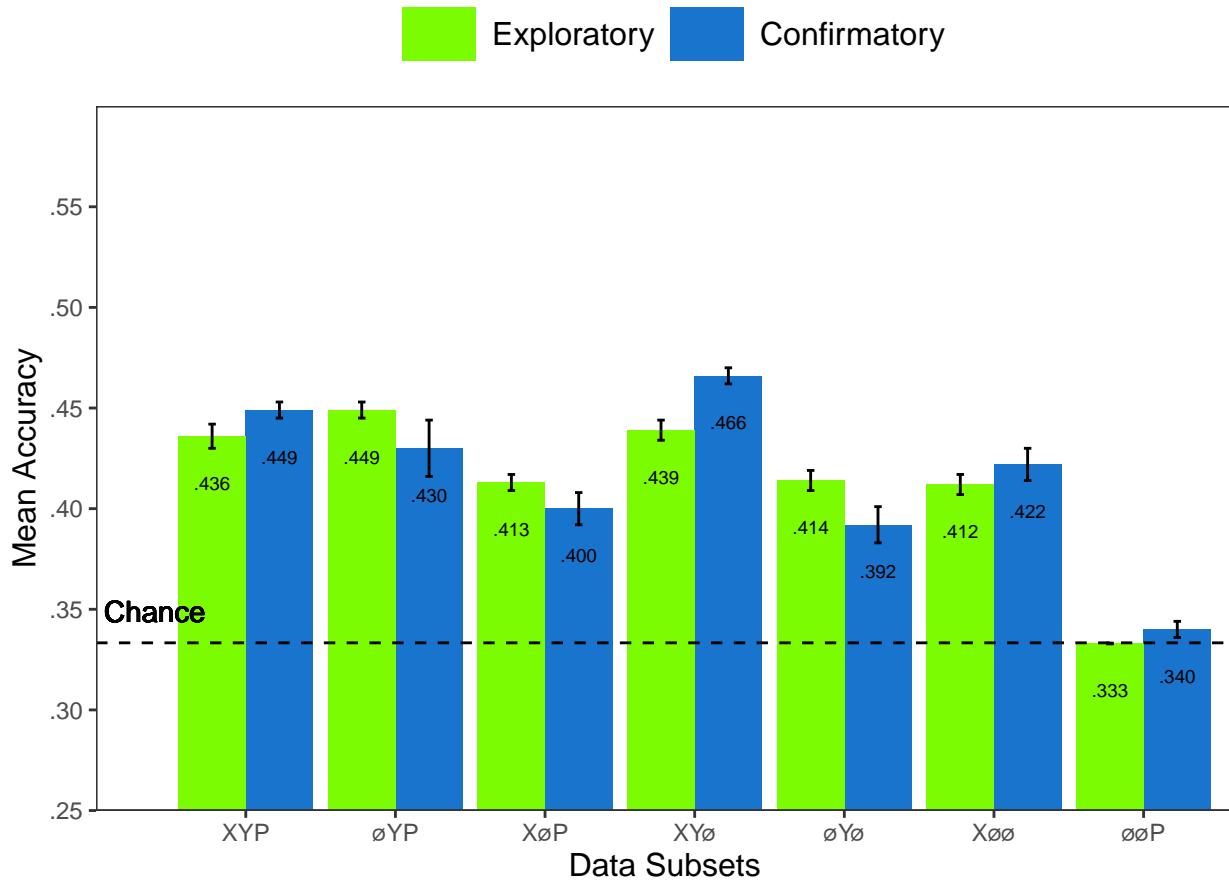


Figure 6. All of the data subsets except for the Exploratory ØØP dataset were decoded at levels better than chance (.33). Each subset is labeled with the mean accuracy. The error bars represent standard errors.

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

396 subsets.

397 **Confirmatory.** Classification accuracies for the XYP confirmatory image dataset
 398 were well above chance ($M = .449$, $SD = 0.012$, $t_9 = 31.061$, $p < .001$), but were less
 399 accurate than the classifications of the confirmatory timeline dataset ($t_{18} = 11.167$, $p <$

