

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

<sup>3</sup> Zachary J. Cole<sup>1</sup>, Karl M. Kuntzelman<sup>1</sup>, Michael D. Dodd<sup>1</sup>, &  
<sup>4</sup> Matthew R. Johnson<sup>1</sup>

<sup>5</sup> <sup>1</sup> University of Nebraska-Lincoln

<sup>6</sup> Abstract

Previous attempts to classify task from eye movement data have relied on model architectures designed to emulate theoretically defined cognitive processes, and/or data that has been processed into aggregate (e.g., fixations, saccades) or statistical (e.g., fixation density) features. *Black box* convolutional neural networks (CNNs) are a model architecture capable of identifying relevant features in raw and minimally processed data and images, but difficulty interpreting the mechanisms underlying these model architectures have contributed to challenges in generalizing lab-trained CNNs to applied contexts. In the current study, a CNN classifier was used to classify task from two eye movement datasets (Exploratory and Confirmatory) in which participants searched, memorized, or rated indoor and outdoor scene images. The Exploratory dataset was used to tune the hyperparameters of the model, and the resulting model architecture was re-trained, validated, and tested on the Confirmatory dataset. The data were formatted into raw timeline data (i.e., x-coordinate, y-coordinate, pupil size) and minimally processed images.  
<sup>7</sup> To further understand the relative informational value of the raw components of the eye movement data, the timeline and image datasets were broken down into subsets with one or more of the components of the data systematically removed. Average classification accuracies were compared between datasets and subsets. Classification of the timeline data consistently outperformed the image data. The Memorize condition was most often confused with the Search and Rate conditions. Pupil size was the least uniquely informative eye movement component when compared with the x- and y-coordinates. The overall pattern of results for the Exploratory dataset was replicated in the Confirmatory dataset, indicating the potential generalizability of this black box approach.

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

Word count: 6222

8

9       The association between eye movements and mental activity is a fundamental topic of  
10 interest in attention research that has provided a foundation for developing a wide range  
11 of human assistive technologies. Foundational work by Yarbus (1967) showed that eye  
12 movement patterns appear to differ qualitatively depending on the task-at-hand (for a review  
13 of this work, see Tatler, Wade, Kwan, Findlay, & Velichkovsky, 2010). A replication of this  
14 work by DeAngelus and Pelz (2009) shows that the differences in eye movements between  
15 tasks can be quantified, and appear to be somewhat generalizable. Technological advances  
16 and improvements in computing power have allowed researchers to make inferences regarding  
17 the mental state underlying eye movement data, also known as the “inverse Yarbus process”  
18 (Haji-Abolhassani & Clark, 2014). Current state-of-the-art machine learning and neural  
19 network algorithms are capable of identifying diagnostic patterns for the purpose of decoding  
20 a variety of data types, but the inner workings of the resulting model solutions are difficult  
21 or impossible to interpret. Algorithms that provide uninterpretable solutions are referred  
22 to as *black box* models. Dissections of black box models have been largely uninformative  
23 (Zhou, Bau, Oliva, & Torralba, 2019), limiting the potential for researchers to apply the  
24 mechanisms underlying successful classification of the data. Still, black box models provide a  
25 convenient solution for technological applications such as human-computer interfaces (HCI;  
26 for a review, see Lukander, Toivanen, & Puolamäki, 2017). While the internal operations of  
27 the model solutions used for HCI applications do not necessarily need to be interpretable  
28 to serve their purpose, Lukander et al. (2017) pointed out that “the black box nature of  
29 the resulting solution impedes generalizability, and makes applying methods across real life  
30 conditions more difficult” (p. 44). To ground these solutions, researchers guide decoding  
31 efforts by using eye movement data and/or models with built-in theoretical assumptions.  
32 For instance, eye movement data is processed into meaningful aggregate properties such as  
33 fixations or saccades, or statistical features such as fixation density, and the models used  
34 to decode these data are structured based on the current understanding of relevant cognitive  
35 or neurobiological processes (e.g., MacInnes, Hunt, Clarke, & Dodd, 2018).

36       At this point, there is no clear evidence to support the notion that the standard  
37 theoretically grounded inferences actually enhance or clarify black box solutions beyond  
38 what could be inferred from an unconstrained model. Consider the case of Greene, Liu, and  
39 Wolfe (2012), who were not to classify the task from commonly used aggregate eye movement  
40 features (i.e., number of fixations, mean fixation duration, mean saccade amplitude, percent  
41 of image covered by fixations) using correlations, a linear discriminant model, and a support  
42 vector machine (see Table 1). This led Greene and colleagues to question the robustness  
43 of Yarbus’s (1967) findings, inspiring a slew of responses that successfully decoded the  
44 same dataset by aggregating the eye movements into different feature sets or implementing  
45 different model architectures (see Table 1; i.e., Haji-Abolhassani & Clark, 2014; Borji &  
46 Itti, 2014; Kanan, Ray, Bseiso, Hsiao, & Cottrell, 2014). The subsequent re-analyses of

---

This work was supported by NSF/EPSCoR grant #1632849 and NIH grant GM130461 awarded to MRJ and colleagues.

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

47 these data support Yarbus (1967) and the notion that mental state can be decoded from eye  
48 movement data using a variety of combinations of data features and model architectures.  
49 Despite using theoretically informed models to classify the data, these re-analyses do not  
50 explain the outcome of Greene et al. (2012) any more definitively than a black box approach.

51 Eye movements can only delineate tasks to the extent that the cognitive processes  
52 underlying the tasks can be differentiated (Król & Król, 2018). Every task is associated  
53 with a unique set of cognitive processes (Coco & Keller, 2014; Król & Król, 2018), but in  
54 some cases, the cognitive processes for different tasks may produce indistinguishable eye  
55 movement patterns. To differentiate the cognitive processes underlying task-evoked eye  
56 movements, some studies have chosen to classify tasks that rely on stimuli that prompt  
57 easily distinguishable eye movements, such as reading text and searching pictures (e.g.,  
58 Henderson, Shinkareva, Wang, Luke, & Olejarczyk, 2013). The eye movements elicited by  
59 salient stimulus features facilitate task classifications, but because these eye movements  
60 are the consequence of a feature, or features, inherent to the stimulus rather than the task,  
61 it is unclear if these classifications are attributable to the stimulus or a complex mental  
62 state (e.g., Henderson et al., 2013; Boisvert & Bruce, 2016). Additionally, the distinct  
63 nature of exogenously elicited eye movements prompts decoding algorithms to prioritize  
64 these bottom-up patterns in the data over higher-level top-down effects (Borji & Itti, 2014).  
65 This means that these models are identifying the type of information that is being processed,  
66 but are not necessarily reflecting the mental state of the individual observing the stimulus.  
67 Eye movements that are the product of bottom-up processes have been reliably decoded,  
68 which is relevant for some HCI applications, but does not fit the nature of the inverse Yarbus  
69 problem which is concerned with decoding high-level abstract mental operations that are  
70 not dependent on particular stimuli.

