Exploring the Effects of Scanpath Feature Engineering for Supervised Image Classification Models

SEAN ANTHONY BYRNE, MoMiLab, IMT School for Advanced Studies Lucca, Italy
VIRMARIE MAQUILING, Human-Computer Interaction, University of Tübingen, Germany
ADAM PETER FREDERICK REYNOLDS, MoMiLab, IMT School for Advanced Studies Lucca, Italy
LUCA POLONIO, Department of Economics, Management and Statistics, Università degli Studi di Milano Bicocca, Italy
NORA CASTNER, Human-Computer Interaction, University of Tübingen, Germany
ENKELEJDA KASNECI, Human-Centered Technologies for Learning, Technical University of Munich, Germany

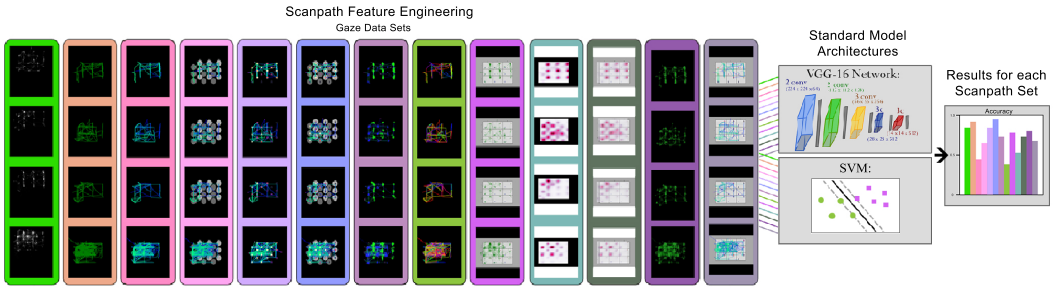

Fig. 1. The workflow of our experiment: The raw gaze data is transformed into sets of scanpath images, each set contains different engineered features commonly found in the eye-tracking literature. The sets are each used as model input into a VGG-16 Convolutional Neural Network and evaluated using k-fold cross-validation and, in a separate trial, using a holdout test set. We also evaluate the sets of scanpaths using a SVM Image Classifier. The resulting metrics from each model experiment are reported for each set of scanpath images.

Image classification models are becoming a popular method of analysis for scanpath classification. To implement these models, gaze data must first be reconfigured into a 2D image. However, this step gets relatively little attention in the literature as focus is mostly placed on model configuration. As standard model architectures have become more accessible to the wider eye-tracking community, we highlight the importance of carefully

Authors' addresses: Sean Anthony Byrne, sean.byrne@imtlucca.it, MoMiLab, IMT School for Advanced Studies Lucca, Piazza S.Francesco, 19, Lucca, LU, Italy, 55100; Virmarie Maquiling, virmarie.maquiling@student.uni-tuebingen.de, Human-Computer Interaction, University of Tübingen, Sand 14, Tübingen, Germany, 72076; Adam Peter Frederick Reynolds, adam.reynolds@imtlucca.it, MoMiLab, IMT School for Advanced Studies Lucca, Piazza S.Francesco, 19, Lucca, LU, Italy, 55100; Luca Polonio, luca.polonio@unimib.it,, Department of Economics, Management and Statistics, Università degli Studi di Milano Bicocca, Piazza dell'Ateneo Nuovo, Milan, Italy, 20126; Nora Castner, nora.castner@uni-tuebingen.de, Human-Computer Interaction, University of Tübingen, Sand 14, Tübingen, Germany, 72076; Enkelejda Kasneci, enkelejda.kasneci@tum.de, Human-Centered Technologies for Learning, Technical University of Munich, Marsstraße 20-22, Munich, Germany, 80335.

2573-0142/2023/5-ART161 $15.00
https://doi.org/10.1145/3591130

choosing feature representations within scanpath images as they may heavily affect classification accuracy. To illustrate this point, we create thirteen sets of scanpath designs incorporating different eye-tracking feature representations from data recorded during a task-based viewing experiment. We evaluate each scanpath design by passing the sets of images through a standard pre-trained deep learning model as well as a SVM image classifier. Results from our primary experiment show an average accuracy improvement of 25 percentage points between the best-performing set and one baseline set.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Feature selection*; Computer vision representations; • **Applied computing** → *Psychology*.

Additional Key Words and Phrases: Scanpaths, Feature engineering, Computer vision, Image processing, Signal processing, Eye movements and cognition, Visual search behavior, Machine learning

ACM Reference Format:
Sean Anthony Byrne, Virmarie Maquiling, Adam Peter Frederick Reynolds, Luca Polonio, Nora Castner, and Enkelejda Kasneci. 2023. Exploring the Effects of Scanpath Feature Engineering for Supervised Image Classification Models. *Proc. ACM Hum.-Comput. Interact.* 7, ETRA, Article 161 (May 2023), 18 pages. https://doi.org/10.1145/3591130

1 INTRODUCTION

Eye-tracking research involves high level abstraction of complex cognitive processes from the eye movement events. The scanpath – the patterns of the eye movements evoked by specific tasks – offers further insight into these complex strategies by providing a more detailed depiction of either temporal, spatial, or both characteristics. Comparing scanpaths can create groupings based on sequence similarity, transition frequencies, etc. However, classification based on the scanpath features can predict which group or task a scanpath belongs to based on the learned patterns attributed to specific groups. Scanpath classification has been used to successfully determine experts from novices [11, 34, 35], neurological disorders [18, 72, 75, 76], and cognitive states [8, 50]. These are only a small facet of what scanpath classification can be used for. Thus, there is still a growing demand for novel approaches to scanpath classification.

The research on scanpath classification models has evolved over the past years in order to keep up with the state of the art from a number of fields. For instance, perspectives from computer vision (i.e., saliency models) and bioinformatics (i.e., Needleman-Wunsch algorithm) promoted new insights into how the the gaze behavior is represented in a model. More recently, scanpath classification has benefited greatly from advancements in machine learning and deep learning. These models have promoted scanpath classification so that they can handle high dimensional data and are able to easily recognize patterns [2]. However, with these models, the concept of scanpath representation becomes even more crucial. One ever-growing perspective that is gaining traction in scanpath classification is image classification. This work takes an in-depth look into how scanpaths can be represented as images for the most optimal feature input into image classification models.

Supervised image classification models have enjoyed a rise in popularity for a little over a decade now. This can be attributed to a variety of factors such as the creation of new machine learning frameworks such as Keras[14] designed to empower people from non-technical backgrounds with the ability to implement these models. In addition, advances in training regimes such as the introduction of transfer learning and fine-tuning strategies overcome both the problems of limited data, which prevented the use of these models in many domains, and the time-consuming task of configuring a network from scratch [22, 52, 53, 82]. Taken together, these factors have encouraged the implementation of image classification algorithms across a variety of different domains as wide-ranging as medical applications to cybersecurity, with results often outperforming traditional methods of analyses [4, 65]. The expanded use of image classification models has led to the rather interesting phenomenon of researchers devising innovative and creative methods to

represent features of their recorded data as an image in order to take advantage of the power of image classification models. Examples of this practice include, music classification tasks where the frequencies of the music are transformed into spectrograms and used as model input for a convolutional neural network (CNN)[15], electrocardiogram (ECG) arrhythmia classification where researchers transformed the recorded signal into an image[41], and algorithms designed to convert tabular data into images [86], opening the potential of image classification models to many more domains of science.