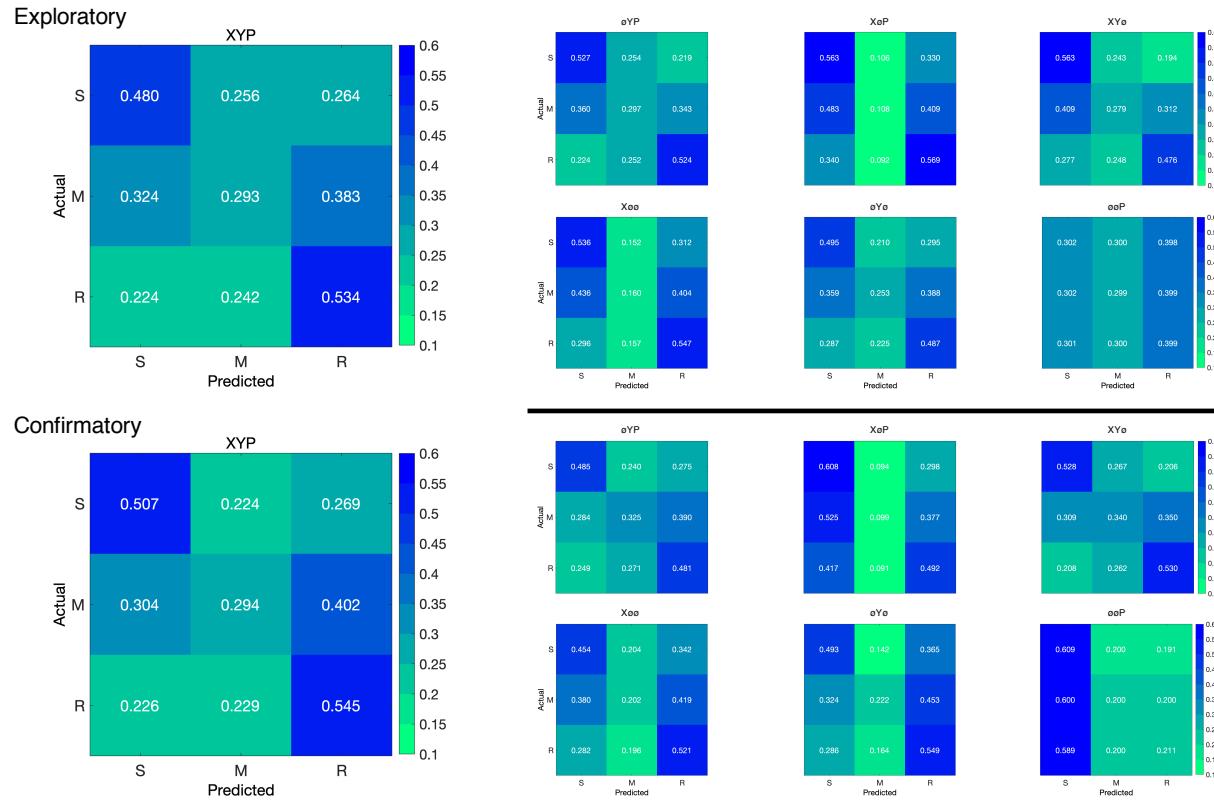


Figure 7. The confusion matrices represent the average classification accuracies for each condition of the image data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents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

.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 mistaken most often, and was relatively equally 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).

To test the stability of the model architecture, the classification accuracies for the XYP Exploratory and Confirmatory plot image datasets were compared. The independent samples *t*-test comparing the classification accuracies for the Exploratory and Confirmatory plot image datasets did not show a significant difference, $t_{18} = 1.777$, $p = .092$, Cohen's $d = 0.795$.

410

Discussion

411 The present study aimed to produce a practical and reliable example of a black box
412 solution to the inverse Yarbus problem. To implement this solution, we classified raw
413 timeline and minimally processed plot image data using a CNN model architecture. To our
414 knowledge, this study was the first to provide a solution to determining task from eye
415 movement data using each of the following: (1) Non-aggregated eye tracking data (i.e., raw
416 x-coordinates, y-coordinates, pupil size), (2) timeline and image data formats (see Figure 2),
417 and (3) a black box CNN architecture. This study probed the independent contributions of
418 the x-coordinate, y-coordinate, and pupil size components of the eye movement data using a
419 CNN. The CNN was able to decode the timeline and plot image data better than chance,
420 although only the timeline datasets were decoded with accuracies comparable to other
421 state-of-the-art approaches. Datasets with lower classification accuracies were not able to
422 differentiate the cognitive processes underlying the Memorize task from the cognitive
423 processes underlying the Search and Rate tasks. Decoding subsets of the data revealed that
424 pupil size was the least uniquely informative component of the eye movement data. This
425 pattern of findings was consistent between the Exploratory and Confirmatory datasets.