71 Currently, the capacity to solve the problem of classifying cognitive tasks from eye  
72 movement data has not been clearly established. Prior evidence has shown that the task-  
73 at-hand is capable of producing distinguishable eye movement features such as the total  
74 scan path length, total number of fixations, and the amount of time to the first saccade  
75 (Castelhano, Mack, & Henderson, 2009; DeAngelus & Pelz, 2009). Decoding accuracies  
76 within the context of determining task from eye movements typically range from chance  
77 performance (between 14.29% and 33%) to 59.64% (see Table 1). In one case, Coco and  
78 Keller (2014) categorized the same eye movement features used by Greene et al. (2012) with  
79 respect to the relative contribution of latent visual or linguistic components of three tasks  
80 (visual search, name the picture, name objects in the picture) with 84% accuracy. While  
81 this manipulation is reminiscent of other experiments relying on the bottom-up influence of  
82 words and pictures (e.g., Henderson et al., 2013; Boisvert & Bruce, 2016) the eye movements  
83 in the Coco and Keller (2014) tasks were entirely the product of top-down processes. A  
84 conceptually similar follow-up to this study classified tasks along two spatial and semantic  
85 dimensions, resulting in 51% classification accuracy (chance = 25%; Król & Król, 2018). A  
86 closer look at these results showed that the categories within the semantic dimension were  
87 consistently mixed up, suggesting that this level of distinction may require a richer dataset,  
88 or a more powerful decoding algorithm. Altogether, there is no measureable index of relative  
89 top-down influence, but this body of literature suggests that the locus of eye movement  
90 control is an important factor to consider when classifying mental state from eye movement

91 data.

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 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 saccade latency, saccade duration, saccade amplitude, peak saccade velocity, absolute saccade angle, pupil size	MM, LASSO, SVM	84% (33%)
MacInnes et al. (2018)	view, memorize, search, rate		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%)

92 As shown in Table 1, when eye movement data are prepared for classification, fixation  
 93 and saccade statistics are typically aggregated along spatial or temporal dimensions, resulting

in variables such as fixation density or saccade amplitude (Castelhano et al., 2009; MacInnes et al., 2018; Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011). The implementation of these statistical methods is meant to explicitly provide the decoding algorithm with characteristics of the eye movement data that are representative of theoretically relevant cognitive processes. For example, MacInnes et al. (2018) attempted to provide an algorithm with data designed to be representative of inputs to the frontal eye fields. In some instances, such as the case of Król and Król (2018), grounding the data using theoretically driven aggregation methods may require sacrificing granularity in the dataset. This means that aggregating the data has the potential to wash out certain fine-grained distinctions that could otherwise be detected. Data structures of any kind can only be decoded to the extent which the data is capable of representing differences between categories. Given that the cognitive processes underlying distinct tasks are often overlapping (Coco & Keller, 2014), decreasing the granularity of the data may actually limit the potential of the algorithm to make fine grained distinctions between diagnostic components underlying the target task and the other tasks.

The current state of the literature does not provide any firm guidelines for determining what eye movement features are most meaningful, or what model architectures are most suited for determining mental state from eye movements. The examples provided in Table 1 used a variety of eye movement features and model architectures, most of which were effective to a similar extent (with the exception of Greene et al., 2012). A proper comparison of these outcomes is impractical because these datasets vary in levels of chance and data quality. Datasets with more tasks to be classified have lower levels of chance, lowering the threshold denoting successful classification. Additionally, datasets with a lower signal-to-noise ratio will have a lower achievable classification accuracy. For these reasons, outside of re-analyzing the same datasets, there is no concensus on how to establish direct comparisons of these model architectures.

The current study explored pragmatic solutions to the problem of classifying task from eye movement data by submitting unprocessed x-coordinate, y-coordinate, and pupil size data to a convolutional neural network (CNN) model. Instead of transforming the data into theoretically defined units, we allowed the network to learn meaningful patterns in the data. CNNs have a natural propensity to develop low-level feature detectors similar to primary visual cortex (e.g., Seeliger et al., 2018). For this reason, CNNs are commonly implemented for image classification. To test the possibility that the image data are better suited to the CNN classifier, the data were also transformed into raw timeline and simple image representations. To our knowledge, no study has attempted to address the inverse Yarbus problem using any combination of the following methods: (1) Non-aggregated data, (2) image data format, and (3) a CNN architecture. Given that CNN architectures are capable of learning features ingrained in raw data, and are well-suited to decoding multidimensional data that have a distinct spatial or temporal structure, we expected that a non-theoretically-constrained CNN architecture could be capable of decoding data at levels consistent with the current state of the art. Furthermore, we expected that despite evidence that black box approaches to the inverse Yarbus problem can impede generalizability (Lukander et al., 2017), our initial findings would replicate when tested on an entirely separate dataset.

137

## Methods

138

### Participants

139

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

147

### Materials and Procedures

148

Each participant viewed a series of XX indoor and outdoor scene images while carrying out a search, memorization, or rating task. For the search task, participants were instructed to find a “Z” or “N” embedded in the image. If the letter was found, the participants were instructed to press a button, which terminated the trial. For the memorization task, participants were instructed to memorize the image for a test that would take place when the task was completed. For the rating task, participants were asked to think about how they would rate the image on a scale from 1 (very unpleasant) to 7 (very pleasant). The participants were prompted for their rating immediately after viewing the image. The same materials were used in both experiments with a minor variation in the procedures. In the Confirmatory experiment, participants were directed as to where search targets might appear in the image (e.g., on flat surfaces). No such instructions were provided in the Exploratory experiment.

160

In both experiments, trials were presented in one mixed block, and three separate task blocks. For the mixed block, the trial types were randomly intermixed within the block. For the three separate task blocks, each block was XX trials consisting entirely of one of the three conditions (Search, Memorize, Rate). Each trial was presented for 10 seconds. The inter-trial interval lasted XX seconds. The participants were seated XX inches from a XXresolutionXX monitor. The pictures were 1024 x 768 pixels, subtending a visual angle of XX degrees.