The idea of representing the scanpath data as an image conforms to 2D spatial understanding. Where some other approaches reshape the scanpath to an array (Needleman-Wunsch and other string comparison approaches), the 2D approaches normalize the scanpath to align with the common visualization methods. For instance, heatmaps and other attention distribution visualizations have been a prominent metric for scanpath assessment [7, 10, 28, 29, 32, 33, 37, 61, 84]. Moreover, incorporating saliency models[1] offers another layer of semantic information (e.g. scene understanding and perceivability) [12, 24, 29, 30, 47, 55, 63, 79]. This harmony of scene semantics and attentional effects provides input for deep learning that works towards human perception [69]. The recent literature on deep learning for image understanding (w.r.t. segmentation, classification, etc.) has also contributed a higher level of automated semantic extraction for scanpath analysis [5, 46, 74, 83]. However, using image classification models on the scanpath rather than the image, bring the focus back to attentional models not confined to specific images or features.

The field of eye-tracking has seen a development of image classification models that achieve state-of-the-art performance across both free-viewing and task-based viewing experiments. We refer the readers to the related work section for further details on image classification models. While these models produce competitive results compared to more traditional approaches, we aim to showcase a gap in the literature regarding the construction of the input feature space when using gaze data as input to image classification algorithms. To address this gap in the literature, we run a series of experiments by first creating sets of scanpath images; all of the images are generated from the same data, yet differ in how they represent the gaze data. We systematically test the effects of building three of the most commonly associated features of gaze data into the images, namely saccades, fixations, and Areas of interest (AOIs). Additionally, we investigate aspects such as sequential coloring of saccadic information or aggregating fixations, which make these features more salient to a kernel-based model. Then, we test the impact of adding these features to the input and report model accuracy and other metrics. We conduct a series of model tests using a pre-trained VGG-16 and a simple SVM model and compare metrics such as Accuracy, F1-score, and AUC to assess performance differences across the different sets of scanpath images.

We explore this scanpath feature engineering and model assessment on data published by Marchiori et al.. This data consists of a task-based viewing experiment where participants play normal-form matrix games against a computer employing a strategy based on the Nash Equilibrium. The Nash Equilibrium is a game theoretical concept where, in a given game, no player has anything to gain by changing their own strategy w.r.t the known strategy of the counterpart [26]. We transformed this task into a binary classification task attempting to classify only if the participant selects the Nash Equilibrium or not. Further details of the dataset are found in section 3.1. We chose this experiment as the environment is sparse, meaning that the scanpaths are relatively simple. Also, the relationship between choices and gaze behavior is well established in the literature [20, 48, 49, 58, 59]. We provide access to the code at the following data used at the following links https://github.com/vbmaq/ImageMaker & https://osf.io/fhmjy .

[1]Computer vision based models that reflect biological processes of how humans visually process a scene.

2 RELATED WORK

2.1 Scanpath as an Image for Classification

As deep learning models for image classification covers a vast range of literature in scanpath prediction and generation, we restrict our literature review classification of human scanpaths. Defining saliency as an attention representation, fixations maps created from the whole scanpath were used as input for a CNN and clustering of scanpath behaviors related to autism spectrum disorder (ASD) [23, 60]. Heatmaps have also been used as input into multiple deep learning models for robust schizophrenia classification [42] and task classification [78]. To incorporate temporal information, CNN-LSTM networks using scanpath-based patches from a saliency-predicted map could accurately classify autism spectrum disorder [13, 75]. [25] created collective *snapshots* of the gaze based on group attention at time windows as input for an LSTM. Somewhat similar, gaze coupled with activation maps fed into a neural network was able to reconstruct the image from semantic features extracted from the scanpath [66, 73]. These attention map approaches often employ techniques, such as Gaussian blurring, to the raw data. This approach offers well supported means of normalizing noisy signals. Whereas others have employed different creations of scanpaths images for CNN classifications: For instance, Markov Transition Fields [78, 81], graphs-based [17, 30, 77], and PCA [42, 45].

Another representation alternative is to create images from the raw scanpath data. This approach avoids pre-processing, which could remove potentially relevant information for the model [42]. [71] created scanpath images from the raw gaze (also, lines connecting x,y coordinates together) for the whole duration and for five second intervals and used them for input for a CNN and RNN, respectively for combined classification of confusion.[1] created grayscale scanpath images by connecting saccades and using a an intensity weighting based on the fixation densities were able to classify ASD gaze using a CNN. Researchers developed a generative model for scanpath classification that transformed gaze data into emojis and then used the emojis to classify scanpaths [27]. Initially, it encodes gaze data as a compact image with the spatial, temporal, and connectivity represented as pixel values in the red, green, and blue channels, respectively. Recently, [3] looked at different representations of scanpaths for ASD classification. They employed temporal coloring to represent saccade velocities and fed this input into neural networks to achieve high classification accuracy. In a similar vein to our experiment, [6] represented the scanpath as an image using symbols to encode different fixation durations (e.g. circle for less than 200ms, star for around 300ms). Their model achieved up to 80% accuracy in predicting text relevancy via eye movement behavior.

To date, there is only one scientific report that delves into feature engineering of scanpath data. [85] examine different scanpath feature engineering approaches for CNN input. They tested multiple models including a VGG model pre-trained on Imagenet, where they found that the VGG model and training regime works well for scanpath classification. Our work differs from their approach as we examine features such as saccades and fixations that are more familiar and therefore more accessible to the wider eye-tracking audience, while [85] use more data driven methods.

2.2 Eye tracking in Economic Games

For decades, economists have conducted eye-tracking studies to use the gaze data as a proxy measurement for cognition. Numerous classification methods have been used to establish a relationship between gaze patterns and the cognitive process of individuals while attempting to classify their decision strategy in economic games[20, 43, 44, 48, 51]. Early attempts have used techniques such as logistic regression or cluster analysis of fixation points[59]. More recently, these studies have taken advantage of machine learning models. For example, [48] attempted to use a Salience Attentive

Model (SAM) with saliency maps as input trying to classify the equilibrium choice in two by two normal-form games. In a separate study, [44] used a Multilayer Perceptron (MLP) to detect whether a single game played by a participant was either of the "predictable" or "unpredictable" type. Scanpaths have also been used to model gaze behavior in matrix games by Byrne et al. [9], who transformed subsequences of the gaze data into scanpath images to predict choice behavior in matrix games.

3 PROPOSED APPROACH

Our contribution highlights how eye-tracking features commonly found in the literature can greatly improve classification accuracy when using supervised image classification models. The findings detailed seek to democratize the use of these classification models and support eye-tracking researchers, regardless of machine learning experience, by demonstrating the effectiveness of using the most standard model architectures combined with a feature engineering strategy based on domain knowledge of scanpaths. As scanpath features are well understood by the eye-tracking community, we feel that the presented strategies used to improve model accuracy can offer explainability of these models. Our contributions are as follows:

(1) We demonstrate that easy to implement transfer learning strategies can be applied successfully to gaze data. This finding is important because it shows that high classification accuracies can still be achieved without complex, specialized training regimes or custom architectures.
(2) We suggest that the traditional features in the eye-tracking literature are useful priors in the realm machine learning based scanpath classification.
(3) Our results hold for both deep learning and machine learning methods suggesting robustness in our scanpath design approach.
(4) The feature engineering strategy we employ can be easily understood and replicated across many other datasets and experiments by eye-tracking researchers wishing to use image classification models.

3.1 The Dataset

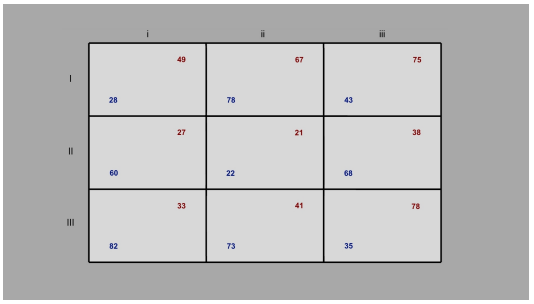

Fig. 2. An example of a gameboard used in the experiment by [51]. The payoffs of the participant are in blue, the payoffs of the other player, which is a computer algorithm, are in red. As the participants are made aware that the computer will always select a choice consistent with the Nash Equilibrium, it stands that any choice the participant makes that is not consistent with the Nash Equilibrium will lead to a sub-optimal outcome. In order for the participants to maximise their payoff, they must perform a complex visual search across the gameboard to find the Nash Equilibrium which is located at position [Row 2, Column 3].