426 Although several aggregate eye movement features have been tested as task predictors,
427 to our knowledge, no other study has assessed the predictive value of the data format (viz.,
428 data in the format of a plot image). Our results suggest that although CNNs are robust
429 image classifiers, eye movement data is decoded in the standard timeline format more
430 effectively than in image format. This may be because the image data format contains less
431 decodable information than the timeline format. Over the span of the trial (six seconds), the
432 eye movements occasionally overlapped. When there was an overlap in the image data
433 format, the more recent data points overwrote the older data points. This resulted in some
434 information loss that did not occur when the data were represented in the raw timeline
435 format. Despite this loss of information, the plot image format was still decoded with better

436 than chance accuracy. To further examine the viability of classifying task from eye
437 movement image datasets, future research might consider representing the data in different
438 forms such as 3-dimensional data formats, or more complex color combinations capable of
439 representing overlapping data points.

440 When considering the superior performance of the timeline data (vs., plot image data),
441 we must also consider the differences in the model architectures. Because the structures of
442 the timeline and plot image data formats were different, the models decoding those data
443 structures also needed to be different. Both model architectures were optimized individually
444 on the Exploratory dataset before being tested on the Confirmatory dataset. For both
445 timeline and plot image formats, there was good replicability between the Exploratory and
446 Confirmatory datasets, demonstrating that these architectures performed similarly from
447 experiment to experiment. An appropriately tuned CNN should be capable of learning any
448 arbitrary function, but given that the upper bound for decodability of these datasets is
449 unknown, there is the possibility that a model architecture exists that is capable of classifying
450 the plot image data format more accurately than the model used to classify the timeline
451 data. Despite this possibility, the convergence of these findings with other studies (see Table
452 1) suggests that the results of this study are approaching a ceiling for the potential to solve
453 the inverse Yarbus problem with eye movement data. We attempted to replicate some of
454 those other studies' methods on our own dataset, but were only able to do so with the
455 methods of Coco and Keller (2014), due to lack of publicly available code or incompatibility
456 with our data; for Coco and Keller's methods, we did not achieve better-than-chance
457 classification in our data. **We believe that the below chance outcome for this replication**
458 **analysis is likely attributable to Coco and Keller's focus on differentiating the eye movements**
459 **for separate task sets based on the assumed underlying mental operations rather than relying**
460 **on distinct features in the data or a complex model architecture.** Although the true capacity
461 to predict **task** from eye movement data is unknown, standardizing datasets in the future
462 could provide a point for comparison that can more effectively indicate which methods are

463 most effective at solving the inverse Yarbus problem. As a gesture towards this goal, we have
464 made the data and code from the present study publicly available at: <https://osf.io/dyq3t>.

465 In the current study, the Memorize condition was classified less accurately than the
466 Search and Rate conditions, especially for the datasets with lower overall accuracy. This
467 suggests that the eye movements associated with the Memorize task were potentially lacking
468 unique or informative features to decode. This means that eye movements associated with
469 the Memorize condition were interpreted as noise, or were sharing features of underlying
470 cognitive processes that were represented in the eye movements associated with the Search
471 and Rate tasks. Previous research (e.g., Król & Król, 2018) has attributed the inability to
472 differentiate one condition from the others to the overlapping of sub-features in the eye
473 movements between two tasks that are too subtle to be represented in the eye movement
474 data.

475 To more clearly understand how the different tasks influenced the decodability of the
476 eye movement data, additional analyses were conducted on the Exploratory and
477 Confirmatory timeline datasets (see Appendix). For the main supplementary analysis, the
478 data subsets were re-submitted to the CNN and re-classified as 2-category task sets. In
479 addition to the main supplementary analysis, the results from the primary analysis were
480 re-calculated from 3-category task sets to 2-category task sets. In the primary analyses, the
481 Memorize condition was predicted with the lowest accuracy, but mis-classifications of the
482 Search and Rate trials were most often categorized as Memorize. As a whole, this pattern of
483 results and the main supplementary analysis indicated a general bias for uncertain trials to
484 be categorized as Memorize. As expected, the main supplementary analysis also showed that
485 the 2-category task set that included only Search and Rate had higher accuracies than both
486 of the 2-category task sets that included the Memorize condition. The re-calculation analysis
487 generally replicated the pattern of results seen in the main supplementary analysis but with
488 larger variance, suggesting that including lower-accuracy trial types during model training

489 can decrease the consistency of classifier performance. Overall, the findings from this
490 supplemental analysis show that conclusions drawn from comparisons between approaches
491 that do not use the same task sets, or the same number of tasks, could be potentially
492 uninterpretable because the features underlying the task categories are interpreted differently
493 by the neural network algorithm.