167

### Datasets

168

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

175

The raw x-coordinate, y-coordinate, and pupil size data collected at every sampling time point in the trial were used as inputs to the deep learning classifier. These data were

177 also used to develop plot image datasets that were classified separately from the raw timeline  
 178 datasets. For the plot image datasets, the timeline data for each trial were converted into  
 179 scatterplot diagrams. The x- and y- coordinates and pupil size were used to plot each data  
 180 point onto a scatterplot (e.g., see Figure 1). The coordinates were used to plot the location  
 181 of the dot, pupil size was used to determine the relative size of the dot, and shading of  
 182 the dot was used to indicate the time-course of the eye movements throughout the trial.  
 183 The background of the plot images and first data point was white. Each subsequent data  
 184 point was one shade darker than the previous data point until the final data point was  
 185 reached. The final data point was black. For standardization, pupil size was divided by  
 186 10, and one **XXunit?XX** was added. The plots were sized to match the dimensions of the  
 187 data collection monitor (1024 x 768 pixels) then shrunk to (240 x 180 pixels) in an effort to  
 188 reduce the dimensionality of the data.

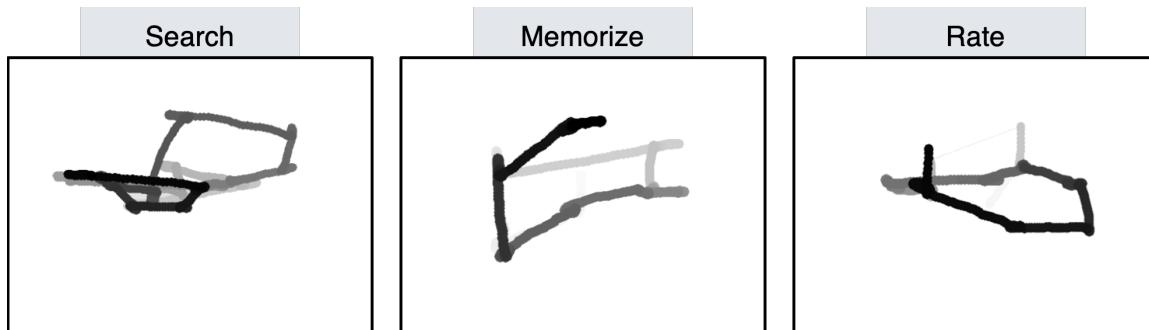


Figure 1. Each trial was represented as an image. Each sample collected within the trial was plotted as a dot in the image. Pupil size was represented by the size of the dot. The time course of the eye movements was represented by the gradual darkening of the dot over time.

189 **Data Subsets.** The full timeline dataset was structured into three columns repre-  
 190 senting the x- and y- coordinates, and pupil size for each data point collected in the first  
 191 six seconds of each trial. To systematically assess the predictive value of each XYP (i.e.,  
 192 x-coordinates, y-coordinates, pupil size) component of the data, the timeline and image  
 193 datasets were batched into subsets that excluded one of the components (i.e., XYØ, XØY,  
 194 ØYP), or contained only one of the components (i.e., XØØ, ØYØ, ØØP). For the timeline  
 195 datasets, this means that the columns to be excluded in each data subset were replaced with  
 196 zeros. The data were replaced with zeros because removing the columns would change the  
 197 structure of the data. The same systematic batching process was carried out for the image  
 198 dataset. See Figure 2 for an example of each of these image data subsets.

## 199 Classification

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

206 The development of these models were not guided by any formal theoretical assumptions

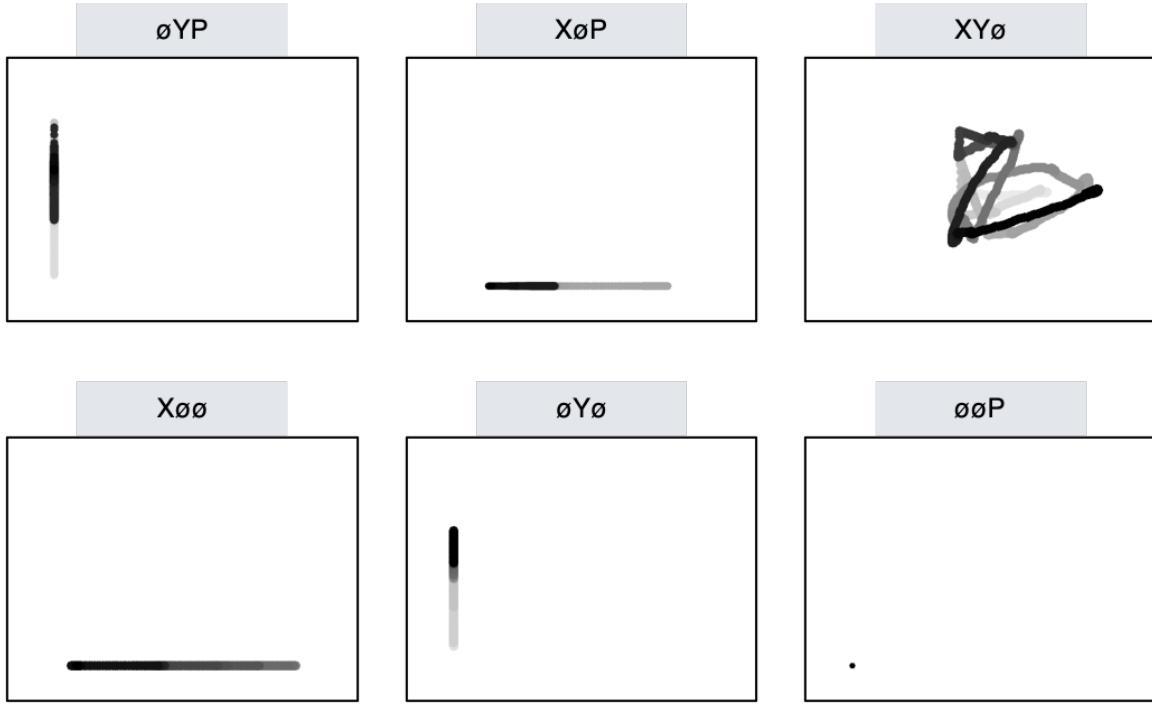


Figure 2. Plot images were used to represent each type of data subset. As with the trials in the full XYP dataset, the time course of the eye movements was represented by the shading of the dot. The first sample of each trial was white, and the last sample was black.

regarding the patterns or features likely to be extracted by the classifier. The models were developed using version 0.3b of the DeLINEATE toolbox, which operates over a Keras backend (<http://delineate.it>). Each implementation of the model randomly split the data so that 70% of the trial data were allocated to training, 15% of the trial data were allocated to validation, and 15% of the trial data were allocated to testing. Training of the model was stopped when validation accuracy did not improve over the span of 100 iterations. Once the early stopping threshold was reached, the resulting model was tested on the held out test data. This process was repeated 10 times for each model, resulting in 10 classification accuracy scores for each implementation of the model. The average of the resulting accuracy scores were the subject of comparisons against chance and other datasets or data subsets.