To test the effects of using different scanpath designs, we use data from [51]. The data was recorded during a behavioral experiment where the eye movements of participants are captured

while playing a set of economic games. The data contains 243 participants (81 males, mean age = 24.1, SD age = 4.6). The participants played a set of two-person 3x3 matrix games presented in normal-form against a computer programmed to always play the action consistent with an optimal strategy that in game theory is named the Nash Equilibrium strategy [54]. Participants were informed that the computer would always play rationally and try to maximize its own payoff. The payoffs of both player and computer in the game matrix were presented in different colors to facilitate comprehension. To identify the Nash Equilibrium in the selected games, participants must use what is known in behavioral game theory as strategic sophistication [59], which is the attempt to predict the other's decisions by taking their own incentives into account and best respond to it [16]. The Nash Equilibrium action was randomized across the games in an effort to avoid any patterns which could be identified by the participants.

During the experiment, the participants played 15 independent games. For the current study, we model scanpaths using ten of these games to generate a total of 2430 scanpaths. We did not consider the other five games because they did not require strategic sophistication and the participants could play optimally without considering the other player's incentives. As this is a supervised classification task, the physical choices made by the participant using a keyboard are used as the labels in our experiments. The participants used the keyboard to select one of the three rows each game knowing that the payoff they would receive would depend on the selection made by the computer. For instance, in Figure 2 the highest payoff for the participant is located in row three. However, this payoff should be seen as unattainable by the participant as they know that a rational opponent will never select the first column. In order to best respond to the actions of the computer knowing that it will play rationally, the participant should choose row two under the assumption that the computer will choose column three. In this experiment, only one Nash Equilibrium exists in each game. To frame this experiment as a binary classification problem, we label if the participant selects the Nash Equilibrium choice as one class and any other choice as the second class. The experiment was conducted at the Experimental Psychology Laboratory of the University of Trento (Italy) and lasted around one hour.

3.2 Gaze Data and Scanpath Creation

The gaze data was recorded at a sampling rate of 1000 HZ using an Eyelink 1000 tower mount (SR research, Ontario, Canada). We transformed the recorded gaze data into scanpath images using the PyGaze library[19]. We created the scanpath sets as systematically as possible with the strategy of integrating different combinations of commonly used eye-tracking features - namely saccades, fixations, and AOIs - into the image with each iteration. We followed the hypothesis that the more gaze information we represented as features in the image, the higher the classification accuracy[9]. Subsequently, we investigated strategies to make the gaze features more salient for image classification models by using different coloring and fixation aggregation strategies. The following subsections detail how we represent saccades and fixations in the images along with our investigation of how to make these features more salient to image classification models. Broadly speaking, the tests fall into four categories.

The first category consists of the the baseline cases (see Fig. 3, Category 1). We compare our results against our two baseline cases, both of which involve simply plotting each recorded gaze-point. For one baseline set, we plot the raw gaze data as white colored dots onto a black background (Fig. 3 ii.). For the other baseline set, we test the effect of the game environment by illustrating the gaze as green colored dots over the gameboard (Fig. 3 iii.); we chose green as it does not occur anywhere on the background. To avoid running too many redundant experiments, we exclude the gameboards from all other scanpath design except in one test where we remodel the best

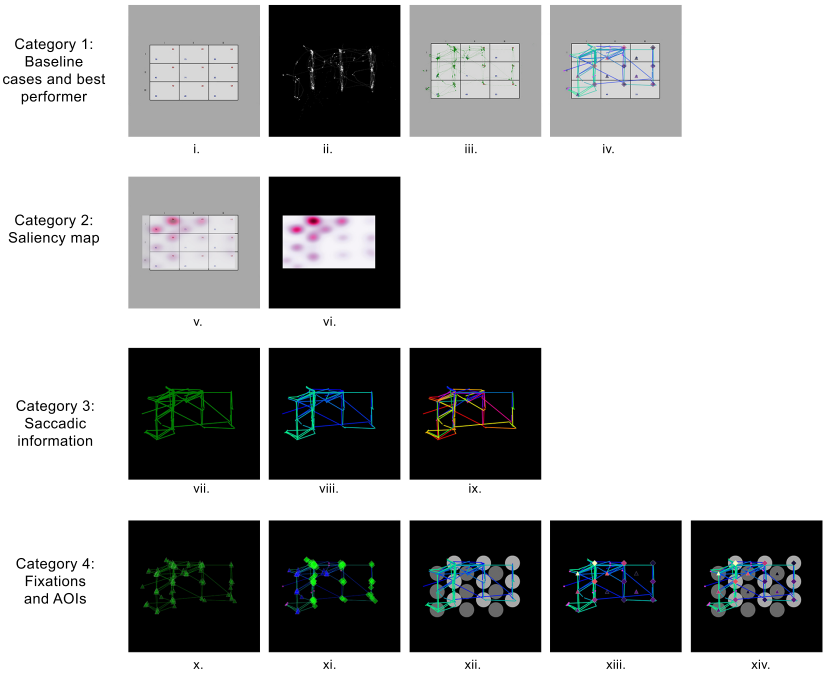

Fig. 3. Scanpath sets arranged by incorporated visual data. Category 1 (from left to right): the empty gameboard serving as the stimulus in the experiment, the raw gaze data, the raw gaze data overlayed on the gameboard image, the best performing scanpath set overlayed on the gameboard image. Category 2: a saliency map overlayed on the gameboard image, a simple saliency map with a black background. Category 3: saccadic information, sequentially-colored saccadic information, non-sequentially-colored saccadic information. Category 4: Non-aggregated fixations with saccades (with uniform shape and color), non-aggregated fixations with sequentially-colored saccades, sequentially-colored saccades over AOIs, sequentially-colored saccades with aggregated fixations, sequentially-colored saccades and aggregated fixations over AOIs.

performing scanpath design with the gameboard background included. It should be noted that we changed the opacity of the scanpath so that the gameboard is visible in this trial.

In our second category, we tested different ways of representing the fixation data as saliency maps. While saliency maps are not scanpaths, as they do not contain any temporal dimension regarding the gaze data, they have been used in the past to represent gaze behavior as an image [29, 48]. Saliency maps are created by plotting the density of fixations across an image. Since they are technically a different class of image, we decided to plot them with and without the game background for completeness (Fig. 3 v.-vi.).

In the third category, we tested different methods of representing saccadic information. In all cases, the saccadic information was plotted using linear saccades that occur between two fixations. In our first saccade design, we plotted the saccades in a single green color against a black background (Fig. 3 vii.). Next, we attempted to make the temporal dynamics of the saccades salient to the model by plotting the saccades using a sequential colormap from the matplotlib library [36](Fig. 3 viii.). Using this colormap, the lightness value increases monotonically, saccades formed at the start of the recording are plotted in a dark blue color and with later saccades being plotted in a light green color. To implement a counterfactual test, we also tested a scanpath design plotting the saccades

using a qualitative or non-sequential colormap where the colors have no order or relationship (Fig. 3 ix.).