494 When determining the unique contributions of the eye movement features used in
495 this study (x-coordinates, y-coordinates, pupil size), the pupil size data was consistently the
496 least uniquely informative. When pupil size was removed from the Exploratory and
497 Confirmatory timeline and plot image datasets, classification accuracy remained stable (vs.,
498 XYP dataset). Furthermore, classification accuracy of the $\emptyset\emptyset P$ subset was the lowest of all
499 of the data subsets, and in one instance, was no better than chance. Although these findings
500 indicate that, in this case, pupil size was a relatively uninformative component of the eye
501 movement data, previous research has associated changes in pupil size as indicators of
502 working memory load (Kahneman & Beatty, 1966; Karatekin, Couperus, & Marcus, 2004),
503 arousal (Wang et al., 2018), and cognitive effort (Porter, Troscianko, & Gilchrist, 2007). The
504 results of the current study indicate that the changes in pupil size associated with these
505 underlying processes were not useful in delineating the tasks being classified (i.e., Search,
506 Memorize, Rate), potentially because these tasks did not evoke a reliable pattern of changes
507 in pupil size. Additionally, properties of the stimuli known to influence pupil size, such as
508 luminance and contrast, were not controlled in these datasets. Given that stimuli were
509 randomly assigned, there is the possibility that uncontrolled stimulus properties known to
510 affect pupil size impeded the CNN's capacity to detect patterns in the pupil size data.

511 The findings from the current study support the notion that black box CNNs are a
512 viable approach to determining task from eye movement data. In a recent review, Lukander,
513 Toivanen, and Puolamäki (2017) expressed concern regarding the lack of generalizability of
514 black box approaches when decoding eye movement data. Overall, the current study showed

515 a consistent pattern of results for the XYP timeline and image datasets, but some minor
516 inconsistencies in the pattern of results for the x- and y- coordinate subset comparisons.
517 These inconsistencies may be a product of overlap in the cognitive processes underlying the
518 three tasks. When the data are batched into subsets, at least one dimension (i.e.,
519 x-coordinates, y-coordinates, or pupil size) is removed, leading to a potential loss of
520 information. When the data provide fewer meaningful distinctions, finer-grained inferences
521 are necessary for the tasks to be distinguishable. As shown by Coco and Keller (2014), eye
522 movement data can be more effectively decoded when the cognitive processes underlying the
523 tasks are explicitly differentiable. While the cognitive processes distinguishing memorizing,
524 searching, or rating an image are intuitively different, the eye movements elicited from these
525 cognitive processes are not easily differentiated. To correct for potential mismatches between
526 the distinctive task-diagnostic features in the data and the level of distinctiveness required to
527 classify the tasks, future research could more definitively conceptualize the cognitive
528 processes underlying the task-at-hand.

529 Classifying **task** from eye movement data is often carried out in an effort to advance
530 technology to improve educational outcomes, strengthen the independence of physically and
531 mentally handicapped individuals, or improve HCI's (Koochaki & Najafizadeh, 2018). Given
532 the previous questions raised regarding the reliability and generalizability of black-box CNN
533 classification, the current study first tested models on an exploratory dataset, then confirmed
534 the outcome using a second independent dataset. Overall, the findings of this study indicate
535 that this black-box approach is capable of producing a stable and generalizable outcome.
536 Additionally, the supplementary analyses showed that different task sets, or a different
537 number of tasks, could lead the algorithm to interpret features differently, which should be
538 taken into account when comparing task classification approaches. Future studies that
539 incorporate features from the stimulus might have the potential to surpass current
540 state-of-the-art classification. According to Bulling, Weichel, and Gellersen (2013),
541 incorporating stimulus feature information into the dataset may improve accuracy relative to

542 decoding gaze location data and pupil size. Alternatively, Borji and Itti (2014) suggested
543 that accounting for salient features in the stimulus might leave little to no room for
544 theoretically defined classifiers to consider mental state. Future research should examine the
545 potential for the inclusion of stimulus feature information in addition to the eye movement
546 data to boost black-box CNN classification accuracy of image data beyond that of timeline
547 data.

References

- Bashivan, P., Rish, I., Yeasin, M., & Codella, N. (2016). Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks. *arXiv:1511.06448 [Cs]*. Retrieved from <http://arxiv.org/abs/1511.06448>
- Boisvert, J. F. G., & Bruce, N. D. B. (2016). Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing*, 207, 653–668. <https://doi.org/10.1016/j.neucom.2016.05.047>
- Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision*, 14(3), 29–29. <https://doi.org/10.1167/14.3.29>
- Bulling, A., Weichel, C., & Gellersen, H. (2013). EyeContext: Recognition of high-level contextual cues from human visual behaviour. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (p. 305). Paris, France: ACM Press. <https://doi.org/10.1145/2470654.2470697>
- Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3), 6–6. <https://doi.org/10.1167/9.3.6>
- Coco, M. I., & Keller, F. (2014). Classification of visual and linguistic tasks using eye-movement features. *Journal of Vision*, 14(3), 11–11. <https://doi.org/10.1167/14.3.11>
- DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, 17(6-7), 790–811. <https://doi.org/10.1080/13506280902793843>
- Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to

