Confusion Prediction from Eye-Tracking Data: Experiments with Machine Learning

Joni Salminen
Qatar Computing Research Institute, HBKU; and Turku School of Economics
Doha, Qatar
jsalminen@hbku.edu.qa

Haewoon Kwak
Qatar Computing Research Institute, HBKU
Doha, Qatar
hkwak@hbku.edu.qa

Soon-gyo Jung
Qatar Computing Research Institute, HBKU
Doha, Qatar
sjung@hbku.edu.qa

Mridul Nagpal
International Institute of Information Technology
Hyderabad, India
mridulnagpal07@gmail.com

Jisun An
Qatar Computing Research Institute, HBKU
Doha, Qatar
jan@hbku.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute, HBKU
Doha, Qatar
bjansen@hbku.edu.qa

ABSTRACT

Predicting user confusion can help improve information presentation on websites, mobile apps, and virtual reality interfaces. One promising information source for such prediction is eye-tracking data about gaze movements on the screen. Coupled with think-aloud records, we explore if user's confusion is correlated with primarily fixation-level features. We find that random forest achieves an accuracy of more than 70% when prediction user confusion using only fixation features. In addition, adding user-level features (age and gender) improves the accuracy to more than 90%. We also find that balancing the classes before training improves performance. We test two balancing algorithms, Synthetic Minority Over Sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) finding that SMOTE provides a higher performance increase. Overall, this research contains implications for researchers interested in inferring users' cognitive states from eye-tracking data.

KEYWORDS

Eye tracking, confusion detection, machine learning

1 INTRODUCTION

Predicting confusion of users is important for improving information presentation on websites, mobile displays, and virtual reality (VR) interfaces [1]. In eye tracking user studies, it is common to record the cognitive states of the participants, such as confusion about presented information, via the think-aloud technique [2], which gives researchers an understanding of users' state of the mind [3]. Additionally, coding the think-aloud records for confusion provides a clear dependent variable for predictive modeling using machine learning. Fixation variables, such as scanpath length and duration, have been observed to reflect users' information processing and mental states [4][5] [6].

In this research, we use fixation data, i.e., the quantity, duration, accuracy, and position of eye fixations, to model how this data relates to users' expressed confusion. Previous studies have found that user attributes, such as gender, age, and level of experience are impactful for individuals' viewing behavior [7] [8] [9]. However, such user-level data is often unavailable due to factors such as privacy concerns. Moreover, fixation patterns of users tend to be noisy [10], especially when associated with cognitive states of the user [11]. The opportunity for neural networks and machine learning in this context lies in uncovering the hidden complexity of fixation paths (i.e., eye movements from one area of the screen to another). The number of fixations from an eye-tracking trial tends toward thousands of fixations, depending on task complexity and trial duration, and the individual paths are highly dissimilar [7] [10]. Therefore, finding patterns using manual analysis is extremely difficult [12] [13], whereas a machine learning model could discover hidden patterns.

In this research, we aim to predict user confusion by using fixation information from a user study as features for machine learning and compare the performance to a model using both participant level information and the fixation information. Our research questions are:

1. *Given only the fixation features, how well can we predict users' perceived confusion?*
2. *How does adding user information impact the prediction accuracy?*
3. *Can the model discriminate between confusion and non-confusion at an AOI-level?*

The research problem is not trivial because the fixation patterns tend to be complex, consisting of users' gaze jumping

ICIST '19, March 24–26, 2019, Cairo, Egypt

ACM ISBN 978-1-4503-6292-4/19/03...$15.00
DOI: 10.1145/3361570.3361577

from one area of the screen to another in a sporadic fashion. Beyond basic metrics, such as number and duration of fixations, one also should consider the sequence of AOIs that intuitively should matter for the prediction [7], but it is generally complex and unpredictable. The problem lies in getting from the complex sequence and duration representation of confused users to a predictive pattern. Our premise is that the users' eye fixation pattern should reveal they are confused, but this pattern is not easily captured by traditional means of analysis.