In the fourth category, we experimented with the representation of fixations and AOIs centered on the 18 payoffs. First, we plotted all of the fixations using a single color and shape (Fig. 3 x.), then added sequential colormapping on the saccades (Fig. 3 xi.). We also displayed saccades with temporal information where we colored just the AOIs depending on if they belonged to the participants or the computers payoffs (Fig. 3 xii.) Next, to try and control overcrowding and overlapping fixations in the scanpath design, we plotted a single shape located in the center of the AOI where the fixations occurred (Fig. 3 xiii.-xiv.). We use a triangle if the participant was fixating on a payoff that he could receive and diamond for when the participant gazes at the counterpart's payoff. In this design, we again made use of sequential colormapping; this time to count the number of times an AOI was visited. Fixations that occur outside of the AOIs are represented by fuchsia dots that are 57% the size of the fixation shapes within an AOI.

3.3 Model Selection and Training Regime

To compare the effects of feature engineering in the scanpath images, we passed the datasets through a VGG-16 model pre-trained on the Imagenet dataset[70]. We chose this model due to both its popularity and ease to implement, making it natural choice for many researchers, regardless of their level of deep learning experience. The VGG-16 model was created using the Pytorch library [56], which makes the Imagenet weights available for VGG models as well as many other standard architectures. We choose Stochastic Gradient Descent (SDG) as the optimiser with a momentum of 0.9. As we are interested in comparing the impact of different representations of the input data - rather than optimising the model, we opted to train the model using a simple transfer learning strategy, freezing all of the layers of the network up to the classification layer, which we replaced to solve a binary classification problem. We choose this very vanilla training regime as it is easy to reproduce and one of the most commonly deployed training regimes for transfer learning and also produced good results in the previous eye-tracking studies [9, 11, 85].

To reduce the the possibility of a chance result, our primary experiment consists of running the VGG-16 model using 5-fold cross-validation with a 80:20 split. We set the model to run for 10 epochs each fold. We also balance the training set by under-sampling the majority class, removing a total of 28 images from the majority class. We report both the the average result across the 5-fold and the best fold for each model. We set the learning rate of 1×10^{-3} with a decay factor of 0.1. We report both the average result across folds and the best performing fold for each dataset.

As a second experiment, we split the data using a 70-20-10 train-validate-test split and ran the VGG-16 model on each set to see how it would perform on a holdout set. We split the data at the participant level to avoid any contamination leaking from the training set. Meaning, the scanpaths recorded from a given participant could only appear in a single split. During these experiments, we did not drop any of the recorded data, thus keeping a slight class imbalance, which is more indicative of real-world problems [40, 80]. Additionally, we incorporated an early-stopping mechanism into the model with a patience of 3 and set the maximum amount of epochs the model could run for to 15. However, no model reached this number of epochs before the early stopping halted the training. We used the same optimiser and learning rate as in the primary experiment.

For the final stage of analysis, we compared the results of the VGG-16 model to a Support Vector Machine (SVM) image classifier. The model is created using the popular SKlearn library [57]. Similar to its VGG-16 counterpart above, we used a 5-fold cross-validation and focused our attention on its average results. For each scanpath set, we apply an exhaustive parameter grid search to select the values of the hyperparameters from the following options: Regularization parameter: $C = \{0.1, 10, 100\}$, Kernel coefficient: $gamma = \{10, 0.1, 0.0001\}$, and a default $Kernel = \{rbf\}$. Out

of the thirteen scanpath datasets, the majority resulted in parameters $C = 10, gamma = 0.0001$, but the two variants that include only the raw gazepoints (gazeraw, gazeraw_wimg) only deviated at parameter $C = 100$. While the primary focus of our paper is not to compare various CNN models or training methods, but rather to investigate the impact of utilizing feature engineering on input data, we employed a ResNet-50 model on the best performing scanpath design to see if our results hold across another variety of deep learning model. We conducted a small ablation study comparing our favored transfer learning approach with frozen weights to the one with unfrozen weights and a random weight initialization. Our findings revealed that the frozen weights transfer learning approach produced comparable results to the unfrozen weights and random weight initialization approaches, while necessitating considerably less computational resources and much easier to devise and implement training regimes. Consequently, we elected to emphasize models with frozen weights from Imagenet and a transfer learning approach in this study. This approach strikes an ideal balance between computational efficiency and performance, is straightforward to set up, and may surprise many readers since the target dataset of scanpaths differs substantially from the domain dataset of Imagenet. Similar findings have been documented in the medical imaging domain in [62], which demonstrated that models trained on Imagenet perform similarly to custom lightweight models. We argue that this approach will serve the eye-tracking community better since deploying pre-trained models necessitates less computational expertise than creating a custom model, although future research could investigate scanpath feature engineering in the context of these lightweight models.

4 EVALUATION

We evaluated our experiment using the following metrics: Accuracy, F1-score, and Area Under the Curve (AUC) in Tables 1, 2 , 3 and 4. True Positive/ False Positive Rates are also reported in the form of confusion matrices, which can be seen in Figure 4. When considering all model tests across both classifiers, the scanpath design containing sequentially colored saccades and aggregated fixations stands out as the best-performing. However, the scanpath design containing sequentially colored saccades and AOIs proved to be a competitive adversary throughout all the tests, with almost equivalent scores across all models and, in some tests, even outperforming the scanpath with sequentially colored saccades and aggregated fixations. We chose the former to insert a gameboard underlay, as it made for a less crowded and crisper image compared to the scanpath design which includes AOIs. The colored saccades/aggregated fixations design performed second best in our primary experiment (the VGG-16 configured for k-fold cross-validation) in terms of average accuracy, with a score of 75.98%, narrowly missing out on first place to the sequentially colored saccades with AOIs design by a margin of 0.75% . Further, it achieved an average F1-score of 75.19% and AUC of 83.62%. In terms of best performing fold in our cross-validated model, this design scored third place, losing again to colored saccades with AOIs by a narrow margin of 0.37%. The first place is won by colored saccades/aggregated fixations with a gameboard underlay by a margin of 0.83%. This design also landed on the top three when using the SVM image classifier and scored the best in terms of average accuracy across the folds with a score of 73.94%, and again landed in third place in terms of the best-performing fold, losing against saliency maps with a gameboard underlay and sequentially colored saccades with non-aggregated fixations.

In terms of incorporating the gameboard image into the design of the baseline case, it performed the worst across all tests using the pre-trained VGG-16 model with a score of 51.16% in average accuracy, 54.89% in best accuracy and 70.40% in test accuracy in the model with a train-validate-test split. Regarding the SVM, it outperformed the raw gaze data plotted on a black background, but only marginally with both sets in the bottom five worst performers. Indeed in all tests, the baseline

Fig. 4. Confusion matrices representing the average performance of each scanpath dataset on a 5-fold cross-validated pre-trained VGG-16 model. Each row and figure number corresponds to the four categories and numbering as defined in Figure 3 excluding the simple gameboard (i) as it is not used as input for the models.

cases of plotting the raw gaze data on a black background or gameboard always performed badly and placed within the bottom five results.