- 571 predict observers' task from eye movement patterns. *Vision Res*, 62, 1–8.
- 572 <https://doi.org/10.1016/j.visres.2012.03.019>
- 573 Haji-Abolhassani, A., & Clark, J. J. (2014). An inverse Yarbus process: Predicting
574 observers' task from eye movement patterns. *Vision Research*, 103, 127–142.
575 <https://doi.org/10.1016/j.visres.2014.08.014>
- 576 Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013).
577 Predicting Cognitive State from Eye Movements. *PLoS ONE*, 8(5), e64937.
578 <https://doi.org/10.1371/journal.pone.0064937>
- 579 Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*,
580 154(3756), 1583–1585.
- 581 Kanan, C., Ray, N. A., Bseiso, D. N. F., Hsiao, J. H., & Cottrell, G. W. (2014).
582 Predicting an observer's task using multi-fixation pattern analysis. In *Proceedings*
583 *of the Symposium on Eye Tracking Research and Applications - ETRA '14* (pp.
584 287–290). Safety Harbor, Florida: ACM Press.
585 <https://doi.org/10.1145/2578153.2578208>
- 586 Karatekin, C., Couperus, J. W., & Marcus, D. J. (2004). Attention allocation in the
587 dual-task paradigm as measured through behavioral and psychophysiological
588 responses. *Psychophysiology*, 41(2), 175–185.
589 <https://doi.org/10.1111/j.1469-8986.2004.00147.x>
- 590 Koochaki, F., & Najafizadeh, L. (2018). Predicting Intention Through Eye Gaze
591 Patterns. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*
592 (pp. 1–4). <https://doi.org/10.1109/BIOCAS.2018.8584665>
- 593 Król, M. E., & Król, M. (2018). The right look for the job: Decoding cognitive
594 processes involved in the task from spatial eye-movement patterns. *Psychological*

- Research. <https://doi.org/10.1007/s00426-018-0996-5>
- Kuntzelman, K. M., Williams, J. M., Lim, P. C., Samal, A., Rao, P. K., & Johnson, M. R. (2021). Deep-Learning-Based Multivariate Pattern Analysis (dMVPA): A Tutorial and a Toolbox. *Frontiers in Human Neuroscience*, 15, 638052. <https://doi.org/10.3389/fnhum.2021.638052>
- Lukander, K., Toivanen, M., & Puolamäki, K. (2017). Inferring Intent and Action from Gaze in Naturalistic Behavior: A Review. *International Journal of Mobile Human Computer Interaction*, 9(4), 41–57.
- <https://doi.org/10.4018/IJMHCI.2017100104>
- MacInnes, W., Joseph, Hunt, A. R., Clarke, A. D. F., & Dodd, M. D. (2018). A Generative Model of Cognitive State from Task and Eye Movements. *Cognitive Computation*, 10(5), 703–717. <https://doi.org/10.1007/s12559-018-9558-9>
- Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision*, 11(8), 17–17. <https://doi.org/10.1167/11.8.17>
- Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: Insights from pupillometry. *Quarterly Journal of Experimental Psychology* (2006), 60(2), 211–229. <https://doi.org/10.1080/17470210600673818>
- Seeliger, K., Fritzsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., & van Gerven, M. A. J. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, 180, 253–266. <https://doi.org/10.1016/j.neuroimage.2017.07.018>
- Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus, Eye Movements, and Vision. *I-Perception*, 1(1), 7–27.

619 <https://doi.org/10.1068/i0382>

620 Wang, C.-A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C., & Munoz, D. P.

621 (2018). Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an

622 Emotional Face Task. *Frontiers in Neurology*, 9, 1029.

623 <https://doi.org/10.3389/fneur.2018.01029>

624 Yarbus, A. (1967). Eye Movements and Vision.

625 [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).

626 Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2019). Comparing the Interpretability

627 of Deep Networks via Network Dissection. In W. Samek, G. Montavon, A.