2 RELATED LITERATURE

Earlier research has shown that eye-tracking features, such as duration spent fixated on an AOI vary by use case and user. While longer fixation duration can indicate confusion in information retrieval tasks [12], for tasks such as online shopping, it can indicate higher engagement [14]. The solution for distinguishing between varying cognitive states is to measure them separately and then associate the states with eye-tracking data. Applied approaches include e.g. neurophysiological methods (e.g., EEG), or utterance-based coding (e.g., concurrent think aloud) [15].

However, the nature of the problem implies that general rules about the relationship between fixation patterns and confusion cannot be easily formulated. Rather, we suggest that such relationships are better off being predicted from the data, given that we have labeled data on confusion, such as the one we obtain in this study. For machine learning purposes, this data can be very helpful, as it provides a dependent variable for predictive models.

Earlier works combining neural networks with eye-tracking data focus on two areas: 1) gaze detection and 2) prediction of mental states. For example, Tan et al. [16] use neural networks for gaze detection using facial features. Kim and Kang [17] create a neural network for tracking users' gaze in difficult background situations. Coughlin et al. [18] create neural networks for automatic calibration of eye-tracking. Other gaze detection works include e.g. [19][20][21] [22] [23]. Moreover, preliminary studies have shown promise for learning cognitive events from eye-tracking data [24]. For example, Harada et al. [25] use neural networks to detect the level of distraction from drivers, and Kuperberg and Heckers [26] and Campana et al. [27] use neural networks for classifying schizophrenia.

In summary, while previous works have applied machine learning to eye-tracking data, we could locate no prior research specifically trying to predict user confusion with machine learning. Yet, confusion is a key issue in user interface design [28] and human-computer interaction [29]. Especially the proliferation of the low-cost eye-tracking devices encourages researchers and organizations to collect user data by conducting eye-tracking studies [30], which highlights the importance of developing new modelling approaches to analyzing the actual eye-tracking data.

3 DATA COLLECTION

3.1 Experiment Set-Up

We conducted an eye-tracking user study to test different layouts of an automatic persona generation (APG) system persona profile, with APG reported in [31]–[37]. APG is both a methodology for creating persona automatically from online analytics data [38] [31] [39]. Figure 1 shows an example of such a persona. Personas are profile representations of core customers of an organization [40] and they are based on real user data [41].

In previous research, persona profiles have been found to risk having inconsistency between different information elements [42] as well as risk for confusion [13], [43], [44]. Because of these challenges, also automatically generated persona profiles might raise *confusion* (defined as a *perceived state of disorientation by the end user*) among the users, especially since data is being retrieved from many sources into one persona profile [42]. Thus, we investigate how different persona layouts (Figure 2) affect the confusion of the users of personas toward the persona.

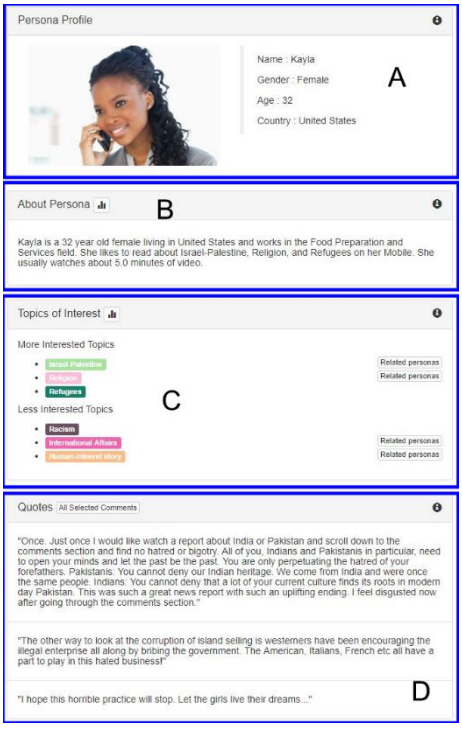

Figure 1: An automatically generated persona with a picture, name, and demographic information [A], text description [B], topics of interest [C], and quotes [D].