Dataset	Accuracy	AUC	F1-Score
Saccades_Temporal_AOI$^{\Delta\star}$	**0.7674**	**0.8423**	**0.7568**
Saccades_Temporal_Fixations$^{\Delta}$	0.7598	0.8362	0.7519
Saccades_Temporal$^{\Delta}$	0.7589	0.8361	0.7510
Saccades_Temporal_Fixations_with_Background$^{\Delta}$	0.7585	0.8292	0.7480
Saccades_Temporal_Fixations_AOI$^{\Delta\star}$	0.7456	0.8242	0.7343
Saccades_Temporal_Non-Aggregated_Fixations$^{\Delta}$	0.7435	0.8131	0.7336
Saccades	0.7369	0.8126	0.7273
Saliency_Map_with_Background $^{\diamond}$	0.7361	0.8064	0.7282
Raw_Gaze	0.7223	0.7952	0.7112
Saccades_Fixations_Single_Shape_Single_Color	0.7173	0.7956	0.7015
Saccades_Temporal_NonSequential_Colormap	0.7090	0.7916	0.6993
Saliency_Map	0.7049	0.7732	0.7154
Raw_Gaze_with_Background $^{\diamond}$	0.5116	0.4964	0.0036

Table 1. Average results of a five-fold cross validation on a VGG-16 model with pre-trained weights sorted from highest to lowest accuracy. Δ denotes temporal information via a sequential color map. $\star$ AOI is included. $\diamond$ With gameboard image placed under the scanpath

The results for the trial's saliency maps were most surprising, as they performed much better than expected in the tests with the VGG-16. In all cases, the saliency maps with the background outperformed the ones on a black background. The saliency maps with the gameboard as a background scored an average accuracy of 73.61% and best accuracy of 75.88% when using the k-fold cross-validation VGG-16 model. We found this result so noteworthy, since comparable research [48] did not find a statistically significant results when attempting to resolve whether salience affects how often people choose the equilibrium strategy in two-by-two matrix games with a similar experimental structure. They used saliency maps plotted onto a game background as input to

Dataset	Accuracy	AUC	F1-Score
Saccades_Temporal_Fixations_with_Background$^{\Delta}$	**0.7879**	**0.8590**	**0.7661**
Saccades_Temporal_AOI$^{\Delta\star}$	0.7833	0.8582	0.7792
Saccades_Temporal_Fixations$^{\Delta}$	0.7796	0.8564	0.7730
Saccades_Temporal$^{\Delta}$	0.7771	0.8533	0.7757
Saccades_Temporal_Fixations_AOI$^{\Delta\star}$	0.7713	0.8470	0.7727
Saccades_Temporal_Non-Aggregated_Fixations$^{\Delta}$	0.7692	0.8287	0.7589
Saccades	0.7651	0.8299	0.7490
Saliency_Map_with_Background $^{\diamond}$	0.7588	0.8231	0.7495
Saliency_Map	0.7542	0.8098	0.7511
Raw_Gaze	0.7354	0.8014	0.7406
Saccades_Fixations_Single_Shape_Single_Color	0.7277	0.8074	0.7265
Saccades_Temporal_NonSequential_Colormap	0.7256	0.8126	0.7295
Raw_Gaze_with_Background $^{\diamond}$	0.5489	0.5174	0.0181

Table 2. Best results of a five-fold cross validation on a VGG-16 model with pre-trained weights sorted from highest to lowest accuracy. Δ denotes temporal information via a sequential color map. $\star$ AOI is included. $\diamond$ With gameboard image placed under the scanpath

a Salience Attentive Model (SAM), As SAMs are usually pre-trained models that fine-tuned on open-access salience datasets such as SALICON [39] this raises questions over how specific training regimes impact model performance when analysing eye-tracking data.

In our primary experiment, we compare the average accuracy across each fold of the VGG-16 configured for k-fold cross-validation model with each different scanpath set as input. The top four performing scanpaths all scored within one percent of each other, with the best score of 76.72%, which outperforms the baseline case of the raw gaze data with the gameboard as a background by over 25%. All top four scanpath designs contain saccades created with the temporal colormap. In all cases, a scanpath design that incorporated some fixation features in combination with temporal saccades became the best-performing model. However, it remains unclear from this experiment how to best represent these fixations because, depending on the experiment, the AOIs with aggregated and non-aggregated fixations all exhibited the highest accuracy. The full results can be seen in Table 1. Our hypothesis that using sequential colouring to helps form meaningful representation for the model is supported because the performance becomes worse when saccadic information is encoded using non-sequential colormaps. These scanpaths performed worse than the baseline raw gaze data in every test using the VGG-16 model, suggesting that convolutional filter extracts meaning from the colorschemes. However, further research is needed to confirm these findings.

Figure 4 shows the confusion matrices based on the average results of all 13 scanpath sets from the primary experiment. Here, it is clear to see that raw gaze data with background image (See Figure 3 iii.) performed the worst with TPR only at 0.18%. While the raw gaze (ii), non-sequentially colored saccades (ix) and single-colored saccades with non-aggregated fixations (x) performed relatively better, they still rank lower compared to scanpath sets containing meaningful representation.

In a second experiment, we trained the VGG-16 model and tested it on a holdout set the design with temporal saccades, aggregated fixations and AOIs scored highest with an accuracy of 78.80%, AUC 87.93%, and an F1-score of 76.23%, making it an excellent classifier by all standards. The results are largely equivalent with the best result for each set from the VGG-16 model using K-fold cross-validation, as seen in Table 3, and Table 2, with the exception being the raw gaze data with a gameboard background, where we see over a 15% decrease in accuracy dropping from 70.40% to 51.16%.

The 5-fold cross-validated SVM model generally yielded comparatively worse results than the VGG-16 variant as well as with a train-validation-test split evaluated on the VGG-16 model. The SVM shows a maximum average accuracy of 73.94% from the scanpath set that includes sequentially colored saccades and aggregated fixations and the maximum best accuracy of 76.79% from

Dataset	Accuracy	AUC	F1-Score
Saccades_Temporal_Fixations_AOI	**0.7880**	**0.8792**	**0.7623**
Saccades_Temporal$^{\Delta}$	0.7880	0.8582	0.7535
Saccades_Temporal_AOI$^{\Delta\star}$	0.7760	0.8583	0.7407
Saccades_Temporal_Fixations_with_Background$^{\Delta}$	0.7680	0.8612	0.7563
saccades	0.7560	0.8317	0.7382
Saliency_Map_with_Background $^{\diamond}$	0.7440	0.8336	0.7168
Saccades_Temporal_Non-Aggregated_Fixations$^{\Delta}$	0.7400	0.8596	0.7368
Saccades_Temporal_Fixations$^{\Delta}$	0.7400	0.8719	0.7410
Saccades_Fixations_Single_Shape_Single_Color	0.7400	0.8387	0.7257
Saliency_Map	0.7360	0.7971	0.7402
Saccades_Temporal_NonSequential_Colormap	0.7280	0.8291	0.7302
Raw_Gaze	0.7160	0.7968	0.7102
Raw_Gaze_with_Background $^{\diamond}$	0.7040	0.8501	0.7176

Table 3. Results of a VGG-16 model with pre-trained weights tested on a hold-out set sorted from highest to lowest accuracy. Δ denotes temporal information via a sequential color map. ⋆ AOI is included. ⋄ With gameboard image placed under the scanpath

sequentially colored saccades with non-aggregated fixations. In comparison, the cross-validated VGG-16, achieved a higher average accuracy by about 3% and a marginally higher best accuracy of about 2%.

Comparing the k-fold cross-validation on the VGG-16 model and SVM in terms of the F1-score, the scanpath set with sequentially colored saccades and AOIs is on top for VGG-16 with an average accuracy of 75.68% and best accuracy off 77.92% and remains in the top five for both the average and best results of the SVM. The dataset with sequentially colored saccades, aggregated fixations, and a gameboard background takes first place in the SVM results with an average accuracy of 73.63% and best accuracy of 76.41% while similarly still remaining in the top five from the VGG-16 results. The dataset containing sequential saccades, aggregated fixations as well as AOIs also consistently place in the top 5 for both SVM and VGG-16 tests.