628 Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting,*

629 *Explaining and Visualizing Deep Learning* (pp. 243–252). Cham: Springer

630 International Publishing. https://doi.org/10.1007/978-3-030-28954-6_12

631

Appendix

632 Additional analyses were conducted in an attempt to clarify the effect of task on
 633 classification accuracy. These supplementary analyses were not seen as central to the current
 634 study, but could prove to be informative to researchers attempting to replicate or extend
 635 these findings in the future. The results from the primary analysis showed that classification
 636 accuracies were the lowest for the Memorize condition. To further understand why
 637 classification accuracy was lower for the Memorize condition than it was for the Search or
 638 Rate condition, the Exploratory and Confirmatory timeline datasets were systematically
 639 batched into subsets with the Search (S), Memorize (M), or Rate (R) condition removed (i.e.,
 640 \emptyset MR, S \emptyset R, SM \emptyset), and then run through the CNN classifier using the same methods as the
 641 primary analysis, but with only two classes.

642 All of the data subsets analyzed in this supplementary analysis were decoded with
 643 better than chance accuracy (see Figure 8a). The same pattern of results was observed in
 644 both the Exploratory and Confirmatory datasets. When the Memorize condition was
 645 removed, classification accuracy improved (see Table 4, Figure 8a). When the Rate condition
 646 was removed, classification was the worst. When the Memorize condition was included (i.e.,
 647 SM \emptyset and \emptyset MR), mis-classifications were biased toward Memorize, and the Memorize
 648 condition was more accurately predicted than the Search and Rate conditions (see Figure 9).

Table 4
Supplementary Subset Comparisons

Comparison	Exploratory		Confirmatory	
	t	p	t	p
\emptyset MR vs. S \emptyset R	3.248	.008	3.094	.012
\emptyset MR vs. SM \emptyset	2.875	.021	2.923	.018
S \emptyset R vs. SM \emptyset	6.123	< .001	6.017	< .001

649 The accuracies for all of the data subsets observed in the supplementary analysis were
 650 higher than the accuracies observed in the main analysis. Although there is a clear difference
 651 in accuracy, the primary analysis was classifying three categories (chance = .33) and the

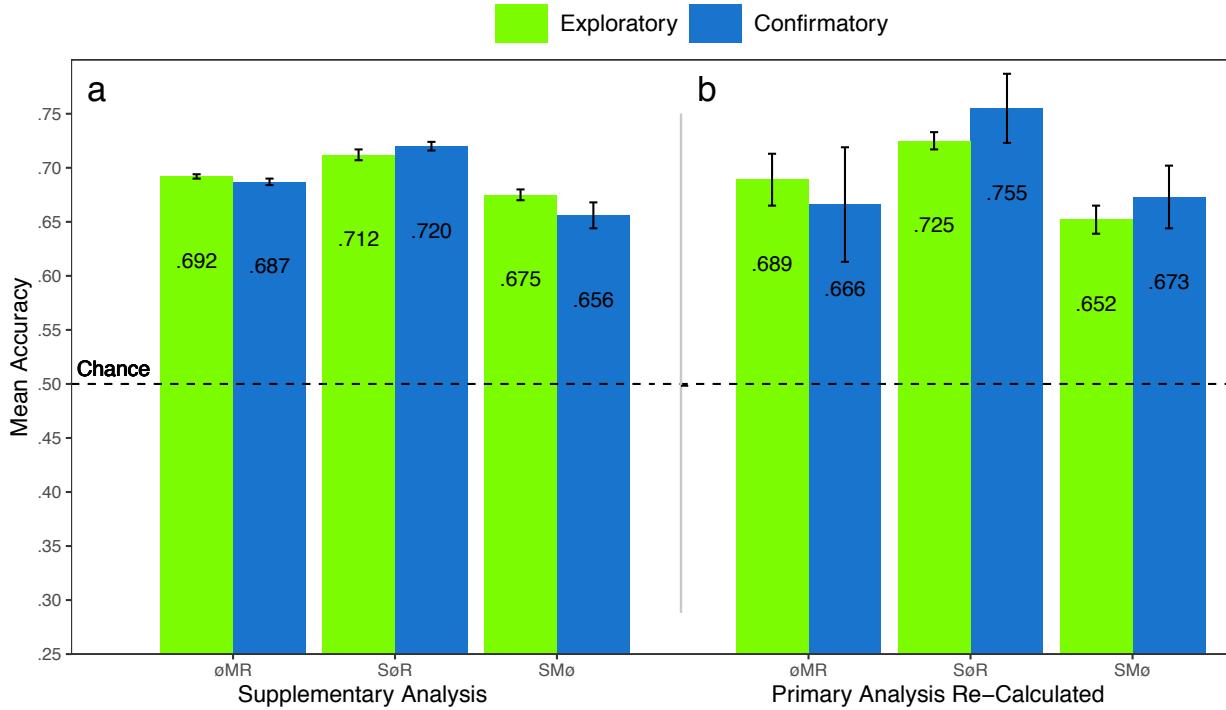


Figure 8. The graph represents the average accuracy reported for each subset of the Exploratory and Confirmatory timeline data for (a) the supplementary analysis, and the (b) re-calculated accuracies from the primary analysis. All of the data subsets were decoded at levels better than chance (.50). The error bars represent standard errors.