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

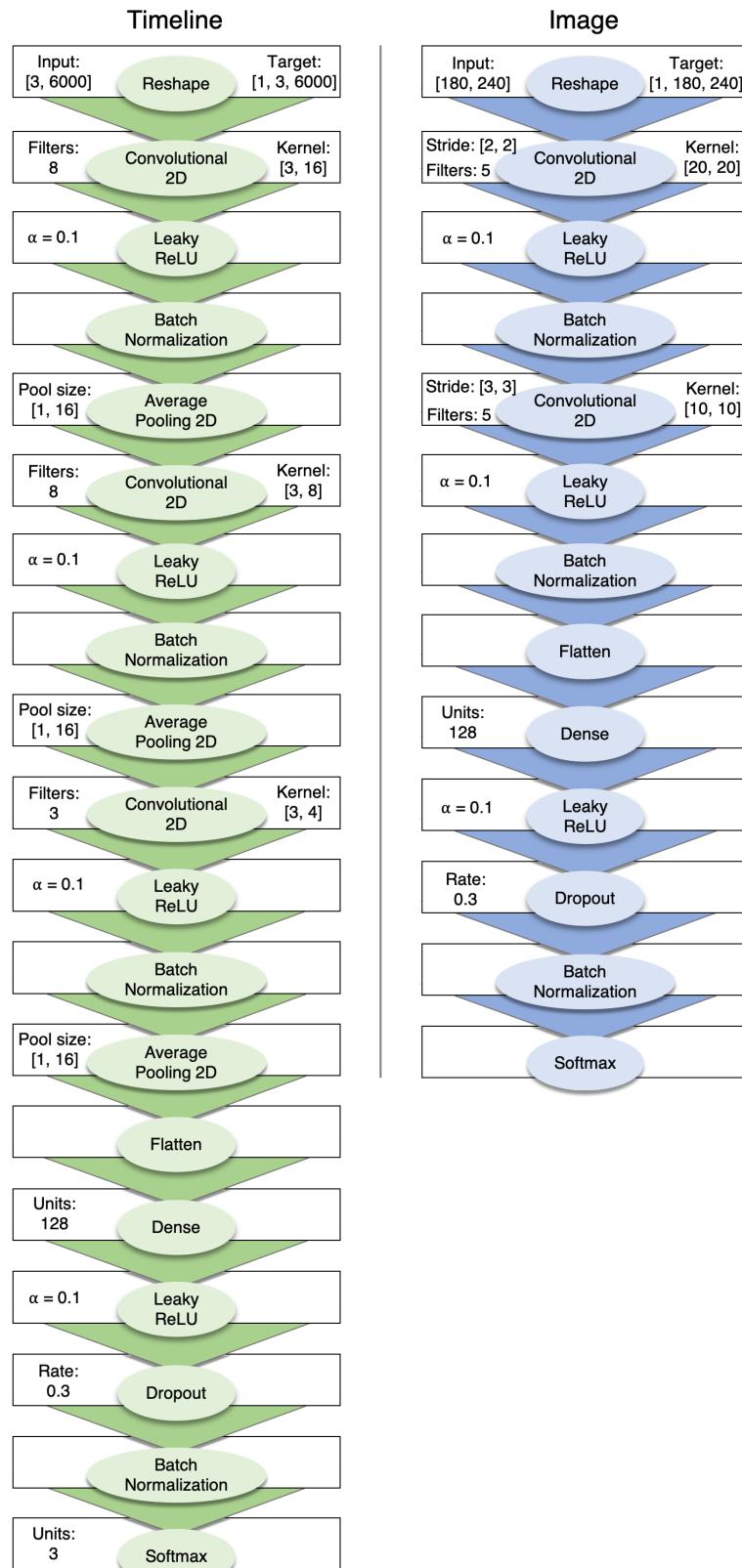


Figure 3. Two different model architectures were used to classify the timeline and image data. Both models were compiled using a categorical crossentropy loss function, and optimized with the Adam algorithm.

**224 Analysis**

225 Results for the CNN architecture that resulted in the highest accuracy on the Ex-  
226 ploratory dataset are reported below. For every dataset tested, a one-sample two-tailed  
227 *t*-test was used to compare the CNN accuracies against chance (33%). The Shapiro-Wilk  
228 test was used to assess the normality for each dataset. When normality was assumed, the  
229 mean accuracy for that dataset was compared against chance using Student's one-sample  
230 two-tailed *t*-test. When normality could not be assumed, the median accuracy for that  
231 dataset was compared against chance using Wilcoxon's Signed Rank test.

232 To determine the relative value of the three components of the eye movement data, the  
233 data subsets were compared within the timeline and plot image data types. If classification  
234 accuracies were lower when the data was batched into subsets, the component that was  
235 removed was assumed to have some diagnostic contribution that the model was using to  
236 inform classification decisions. To determine the relative value of the contribution from each  
237 component, the accuracies from each subset with one component of the data removed were  
238 compared to the accuracies for the full dataset (XYP) using a one-way between-subjects  
239 Analysis of Variance (ANOVA). To further evaluate the decodability of each component  
240 independently, the accuracies from each subset containing only one component of the eye  
241 movement data were compared within a separate one-way between-subject ANOVA. All  
242 post-hoc comparisons were corrected using Tukey's *HSD*.

**243 Results****244 Timeline Data Classification**

245 **Exploratory.** Classification accuracies for the XYP timeline dataset were well above  
246 chance ( $M = .526$ ,  $SD = .018$ ;  $t(9) = 34.565$ ,  $p < .001$ ). Accuracies for classifications  
247 of the batched data subsets were all better than chance (see Figure 4). As shown in the  
248 confusion matrices displayed in Figure 5, the data subsets with lower overall classification  
249 accuracies almost always classified the Memorize condition at or below chance levels of  
250 accuracy. Misclassifications of the Memorize condition were split relatively evenly between  
251 the Search and Rate conditions.

252 There was a difference in classification accuracy for the XYP dataset and the subsets  
253 that had the pupil size, x-coordinate, and y-coordinate data systematically removed ( $F_{(3,36)}$   
254 = 47.471,  $p < .001$ ,  $\eta^2 = 0.798$ ). Post-hoc comparisons against the XYP dataset showed  
255 that classification accuracies were not affected by the removal of pupil size or y-coordinate  
256 data (see Table 2). The null effect present when pupil size was removed suggests that the  
257 pupil size data were not contributing unique information that was not otherwise provided  
258 by the x- and y-coordinates. A strict significance threshold of  $\alpha = .05$  implies the same  
259 conclusion for the y-coordinate data, but the relatively low degrees of freedom ( $df = 18$ )  
260 and the borderline observed *p*-value ( $p = .056$ ) affords the possibility that there exists a  
261 small effect. Moreover, classification for the  $\emptyset$ YP subset was lower than the XYP dataset,  
262 showing that the x-coordinate data were uniquely informative to the classification.