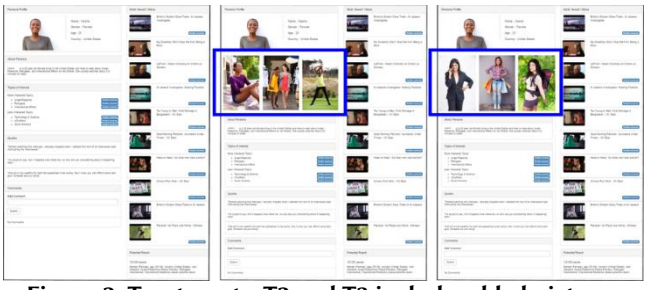

Figure 2: Treatments. T2 and T3 include added pictures (highlighted by blue boxes). We expected the extra pictures to increase confusion because they are not directly related to the persona. The difference between T2 and T3 is that T3 involves pictures of people like the persona (by age and gender), whereas T2 involves pictures of the same person (corresponding to the persona).

To know how the choice of images affects confusion, we carried out an eye-tracking experiment in the premises of an international news and media company that uses personas to increase content producers' customer understanding. The 29 participants included journalists and editors. The average age of the participants was 32.6 years, and prior experience in the news industry 7.3 years. There were 15 males, 14 females.

We set up the EyeTribe tracker, in the organization's premises and conducted eye-tracking trials with the participants. Each participant was administered three treatments (Figure 2) while completing a journalistic task which involved writing a story for the persona. Each session lasted between 20 to 40 minutes, depending on the speed of the participant. We did not set a constraint for viewing time; each treatment was viewed approximately for five minutes. The sequence of showing treatments T1-T3 was counterbalanced, so that each treatment was shown an equal number first, middle, and last.

3.2 Eye Tracking and Confusion Coding

The fixation data was recorded with a sampling rate of 50Hz that grows the number of observations rapidly even with a small number of participants. Here, 29 participants yielded more than 53,000 fixation observations targeting different areas of interest (AOIs) on the screen. AOIs were encoded by the researchers to represent key information in the persona profiles (see Figure 3).

During the experiment, we encouraged participants to actively describe what information they see and how they feel about it while carrying out their tasks. We recorded these think-aloud comments during the experiment and analyzed them using the *cognitive discourse analysis* (CDA) technique by Tenbrink [45]. Following CDA, confusion was coded based on its verbalized cues. For example, a participant might say: "*lost on here; conflicted profile*," "*if I were to pitch based on this profile, I'm confused*," and so on. Such indications of confusion resulted in the confusion of the trial labeled as 1 (true), and lack thereof as 0 (false). For example, P1: Confusion T1 = 1 means that Participant 1 was confused during Treatment 1. The fixations from this trial were then automatically coded as confused fixations.

To mitigate subjectivity, each treatment for each participant was coded by two researchers. Inter-rater agreement was satisfactory (Cohen's Kappa = 0.86). In addition to the fixation observations and the confusion coding, we asked background information from each participant, including age and gender.

4 MODEL DEVELOPMENT

4.1 Feature Selection

Due to our research purpose, we focused on fixation features. These features were retrieved from the output given by the eye tracker (see Table 1). Also, one feature indicated whether a given fixation is targeting a given AOI, and participant level features included age and gender.

As known from prior research, fixation duration can indicate confusion [10] [46]. In addition, the fixation position (x and y coordinates) is an obvious choice because *what* the user sees on the screen, in fact, causes the confusion, and the fixation position captures this association. In addition to including the x and y coordinates separately, we calculated *Distance* as the Euclidian distance of each x and y pair from the origo: $D_i = \sqrt{x_i{}^2 + y_i{}^2}$. This is done because each fixation observation is tied to x and y coordinate (i.e., they do not vary in isolation), and we wanted to make sure each position is captured correctly in the feature vector.

Figure 3: Treatment 3 with AOIs and heatmap showing the fixation density from all the participants (stronger color indicates more fixations).

Table 1: Chosen features for confusion prediction.