Additionally, we performed a small ablation analysis on the best performing scanpath design. We used the scanpath design that was the sequentially colored saccades and aggregated fixations, to test a VGG-16 and a ResNet-50 model under various initialization conditions, such as random initialization and models starting with Imagenet weights with frozen and unfrozen layers. We followed the methodology of our second experiment, training each model and evaluating its performance on a holdout set. Although our paper primarily focuses on the impact of feature engineering on performance, we conducted this additional analysis to emphasize that our transfer learning strategy, can be an effective training regime for scanpath images. The VGG model pretrained on Imagenet with frozen layers returned an accuracy of 0.7880. The model with random weight initialization performs almost equivalently with a score of 0.7800. When running the model with unfrozen layers meaning that model can adjust the weights and biases, we see a drop of performance decrease of 6% to 0.7280. Moving to the ResNet-50 – another popular architecture, we see less stable results, especially when moving to random weights hence further supporting our strategy choice. The following accuracies can be reported for the Resnet-50 Model 0.7680 (unfrozen), 0.6840 (frozen) and 0.6280 (random initialization). Our findings are in line with previous research, for instance, [6] used scanpaths with a comparable design and from a similar-sized dataset into multiple CNN architectures and found that the VGG architecture slightly outperformed the rest.

5 DISCUSSION

In this paper, we demonstrate that transforming the raw gaze data into saccades and fixations can greatly boost the performance of machine learning models. While this result is not surprising as

Dataset	Accuracy	AUC	F1-Score
Saccades_Temporal_Fixations	**0.7394**	**0.7390**	**0.7300**
Saliency_Map_with_Background$^{\diamond}$	0.7385	0.7398	0.7354
Saccades_Temporal_Fixations_AOI$^{\Delta\star}$	0.7381	0.7383	0.7336
Saccades_Temporal_NonAggregated_Fixations$^{\Delta}$	0.7369	0.7370	0.7318
Saccades_Fixations_Single_Shape_Single_Color	0.7369	0.7371	0.7315
Saccades_Temporal_Fixations_with_Background$^{\Delta}$	0.7294	0.7304	0.7363
Saccades_Temporal_AOI$^{\Delta\star}$	0.7273	0.7274	0.7156
Saccades_Temporal$^{\Delta}$	0.7156	0.7161	0.7027
Saccades	0.7140	0.7144	0.7051
Raw_Gaze_with_Background$^{\diamond}$	0.7115	0.7109	0.6894
Saccades_Temporal_NonSequential_Colormap$^{\Delta}$	0.7086	0.7085	0.6990
Saliency_Map	0.7073	0.7070	0.7001
Raw_Gaze	0.7069	0.7062	0.6870

Table 4. Average results of a five-fold cross validation on a simple SVM model sorted from highest to lowest accuracy. Δ denotes temporal information via a sequential color map. $\star$ AOI is included. $\diamond$ With gameboard image placed under the scanpath

feature engineering has been shown to boost model performance in other domains [38, 64, 67], it highly supports how traditional eye-tracking metrics for feature processing impacts the results of our model experiments. Our results suggested that a combination of carefully thought-out representations of saccades and fixations that fit the experimental task produce the best classification results.

Using image classifiers to analyse eye-tracking data has many potential benefits when compared to more common approaches such as processing the gaze data as a sequence. For instance, it avoids any issues surrounding sequence padding to handle uneven sequences, and allows the context that the subject is viewing to be easily integrated into the analyses. Deep learning image classifiers come with the benefit that there exists many standard architectures such as VGG [70], Resnets[31] and Vision Transformers [21], allowing researches to try a plethora of different architectures in order to find which one best suits their needs. Another rather counter-intuitive benefit of using image classifiers is that it has now been shown multiple times that pre-training networks on large datasets such as Imagenet can be a successful strategy when classifying scanpaths [9, 85]. This result has quite an impact, yet is not entirely unexpected. A similar phenomenon is well documented across the medical imaging domain where the source and target datasets also greatly differ [68]. Thus employing this type of training regime can provide a good starting point for eye tracking researchers looking to apply deep learning methods to analyse scanpaths.

Eye tracking research is a multidisciplinary pursuit with researchers from diverse fields as psychology and linguistics to computer science and physics. This brings a large variance in the technical abilities of researchers operating in this field. We do not intend for this paper to be viewed as an exhaustive list of how to build scanpaths for image classification models, or to take away from the merits of using a well-defined model architecture, but rather to demonstrate a strategy of iterating through scanpath features with a combination of knowledge on the task that may help less technically inclined researchers in eye-tracking reap the benefits of image classification models.

6 LIMITATIONS AND FUTURE RESEARCH

Our study has limitations. First, the dataset we used for this study was handpicked as it provides a much sparser environment than most eye-tracking datasets meaning that the constructed features are more pronounced than they may be in a dataset that contains natural images. This sparseness may contribute to the improvements in accuracy as the image becomes fuller so to speak, as we include more eye-tracking features. Second, our list of engineered features is not exhaustive and

there may well be scanpath configurations that yields better results. In future studies, we aim to test the effects of feature engineering in the design of scanpath images on multiple datasets for both free and task-based viewing in order to come up with some guiding principles as currently this work only provides the reader with an example case. Third, we did not explore all of the many different types of image classifiers such as Vision Transformers that may perform better without any feature engineering. Finally, another avenue that needs to be explored in future research involves how to best optimise a training regime that could impact the model performance on any of the given scanpath sets containing engineered features.

7 CONCLUSION

In this paper we demonstrated that by using feature engineering techniques stemming from domain knowledge of general eye-tracking research and task specific knowledge, we were able to create scanpath images that outperformed the baseline cases in terms of accuracy, F1-score, and AUC. Furthermore, we showcase feature engineering strategies, such as using sequential coloring and aggregation techniques, that can further boost performance. Additionally, the results from our experiments illustrate that sub-optimal feature engineering strategies, such as the non-sequential coloring of saccades can lead to a performance decrease compared to plotting the raw data onto an image. As image classification models and machine learning becomes more prevalent in eye-tracking research, we demonstrate how domain knowledge can greatly complement these models, as they become more accessible to everyone in the field.

8 ACKNOWLEDGMENTS

The authors would like to thank Momchil Yordanov & Steve Borchardt for their valuable contributions and suggestions throughout this project.

REFERENCES

[1] Zeyad AT Ahmed and Mukti E Jadhav. 2020. Convolutional Neural Network for Prediction of Autism based on Eye-tracking Scanpaths. *International Journal of Psychosocial Rehabilitation* 24, 05 (2020).

[2] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine* 34, 6 (2017), 26–38.

[3] Adham Atyabi, Frederick Shic, Jiajun Jiang, Claire E Foster, Erin Barney, Minah Kim, Beibin Li, Pamela Ventola, and Chung Hao Chen. 2022. Stratification of Children with Autism Spectrum Disorder through fusion of temporal information in eye-gaze scan-paths. *ACM Transactions on Knowledge Discovery from Data (TKDD)* (2022).

[4] Imon Banerjee, Yuan Ling, Matthew C Chen, Sadid A Hasan, Curtis P Langlotz, Nathaniel Moradzadeh, Brian Chapman, Timothy Amrhein, David Mong, Daniel L Rubin, et al. 2019. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artificial intelligence in medicine* 97 (2019), 79–88.

[5] Michael Barz and Daniel Sonntag. 2016. Gaze-guided object classification using deep neural networks for attention-based computing. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 253–256.

[6] Nilavra Bhattacharya, Somnath Rakshit, Jacek Gwizdka, and Paul Kogut. 2020. Relevance prediction from eye-movements using semi-interpretable convolutional neural networks. In *Proceedings of the 2020 conference on human information interaction and retrieval*. 223–233.

[7] Jonathan FG Boisvert and Neil DB Bruce. 2016. Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing* 207 (2016), 653–668.

[8] Christian Braunagel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2017. Ready for take-over? A new driver assistance system for an automated classification of driver take-over readiness. *IEEE Intelligent Transportation Systems Magazine* 9, 4 (2017), 10–22.