652 supplementary analysis was classifying two categories (chance = .50). Because the baseline
 653 chance performance was different for the primary and supplemental analyses, any conclusions
 654 drawn from a comparison of the results of analyses could be misleading. For this reason, we
 655 revisited the results from the primary analysis and re-calculated the predictions to be
 656 equivalent to a 50% chance threshold. Because the cross-validation scheme implemented by
 657 the DeLINEATE toolbox (<http://delineate.it>) (Kuntzelman et al., 2021) guaranteed an equal
 658 number of trials in the test set were assigned to each condition for each dataset, we were able
 659 to re-calculate 2-category predictions from the 3-category predictions presented in the
 660 confusion matrices from the primary analysis (see Figure 5). The predictions were
 661 re-calculated using the following formula: $\text{Prediction}_{(A,A,A \otimes C)} = \text{Prediction}_{(A,A,ABC)} /$
 662 $(\text{Prediction}_{(A,A,ABC)} + \text{Prediction}_{(A,C,ABC)})$. For example, accuracy for the Search
 663 classification for S \otimes R would be calculated with the following: $\text{Prediction}_{(S,S,S \otimes R)} =$
 664 $\text{Prediction}_{(S,S,SMR)} / (\text{Prediction}_{(S,S,SMR)} + \text{Prediction}_{(S,R,SMR)})$, where $\text{Prediction}_{(S,R,S \otimes R)}$ is

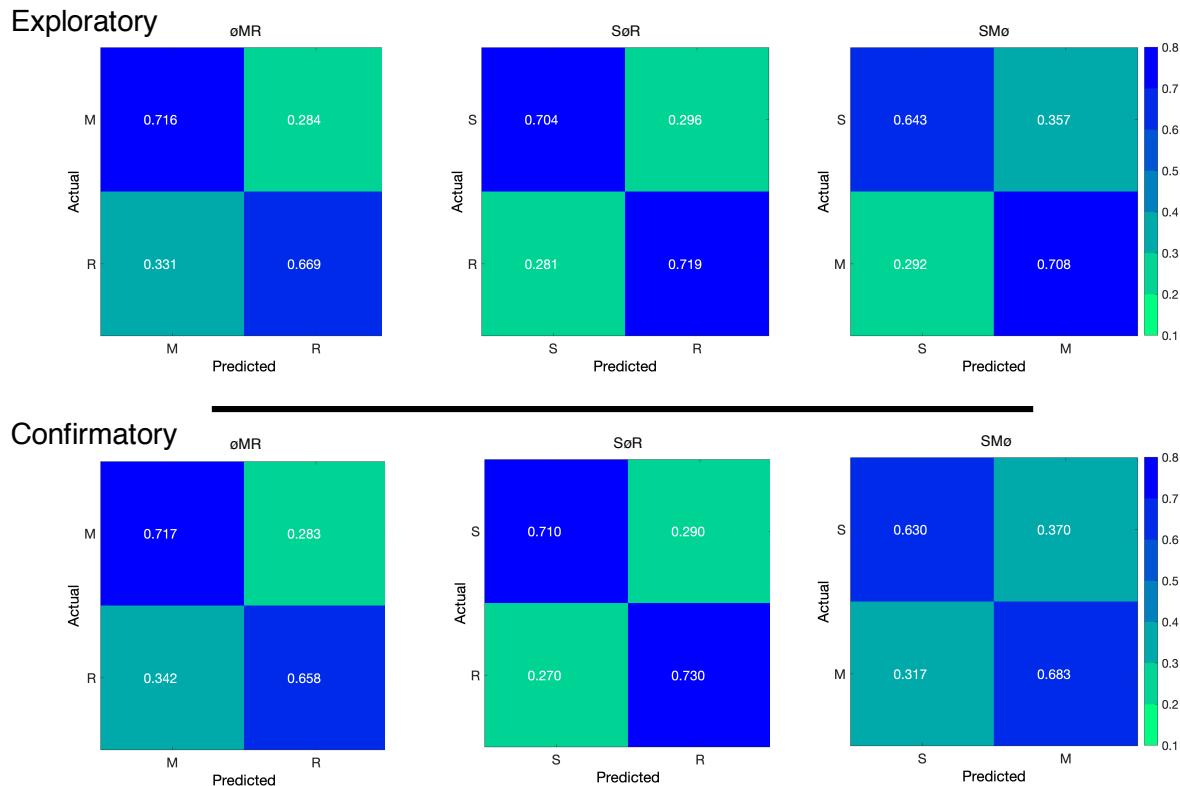


Figure 9. The confusion matrices represent the average classification accuracies for each condition of the timeline data (S = Search, M = Memorize, R = Rate). The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents the condition that was predicted by the model.