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

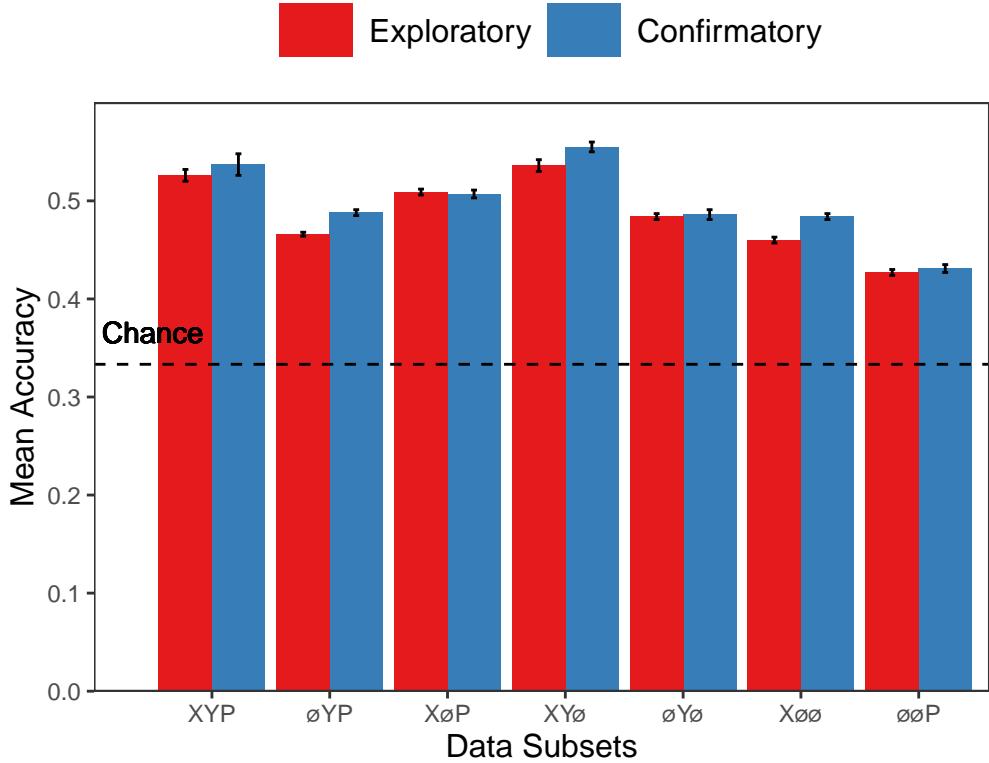


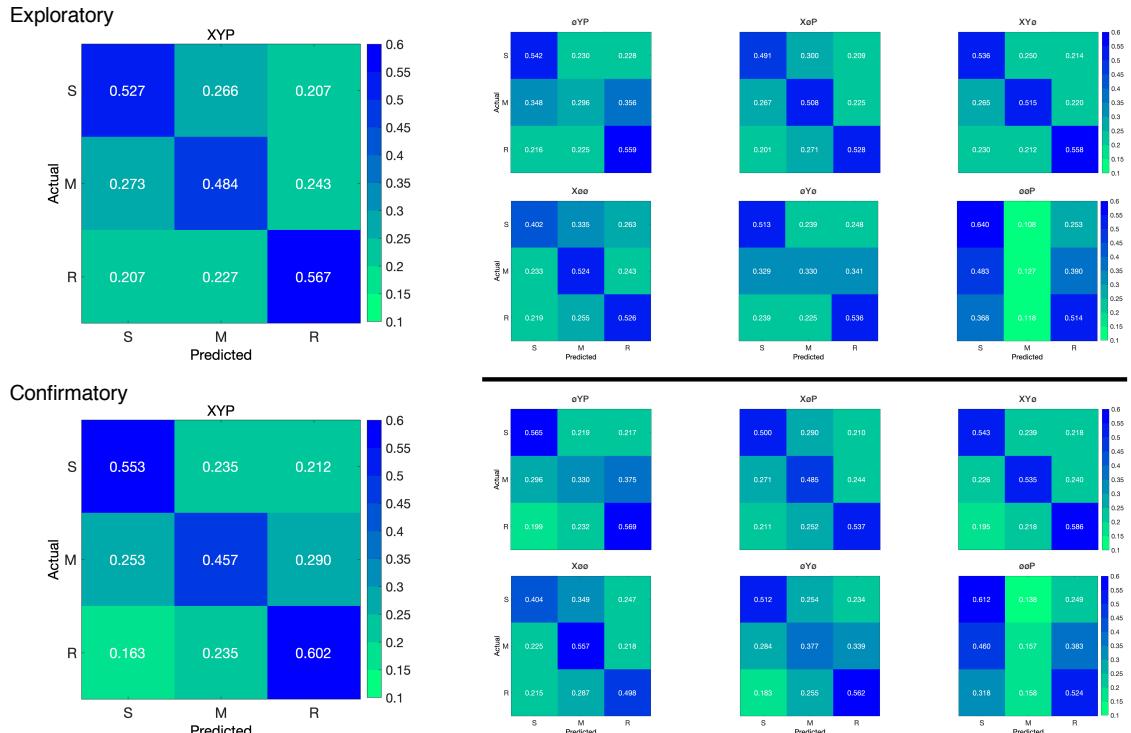
Figure 4. The graph represents the average accuracy reported for each subset of the timeline data. All of the data subsets were decoded at levels better than chance (33%). The error bars represent standard errors.

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

classification accuracy for the ØØP subset was lower than the XØØ and ØYØ subsets. Classification accuracy for the XØØ subset was higher than the ØYØ subset. Altogether, these findings suggest that pupil size data was the least uniquely informative to classification decisions, while the x-coordinate data was the most uniquely informative.

**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



*Figure 5.* The confusion matrices represent the average classification accuracies for each condition of the timeline data. 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.

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 precise (as indicated by the increase in standard deviation).