[9] Sean Anthony Byrne, Adam Peter Frederick Reynolds, Carolina Biliotti, Falco J Bargagli-Stoffi, Luca Polonio, and Massimo Riccaboni. 2023. Predicting choice behaviour in economic games using gaze data encoded as scanpath images. *Scientific Reports* 13, 1 (2023), 4722.

[10] Roberto Caldara and Sébastien Miellet. 2011. iMap: a novel method for statistical fixation mapping of eye movement data. *Behavior Research Methods* 43, 3 (2011), 864–878.

[11] Nora Castner, Thomas C Kuebler, Katharina Scheiter, Juliane Richter, Therese Eder, Fabian Huettig, Constanze Keutel, and Enkelejda Kasneci. 2020. Deep Semantic Gaze Embedding and Scanpath Comparison for Expertise Classification during OPT Viewing. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Full Papers)*. Association for Computing Machinery, New York, NY, USA, Article 18, 10 pages. 9781450371339 https://doi.org/10.1145/3379155.3391320

[12] Moran Cerf, Jonathan Harel, Alex Huth, Wolfgang Einhäuser, and Christof Koch. 2008. Decoding what people see from where they look: Predicting visual stimuli from scanpaths. In *International Workshop on Attention in Cognitive Systems*. Springer, 15–26.

[13] Shi Chen and Qi Zhao. 2019. Attention-based autism spectrum disorder screening with privileged modality. In *Proceedings of the IEEE International Conference on Computer Vision*. 1181–1190.

[14] François Chollet et al. 2015. Keras. https://keras.io.

[15] Yandre MG Costa, Luiz S Oliveira, and Carlos N Silla Jr. 2017. An evaluation of convolutional neural networks for music classification using spectrograms. *Applied soft computing* 52 (2017), 28–38.

[16] Miguel Costa-Gomes, Vincent P. Crawford, and Bruno Broseta. 2001. Cognition and Behavior in Normal-Form Games: An Experimental Study. *Econometrica* 69, 5 (2001), 1193–1235. https://doi.org/10.1111/1468-0262.00239 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00239

[17] Antoine Coutrot, Janet H Hsiao, and Antoni B Chan. 2018. Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods* 50, 1 (2018), 362–379.

[18] David P Crabb, Nicholas D Smith, and Haogang Zhu. 2014. What's on TV? Detecting age-related neurodegenerative eye disease using eye movement scanpaths. *Frontiers in Aging Neuroscience* 6 (2014), 312.

[19] Edwin Dalmaijer, Sebastiaan Mathôt, and Stefan Stigchel. 2013. PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior Research Methods* 46 (11 2013). https://doi.org/10.3758/s13428-013-0422-2

[20] Giovanna Devetag, Sibilla Di Guida, and Luca Polonio. 2016. An eye-tracking study of feature-based choice in one-shot games. *Experimental Economics* 19, 1 (2016), 177–201.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/ARXIV.2010.11929

[22] Cao Vu Dung, Hidehiko Sekiya, Suichi Hirano, Takayuki Okatani, and Chitoshi Miki. 2019. A vision-based method for crack detection in gusset plate welded joints of steel bridges using deep convolutional neural networks. *Automation in Construction* 102 (2019), 217–229.

[23] Mahmoud Elbattah, Romuald Carette, Gilles Dequen, Jean-Luc Guérin, and Federica Cilia. 2019. Learning clusters in autism spectrum disorder: Image-based clustering of eye-tracking scanpaths with deep autoencoder. In *2019 41st Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 1417–1420.

[24] Ramin Fahimi and Neil DB Bruce. 2021. On metrics for measuring scanpath similarity. *Behavior Research Methods* 53, 2 (2021), 609–628.

[25] Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. 2020. How much time do you have? modeling multi-duration saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4473–4482.

[26] Drew Fudenberg and David K Levine. 2016. Whither game theory? Towards a theory of learning in games. *Journal of Economic Perspectives* 30, 4 (2016), 151–70.

[27] Wolfgang Fuhl, Efe Bozkir, Benedikt Hosp, Nora Castner, David Geisler, Thiago C Santini, and Enkelejda Kasneci. 2019. Encodji: encoding gaze data into emoji space for an amusing scanpath classification approach. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 1–4.

[28] Wolfgang Fuhl, TC Kübler, Katrin Sippel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2015. Arbitrarily shaped areas of interest based on gaze density gradient. In *European Conference on Eye Movements*, Vol. 1. 5.

[29] Wolfgang Fuhl, Thomas Kuebler, Thiago Santini, and Enkelejda Kasneci. 2018. Automatic generation of saliency-based areas of interest for the visualization and analysis of eye-tracking data. In *Proceedings of the Conference on Vision, Modeling, and Visualization*. 47–54.

[30] David Geisler, Daniel Weber, Nora Castner, and Enkelejda Kasneci. 2020. Exploiting the GBVS for Saliency Aware Gaze Heatmaps. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Short Papers)*. Association for Computing Machinery, New York, NY, USA, Article 24, 5 pages. 9781450371346 https://doi.org/10.1145/3379156.3391367

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/ARXIV.1512.03385

[32] Helene Hembrooke, Matt Feusner, and Geri Gay. 2006. Averaging scan patterns and what they can tell us. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*. 41–41.

[33] John Heminghous and Andrew T Duchowski. 2006. iComp: a tool for scanpath visualization and comparison. In *Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization*. 152–152.

[34] Benedikt Hosp, Florian Schultz, Enkelejda Kasneci, and Oliver Höner. 2021. Expertise classification of soccer goalkeepers in highly dynamic decision tasks: a deep learning approach for temporal and spatial feature recognition of fixation image patch sequences. *Frontiers in Sports and Active Living* (2021), 183.

[35] Benedikt Hosp, Myat Su Yin, Peter Haddawy, Ratthapoom Watcharopas, Paphon Sa-Ngasoongsong, and Enkelejda Kasneci. 2021. Differentiating Surgeons' Expertise solely by Eye Movement Features. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*. 371–375.

[36] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 3 (2007), 90–95. https://doi.org/10.1109/MCSE.2007.55

[37] Poika Isokoski, Jari Kangas, and Päivi Majaranta. 2018. Useful approaches to exploratory analysis of gaze data: enhanced heatmaps, cluster maps, and transition maps. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. 1–9.

[38] Dipendra Jha, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. 2018. Elemnet: Deep learning the chemistry of materials from only elemental composition. *Scientific reports* 8, 1 (2018), 1–13.

[39] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. SALICON: Saliency in Context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[40] Justin Johnson and Taghi Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6 (03 2019), 27. https://doi.org/10.1186/s40537-019-0192-5

[41] Tae Joon Jun, Hoang Minh Nguyen, Daeyoun Kang, Dohyeun Kim, Daeyoung Kim, and Young-Hak Kim. 2018. ECG arrhythmia classification using a 2-D convolutional neural network. https://doi.org/10.48550/ARXIV.1804.06812

[42] Juraj Kacur, Jaroslav Polec, Eva Smolejova, and Anton Heretik. 2020. An analysis of eye-tracking features and modelling methods for free-viewed standard stimulus: application for schizophrenia detection. *IEEE Journal of Biomedical and Health Informatics* 24, 11 (2020), 3055–3065.

[43] Daniel T. Knoepfle, Colin F. Camerer, and Joseph Tao yi Wang. 2009. Studying Learning in Games Using Eye-Tracking. *Journal of the European Economic Association* 7, 2/3 (2009), 388–398. 15424766, 15424774 http://www.jstor.org/stable/40282757

[44] Michal Krol and Magdalena Krol. 2017. A novel approach to studying strategic decisions with eye-tracking and machine learning. *Judgment and Decision Making* 12, 6 (2017), 596.