Feature	Definition
FIXATION FEATURES	
FixDuration*	The duration of fixation (eye position is relative stable looking at the screen) in milliseconds.
FixX*	The x-axis coordinate of where the fixation was focused on the screen.
FixY*	The y-axis coordinate of where the fixation was focused on the screen.
FixDispersion*	Fixation dispersion, computed by the eye-tracking device.
FixStart*	Fixation starting time in milliseconds, measured from the beginning of the treatment trial.
Distance	Euclidian distance of the fixation calculated from Cartesian zero point.
AOI FEATURES	
AOI-1…24	"True" or "False", depending on if the gaze is targeting the AOI.
USER FEATURES	
Age	Age of the participant.
Gender	Gender of the participant.
*Note: * indicates the values were retrieved from the eye-tracking device.*	

Finally, we include fixation start time that captures the time sequence – this variable tells when the fixation started during a given trial. The ordinal sequence of the fixations obviously matters because it informs the model of the fixation path (i.e., movement between one set of coordinates to another) as well as transition path (i.e., movement from one AOI to another).

4.2 Data Exploration and Augmentation

In total, there are 53,333 fixation observations in the dataset, 1,839 fixations on average per participant (sd = 479). According to the recommended duration of fixations [46], we remove fixation observations exceeding 600 milliseconds, as these are likely to be measurement errors. We analyze correlations between the features and find them to be low (below r = 0.2) apart from x and y coordinates that are strongly correlated (r = 0.6).

We also examine the similarity of gaze patterns between the participants by computing the Levensthein distance [47] of each participant's fixation path (defined as *eye movement from one AOI to another during the experiment*) We find a general tendency toward dissimilarity and, in fact, no two paths are exactly alike. For our data, the Levensthein distance is on average 603.5, meaning one needs to conduct 604 operations to produce two similar fixation paths.

Table 2: Confused and non-confused participants in different treatments.

	Confused	Non-confused	Total
T1	6	23	29
T2	9	20	29
T3	20	9	29
Total	35 (40%)	52 (60%)	87 (100%)

We examine the balance of the predicted variable and observe an imbalance in the share of confusion relative to non-confusion observations, so non-confusion is much more predominant in the sample (see Table 2).

The P-T level data contains more observations of non-confusion per participant and treatment. In other words, there is a class imbalance that might result in model bias. For example, if there are only 9% of confusion observations, predicting non-confusion in all cases provides 91% accuracy, even though 0% of confusion cases are correctly predicted. To correct for imbalance, we generate additional confusion observation using two approaches:

- **SMOTE:** Synthetic Minority Over Sampling Technique generates samples of the minority class by calculating the difference between the nearest neighbors and multiplying it by a random number between 0 and 1 [48].
- **ADASYN:** Adaptive Synthetic generates minority class samples by over-emphasizing samples that are more difficult to learn using a weighted distribution [49].

The purpose of the data augmentation is to "boost the signal"; because non-confused observations are so predominant, it would become harder to detect the signal from noise. Figure 4 describes the data by AOI before and after the data augmentation.

4.3 Algorithm Selection

We chose to model using two types of approach. First, we chose neural networks (NN) because they have shown high performance in a variety of tasks recently [50]. Second, we chose Random Forest (RF) because these models provide interpretability [51] that NN generally lacks. Each algorithm was deployed using the Python programming language.

We train both NN and RF with two input vectors: a) *Vector A* (incl. fixation level features from Table 1), and b) *Vector B* (fixation level features + participant level features from Table 1). The model predicts whether a given fixation observation in the dataset is labeled for confusion or not.

Regarding data augmentation, because SMOTE provided better performance gain in our testing than ADASYN, we apply it to the final models (Table 3). We also experimented with using the original, non-augmented dataset. Interestingly, the overall accuracy was higher with the non-augmented dataset, but the precision and recall were below the acceptable range (0.2–0.4). This shows the value of data augmentation; even though the overall accuracy may drop, precision and recall improve because of a more balanced training set. For prediction tasks dealing with sparsity, such as confusion detection with this dataset, data augmentation is therefore recommended.