287 **Plot Image Classification**

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

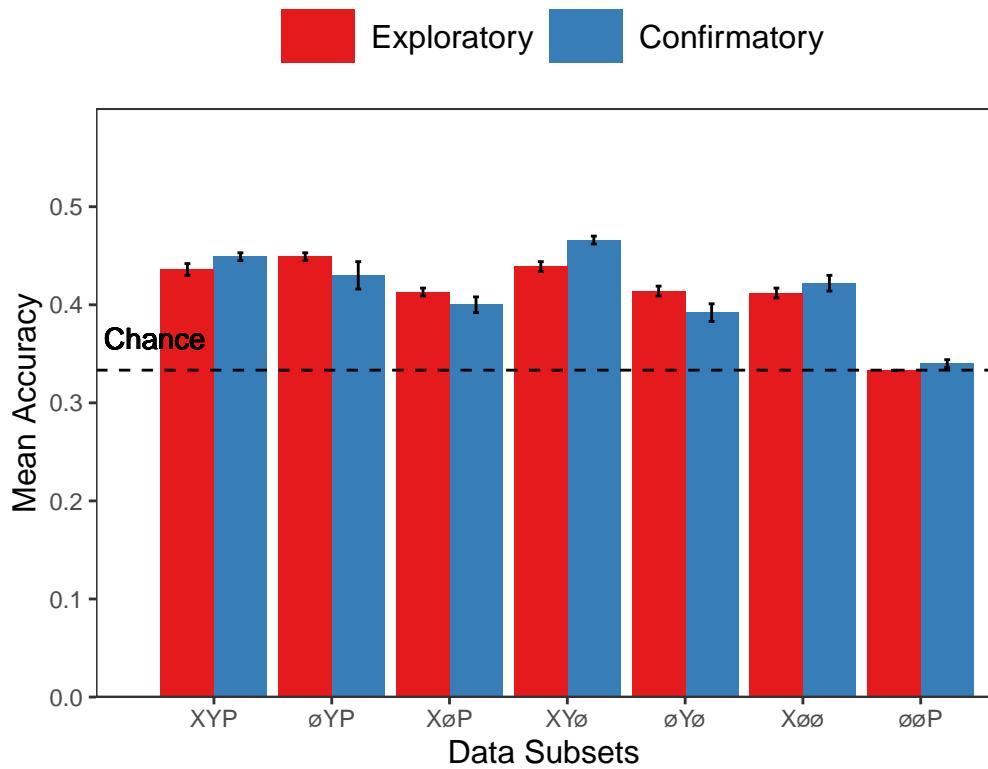


Figure 6. The graph represents the average accuracy reported for each subset of the image data. All of the data subsets except for the Exploratory XY∅ dataset were decoded at levels better than chance (33%). The error bars represent standard errors.

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

300 There was also a difference in classification accuracies between the X∅∅, ∅Y∅,  
 301 and ∅∅P subsets (Levene's test:  $F_{(2,27)} = 3.815, p = .035$ ; Welch correction for lack of  
 302 homogeneity of variances:  $F_{(2,17.993)} = 228.137, p < .001, \eta^2 = .899$ ). Post-hoc comparisons  
 303 showed that there was no difference in classification accuracies for the X∅∅ and ∅Y∅

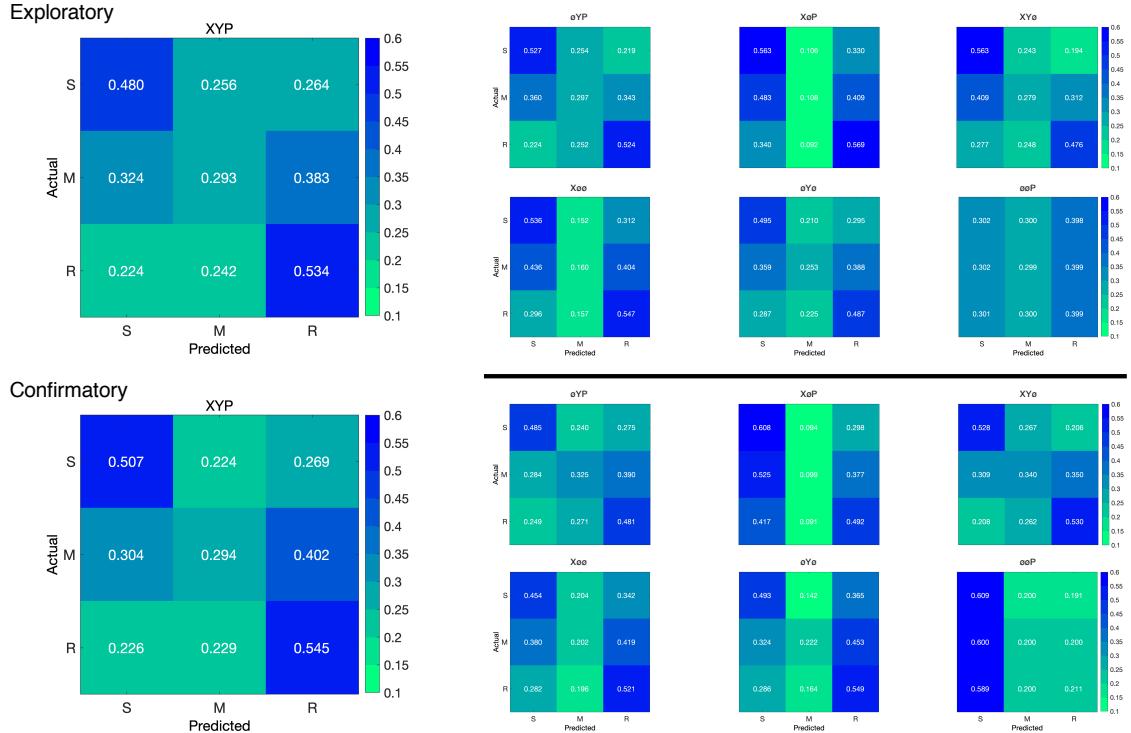


Figure 7. The confusion matrices represent the average classification accuracies for each condition of the image data. The vertical axis of the confusion matrices represents the actual condition for the trial. The horizontal axis of the confusion matrices represents the condition that was predicted by the model.

Table 3  
*Image Subset Comparisons*

Comparison	Exploratory		Confirmatory	
	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>
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

304 subsets, but classification for the ØØP subset were less accurate than the XØØ and ØYØ  
305 subsets.

306 **Confirmatory.** Classification accuracies for the confirmatory image dataset were  
307 well above chance ( $M = .449$ ,  $SD = 0.012$ ,  $t_{(9)} = 31.061$ ,  $p < .001$ ), but were less accurate  
308 than the classifications of the confirmatory timeline dataset ( $t_{(18)} = 11.167$   $p < .001$ ).  
309 Accuracies for classifications of the data subsets were also all better than chance (see Figure  
310 6). The confusion matrices followed the pattern showing that the Memorize condition was  
311 confused most often, and was relatively evenly mis-identified as a Search or Rate trial (see

312 Figure 7). As with the timeline data, the general trend showing that pupil size data was the  
313 least informative to the model was replicated in the Confirmatory dataset (see Table 3).