[45] Ayush Kumar, Prantik Howlader, Rafael Garcia, Daniel Weiskopf, and Klaus Mueller. 2020. Challenges in interpretability of neural networks for eye movement data. In *ACM Symposium on Eye Tracking Research and Applications*. 1–5.

[46] Kuno Kurzhals. 2021. Image-based projection labeling for mobile eye tracking. In *ACM Symposium on Eye Tracking Research and Applications*. 1–12.

[47] Olivier Le Meur and Thierry Baccino. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods* 45, 1 (2013), 251–266.

[48] Xiaomin Li and Colin Camerer. 2020. Predictable Effects of Bottom-up Visual Salience in Experimental Decisions and Games. *Available at SSRN 3308886* (2020).

[49] Xiaomin Li and Colin Camerer. 2021. Hidden Markov Modeling of the Cognitive Process in Strategic Thinking. *Available at SSRN 3838911* (2021).

[50] Alexander Lotz and Sarah Weissenberger. 2018. Predicting take-over times of truck drivers in conditional autonomous driving. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 329–338.

[51] Davide Marchiori, Sibilla Di Guida, and Luca Polonio. 2021. Plasticity of strategic sophistication in interactive decision-making. *Journal of Economic Theory* 196 (2021), 105291.

[52] Amitha Mathew, P Amudha, and S Sivakumari. 2020. Deep learning techniques: an overview. In *International conference on advanced machine learning technologies and applications*. Springer, 599–608.

[53] Jojo Moolayil, Jojo Moolayil, and Suresh John. 2019. *Learn Keras for deep neural networks*. Springer.

[54] John F Nash Jr. 1950. Equilibrium points in n-person games. *Proceedings of the national academy of sciences* 36, 1 (1950), 48–49.

[55] Nabil Ouerhani, Heinz Hügli, René Müri, and Roman Von Wartburg. 2003. Empirical validation of the saliency-based model of visual attention. In *Electronic Letters on Computer Vision and Image Analysis*. 13–23.

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. https://doi.org/10.48550/ARXIV.1912.01703

[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[58] Luca Polonio and Giorgio Coricelli. 2019. Testing the level of consistency between choices and beliefs in games using eye-tracking. *Games and Economic Behavior* 113 (2019), 566–586.

[59] Luca Polonio, Sibilla Di Guida, and Giorgio Coricelli. 2015. Strategic sophistication and attention in games: An eye-tracking study. *Games and Economic Behavior* 94 (2015), 80–96.

[60] KN Praveena and R Mahalakshmi. 2022. Classification of Autism Spectrum Disorder and Typically Developed Children for Eye Gaze Image Dataset using Convolutional Neural Network. *International Journal of Advanced Computer Science and Applications* 13, 3 (2022).

[61] Claudio M. Privitera and Lawrence W. Stark. 2000. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 9 (2000), 970–982.

[62] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. 2019. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems* 32 (2019).

[63] Umesh Rajashekar, Ian Van Der Linde, Alan C Bovik, and Lawrence K Cormack. 2008. GAFFE: A gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing* 17, 4 (2008), 564–573.

[64] Tara Rawat and Vineeta Khemchandani. 2017. Feature engineering (FE) tools and techniques for better classification performance. *International Journal of Innovations in Engineering and Technology* 8, 2 (2017), 169–179.

[65] Monika Roopak, Gui Yun Tian, and Jonathon Chambers. 2019. Deep learning models for cyber security in IoT networks. In *2019 IEEE 9th annual computing and communication workshop and conference (CCWC)*. IEEE, 0452–0457.

[66] Hosnieh Sattar, Mario Fritz, and Andreas Bulling. 2020. Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations. *Neurocomputing* 387 (2020), 369–382.

[67] Devarshi Shah, Jin Wang, and Q Peter He. 2020. Feature engineering in big data analytics for IoT-enabled smart manufacturing–Comparison between deep learning and statistical learning. *Computers & Chemical Engineering* 141 (2020), 106970.

[68] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging* 35, 5 (2016), 1285–1298.

[69] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034 [cs.CV]

[70] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[71] Shane D Sims and Cristina Conati. 2020. A neural architecture for detecting user confusion in eye-tracking data. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 15–23.

[72] Mikhail Startsev and Michael Dorr. 2019. Classifying autism spectrum disorder based on scanpaths and saliency. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 633–636.

[73] Florian Strohm, Ekta Sood, Sven Mayer, Philipp Müller, Mihai Bâce, and Andreas Bulling. 2021. Neural Photofit: Gaze-based Mental Image Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 245–254.

[74] Yusuke Sugano, Yasunori Ozaki, Hiroshi Kasai, Keisuke Ogaki, and Yoichi Sato. 2014. Image preference estimation with a data-driven approach: A comparative study between gaze and image features. *Journal of Eye Movement Research* 7, 3 (2014).

[75] Y. Tao and M. Shyu. 2019. SP-ASDNet: CNN-LSTM Based ASD Classification Model using Observer ScanPaths. In *2019 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*. 641–646. https://doi.org/10.1109/ICMEW.2019.00124

[76] Po-He Tseng, Ian GM Cameron, Giovanna Pari, James N Reynolds, Douglas P Munoz, and Laurent Itti. 2013. High-throughput classification of clinical populations from natural viewing eye movements. *Journal of Neurology* 260, 1 (2013), 275–284.

[77] Pranav Venuprasad, Li Xu, Enoch Huang, Andrew Gilman, Leanne Chukoskie Ph. D, and Pamela Cosman. 2020. Analyzing gaze behavior using object detection and unsupervised clustering. In *ACM Symposium on Eye Tracking Research and Applications*. 1–9.

[78] Lisa-Marie Vortmann, Jannes Knychalla, Sonja Annerer-Walcher, Mathias Benedek, and Felix Putze. 2021. Imaging Time Series of Eye Tracking Data to Classify Attentional States. *Frontiers in Neuroscience* 15 (2021), 664490.

[79] Stephen Anthony Waite, Arkadij Grigorian, Robert G Alexander, Stephen Louis Macknik, Marisa Carrasco, David Heeger, and Susana Martinez-Conde. 2019. Analysis of perceptual expertise in radiology–Current knowledge and a new perspective. *Frontiers in Human Neuroscience* 13 (2019), 213.

[80] Shoujin Wang, Wei Liu, Jia Wu, Longbing Cao, Qinxue Meng, and Paul J. Kennedy. 2016. Training deep neural networks on imbalanced data sets. In *2016 International Joint Conference on Neural Networks (IJCNN)*. 4368–4374. https://doi.org/10.1109/IJCNN.2016.7727770

[81] Zhiguang Wang and Tim Oates. 2015. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.

[82] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3, 1 (2016), 1–40.

[83] Julian Wolf, Stephan Hess, David Bachmann, Quentin Lohmeyer, and Mirko Meboldt. 2018. Automating areas of interest analysis in mobile eye tracking experiments based on machine learning. *Journal of Eye Movement Research* 11, 6 (2018).

[84] David S Wooding. 2002. Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers* 34, 4 (2002), 518–528.

[85] Yuehan Yin, Yahya Alqahtani, Jinjuan Heidi Feng, Joyram Chakraborty, and Michael P McGuire. 2021. Classification of eye tracking data in visual information processing tasks using convolutional neural networks and feature engineering. *SN Computer Science* 2, 2 (2021), 1–26.

[86] Yitan Zhu, Thomas Brettin, Fangfang Xia, Alexander Partin, Maulik Shukla, Hyunseung Yoo, Yvonne A Evrard, James H Doroshow, and Rick L Stevens. 2021. Converting tabular data into images for deep learning with convolutional neural networks. *Scientific reports* 11, 1 (2021), 1–11.

Received November 2022; revised February 2023; accepted March 2023