665 the ratio of Search trials that were misclassified as Rate.

666 The results for the re-calculated predictions followed a pattern similar to the main
 667 supplementary analysis (see Figure 8b). Looking back at the primary analysis, the
 668 3-category classifications predicted the Memorize conditions with the lowest accuracy (c.f.,
 669 Search and Rate conditions), and mis-classifications of the Search and Rate conditions were
 670 most often categorized as Memorize (see Figure 5). Because the Memorize condition was
 671 mis-classified more often than the other conditions in the primary analysis, the removal of
 672 the third class in the re-calculated SMø and øMR subsets resulted in a disproportionate
 673 amount of mis-classified Memorize trials being removed from those data subsets, somewhat
 674 eliminating the tendency to mis-classify Search and Rate trials as Memorize (see Figure 10).
 675 Nevertheless, the re-calculated SMø and øMR subsets were classified less accurately than
 676 SøR, just as in the main supplementary analysis.

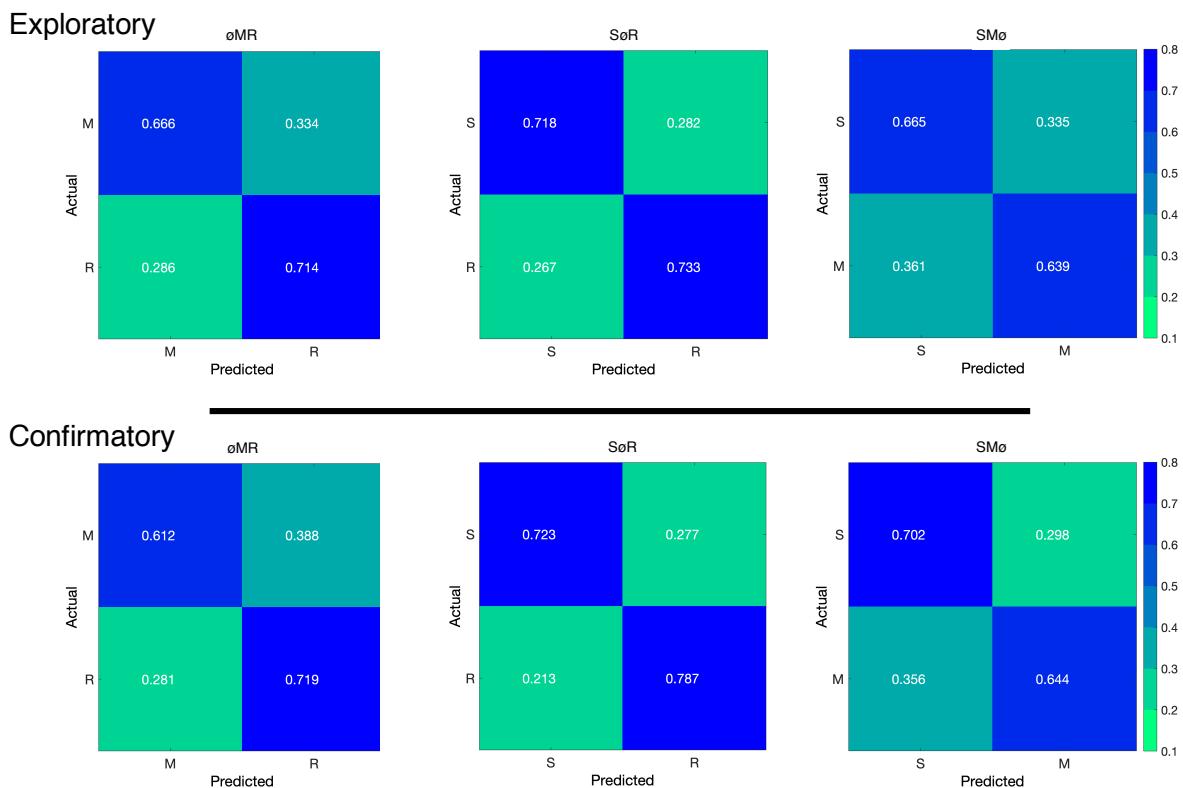


Figure 10. The confusion matrices represent a re-calculation of the classification accuracies for each category from the primary analysis. This re-calculation is meant to make the accuracies presented in the primary analysis (chance = .33) equivalent to the classification accuracies presented in the supplementary analysis (chance = .50).