314 To test the generalizability of the model, the classification accuracies for the XYP  
315 Exploratory and Confirmatory plot image datasets were compared. The independent samples  
316 *t*-test showed that the deep learning model did equally well at classifying the Exploratory  
317 and Confirmatory plot image datasets,  $t_{(18)} = 1.777$ ,  $p = .092$ , Cohen's  $d = 0.795$ .

### 318 Discussion

319 The present study aimed to produce a practical and reliable example of a black box  
320 solution to the inverse Yarbus problem. To implement this solution, we classified raw  
321 timeline and minimally processed plot image data using a CNN model architecture. To  
322 our knowledge, this study was the first to provide a solution to determining mental state  
323 from eye movement data using each of the following: (1) Non-aggregated eye tracking data  
324 (i.e., raw x-coordinates, y-coordinates, pupil size), (2) timeline and image data formats (see  
325 Figure 2), and (3) a black box CNN architecture. This study probed the relative predictive  
326 value of the x-coordinate, y-coordinate, and pupil size components of the eye movement data  
327 using a CNN. The CNN was able to decode the timeline and plot image data better than  
328 chance, although only the timeline datasets were decoded with state-of-the-art accuracy.  
329 Datasets with lower classification accuracies were not able to differentiate the cognitive  
330 processes underlying the Memorize task from the cognitive processes underlying the Search  
331 and Rate tasks. Decoding subsets of the data revealed that pupil size was the least uniquely  
332 informative component of the eye movement data. This pattern of findings was consistent  
333 between the Exploratory and Confirmatory datasets.

334 Although several aggregate eye movement features have been tested as task predictors,  
335 to our knowledge, no other study has assessed the predictive value of the data format  
336 (viz., data in the format of a plot image). Our results suggest that although CNNs are  
337 robust image classifiers, eye movement data is decoded in the standard timeline format more  
338 effectively than in image format. This may be because the image data format may contain  
339 less decodable information than the timeline format. Over the span of the trial (six seconds),  
340 the eye movements occasionally overlapped. When there was an overlap in the image data  
341 format, the more recent data points overwrote the older data points. This resulted in some  
342 information loss that did not occur when the data were represented in the raw timeline  
343 format. Despite this loss of information, the plot image format was still decoded with better  
344 than chance accuracy. To further examine the viability of classifying task from eye movement  
345 image datasets, future research might consider representing the data in different forms such  
346 as 3-dimensional data formats, or more complex color combinations capable of representing  
347 overlapping data points.

348 When considering the superior performance of the timeline data (vs., plot image data),  
349 we must also consider the differences in the model architectures. Because the structures of  
350 the timeline and plot image data formats were different, the models decoding those data  
351 structures also needed to be different. Both models were optimized individually on the  
352 Exploratory dataset before being tested on the Confirmatory dataset. For both timeline and  
353 plot image formats, there was good replicability between the Exploratory and Confirmatory

354 datasets, demonstrating that these architectures performed similarly from experiment to  
355 experiment. An appropriately tuned CNN should be capable of learning any arbitrary  
356 function, but given that the upper bound for decodability of these datasets is unknown,  
357 there is the possibility that a model architecture exists that is capable of classifying the  
358 plot image data format more accurately than the model used to classify the timeline data.  
359 Despite this possibility, the convergence of these findings with other studies (see Table 1)  
360 suggests that the results of this study are approaching a ceiling for the potential to solve  
361 the inverse Yarbus problem with eye movement data. Although the true capacity to predict  
362 mental state from eye movement data is unknown, standardizing datasets in the future could  
363 provide a point for comparison that can more effectively indicate which methods are more  
364 effectively solving the inverse Yarbus problem.

365 In the current study, the Memorize condition was most often confused with the  
366 Search and Rate conditions, especially for the datasets with lower overall accuracy. This  
367 suggests that the eye movements associated with the Memorize task were potentially lacking  
368 unique or informative features to decode. This means that eye movements associated with  
369 the Memorize condition were interpreted as noise, or were sharing features of underlying  
370 cognitive processes that were represented in the eye movements associated with the Search  
371 and Rate tasks. Previous research (e.g., Król & Król, 2018) has attributed the inability  
372 to differentiate one condition from the others to the overlapping of sub-features in the eye  
373 movements between two tasks that are too subtle to be represented in the eye movement  
374 data. An additional analysis was conducted with the Memorize condition removed from the  
375 Exploratory and Confirmatory timeline and plot image data (see Appendix A). With the  
376 Memorize condition removed, **these findings indicate that blahblahblah.**

377 When determining the relative contributions of the the eye movement features used in  
378 this study (x-coordinates, y-coordinates, pupil size), the pupil size data was consistently  
379 the least uniquely informative. When pupil size was removed from the Exploratory and  
380 Confirmatory timeline and plot image datasets, classification accuracy remained stable (vs.,  
381 XYP dataset). Furthermore, classification of the ØØP subset was the lowest of all of the data  
382 subsets, and in one instance, was no better than chance. Although these findings indicate  
383 that, in this case, pupil size was a relatively uninformative component of the eye movement  
384 data, previous research has associated changes in pupil size as indicators of working memory  
385 load (Kahneman & Beatty, 1966; Karatekin, Couperus, & Marcus, 2004), arousal (Wang  
386 et al., 2018), and cognitive effort (Porter, Troscianko, & Gilchrist, 2007). The results of  
387 the current study indicate that the changes in pupil size associated with these underlying  
388 processes are not useful in delineating the tasks being classified (i.e., Search, Memorize,  
389 Rate), potentially because these tasks do not evoke a reliable pattern changes in pupil size.

390 The findings from the current study support the notion that black box CNNs are a  
391 viable approach to determining task from eye movement data. In a recent review, Lukander  
392 et al. (2017) expressed concern regarding the lack of generalizability of black box approaches  
393 when decoding eye movement data. Overall, the current study showed a consistent pattern  
394 of results for the XYP timeline and image datasets, but some minor inconsistencies in the  
395 pattern of results for the x- and y- coordinate subset comparisons. These inconsistencies  
396 may be a product of overlap in the cognitive processes underlying the three tasks. When

397 the data are batched into subsets, at least one dimension (i.e., x-coordinates, y-coordinates,  
398 or pupil size) are removed, leading to a potential loss of information. When the data  
399 provide fewer meaningful distinctions, finer-grained inferences are necessary for the tasks  
400 to be distinguishable. As shown by Coco and Keller (2014), eye movement data can be  
401 more effectively decoded when the cognitive processes underlying the tasks are explicitly  
402 differentiable. While the cognitive processes distinguishing memorizing, searching, or rating  
403 an image are intuitively different, the eye movements elicited from these cognitive processes  
404 are not easily differentiated. To correct for potential mismatches between the distinctive  
405 task-diagnostic features in the data and the level of distinctiveness required to classify the  
406 tasks, future research could more definitively conceptualize the cognitive processes underlying  
407 the task-at-hand.

408 Classifying mental state from eye movement data is often carried out in an effort  
409 to advance technology to improve educational outcomes, strengthen the independence of  
410 physically and mentally handicapped individuals, or improve HCI's (Koochaki & Najafizadeh,  
411 2018). Given the previous questions raised regarding the reliability and generalizability  
412 surrounding the black box nature of CNN classification, the current study first tested models  
413 on an exploratory dataset, then confirmed the outcome using a second independent dataset.  
414 Overall, the findings of this study indicate that this black box approach is capable of  
415 producing a stable and generalizable outcome. Although the timeline data outperformed  
416 the image data format, future studies that incorporate stimulus features might have the  
417 potential to surpass current state-of-the-art classification. According to Bulling, Weichel, and  
418 Gellersen (2013), incorporating stimulus feature information into the dataset may provide  
419 information that is diagnostic beyond decoding spatial location data alone. Alternatively,  
420 Borji and Itti (2014) suggested that accounting for salient features in the stimulus might  
421 leave little to no room for the classifier to consider mental state. Future research should  
422 examine the potential for the inclusion of stimulus feature information in addition to the eye  
423 movement data to boost the classification accuracy of image data beyond that of timeline  
424 data.

- 425 Boisvert, J. F. G., & Bruce, N. D. B. (2016). Predicting task from eye movements: On the  
426 importance of spatial distribution, dynamics, and image features. *Neurocomputing*,  
427 207, 653–668. <https://doi.org/10.1016/j.neucom.2016.05.047>
- 428 Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task.  
429 *Journal of Vision*, 14(3), 29–29. <https://doi.org/10.1167/14.3.29>
- 430 Bulling, A., Weichel, C., & Gellersen, H. (2013). EyeContext: Recognition of high-level  
431 contextual cues from human visual behaviour. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (p. 305). Paris, France:  
432 ACM Press. <https://doi.org/10.1145/2470654.2470697>
- 433 Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye  
434 movement control during active scene perception. *Journal of Vision*, 9(3), 6–6.  
435 <https://doi.org/10.1167/9.3.6>
- 436 Coco, M. I., & Keller, F. (2014). Classification of visual and linguistic tasks using eye-  
437 movement features. *Journal of Vision*, 14(3), 11–11. <https://doi.org/10.1167/14.3.11>
- 438 DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited.  
439 *Visual Cognition*, 17(6-7), 790–811. <https://doi.org/10.1080/13506280902793843>
- 440 Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to  
441 predict observers' task from eye movement patterns. *Vision Res*, 62, 1–8. <https://doi.org/10.1016/j.visres.2012.03.019>
- 442 Haji-Abolhassani, A., & Clark, J. J. (2014). An inverse Yarbus process: Predicting observers'  
443 task from eye movement patterns. *Vision Research*, 103, 127–142. <https://doi.org/10.1016/j.visres.2014.08.014>
- 444 Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013).  
445 Predicting Cognitive State from Eye Movements. *PLoS ONE*, 8(5), e64937. <https://doi.org/10.1371/journal.pone.0064937>
- 446 Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*,  
447 154(3756), 1583–1585. Retrieved from <https://www.jstor.org/stable/1720478>
- 448 Kanan, C., Ray, N. A., Bseiso, D. N. F., Hsiao, J. H., & Cottrell, G. W. (2014). Predicting  
449 an observer's task using multi-fixation pattern analysis. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '14* (pp. 287–290).  
Safety Harbor, Florida: ACM Press. <https://doi.org/10.1145/2578153.2578208>
- 450 Karatekin, C., Couperus, J. W., & Marcus, D. J. (2004). Attention allocation in the dual-  
451 task paradigm as measured through behavioral and psychophysiological responses.  
452 *Psychophysiology*, 41(2), 175–185. <https://doi.org/10.1111/j.1469-8986.2004.00147.x>
- 453 Koochaki, F., & Najafizadeh, L. (2018). Predicting Intention Through Eye Gaze Patterns.  
454 In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (pp. 1–4).  
455 <https://doi.org/10.1109/BIOCAS.2018.8584665>
- 456 Król, M. E., & Król, M. (2018). The right look for the job: Decoding cognitive processes

- 463 involved in the task from spatial eye-movement patterns. *Psychological Research*.  
464 <https://doi.org/10.1007/s00426-018-0996-5>
- 465 Lukander, K., Toivanen, M., & Puolamäki, K. (2017). Inferring Intent and Action from  
466 Gaze in Naturalistic Behavior: A Review. *International Journal of Mobile Human*  
467 *Computer Interaction*, 9(4), 41–57. <https://doi.org/10.4018/IJMHCI.2017100104>
- 468 MacInnes, W., Joseph, Hunt, A. R., Clarke, A. D. F., & Dodd, M. D. (2018). A Generative  
469 Model of Cognitive State from Task and Eye Movements. *Cognitive Computation*,  
470 10(5), 703–717. <https://doi.org/10.1007/s12559-018-9558-9>
- 471 Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011).  
472 Examining the influence of task set on eye movements and fixations. *Journal of*  
473 *Vision*, 11(8), 17–17. <https://doi.org/10.1167/11.8.17>
- 474 Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting:  
475 Insights from pupillometry. *Quarterly Journal of Experimental Psychology* (2006),  
476 60(2), 211–229. <https://doi.org/10.1080/17470210600673818>
- 477 Seeliger, K., Fritzsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., &  
478 van Gerven, M. A. J. (2018). Convolutional neural network-based encoding and  
479 decoding of visual object recognition in space and time. *NeuroImage*, 180, 253–266.  
480 <https://doi.org/10.1016/j.neuroimage.2017.07.018>
- 481 Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus,  
482 Eye Movements, and Vision. *I-Perception*, 1(1), 7–27. <https://doi.org/10.1068/i0382>
- 483 Wang, C.-A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C., & Munoz, D. P. (2018).  
484 Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional  
485 Face Task. *Frontiers in Neurology*, 9. <https://doi.org/10.3389/fneur.2018.01029>
- 486 Yarbus, A. (1967). Eye Movements and Vision. Retrieved January 24, 2019, from [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)
- 487 Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2019). Comparing the Interpretability of  
488 Deep Networks via Network Dissection. In W. Samek, G. Montavon, A. Vedaldi,  
489 L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and*  
490 *Visualizing Deep Learning* (pp. 243–252). Cham: Springer International Publishing.  
491 [https://doi.org/10.1007/978-3-030-28954-6\\_12](https://doi.org/10.1007/978-3-030-28954-6_12)
- 492
- 493