Table 3: Predictive performance with augmented data. SMOTE provides higher performance.

	ADASYN	SMOTE
Accuracy	0.596	0.621
Precision	0.530	0.534
Recall	0.607	0.629
F1	0.587	0.603

To avoid overfitting, we trained both models with 67% of the data, retaining 33% of the data for testing the model's predictive accuracy. We also used 10-fold cross-validation to split our training data into 10 parts and then validated on each set while training on the rest.

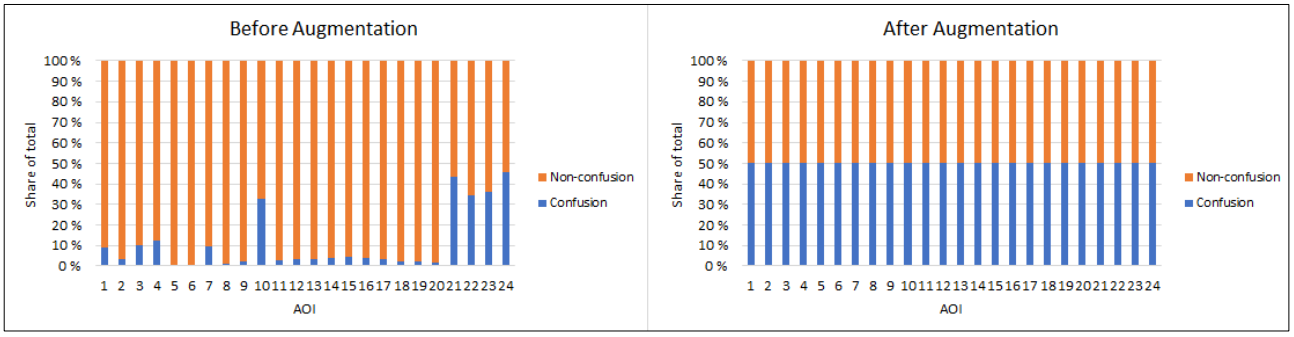

Figure 4: Share of confusion and non-confusion observations.

Regarding NN, we tested many combinations of hidden layers (4–10) and ended up with 3 layers. In between these layers, we applied MaxPool and dropout layers to avoid overfitting of training data. We did not use batch normalization as there were no vanishing gradients [52]. However, we used a validation set for hyperparameters. As a loss function, we used binary cross-entropy as the confusion would be either 1 (*true*) or 0 (*false*). As an optimizer, we used the adam, which is a method for stochastic optimization [53]. The last fully connected layer had the sigmoid activation function. Finally, we added a rectified linear unit (ReLu) activations to introduce non-linearity [50]. The architecture can be described as *NN*: 8 → 8 → 4 → 2 →1, where there are eight inputs to the first layer, outputting eight features, with the outputs halving until the final layer giving the prediction.

Second, we develop a random forest model, where we could tune hyperparameters, namely the number of trees and maximum features. The number of trees was varied from 500 to 100; seeing no significant changes, we used 300 as the model will be computationally more efficient. We set the number of maximum features, as the square root of the number of features, as customary [54].

4.4 Evaluation of Results

The results from NN and RF models are shown in Table 4. They were obtained by applying the models trained with the training set augmented with SMOTE to a non-augmented test dataset (holdout sample), in order to prevent the augmentation from artificially boosting the accuracies.

Table 4: Performance across models. PART refers to including participant information. Random forest performs better than neural network.

	NN	RF	NN+PART	RF+PART
$Accuracy_{T1}$	0.665	0.726	0.602	0.991
$Precision_{T1}$	0.394	0.391	0.789	0.973
$Recall_{T1}$	0.392	0.481	0.396	0.982
$F1_{T1}$	0.437	0.496	0.566	0.984
$Accuracy_{T2}$	0.651	0.712	0.738	0.991
$Precision_{T2}$	0.461	0.469	0.559	0.976
$Recall_{T2}$	0.468	0.554	0.601	0.994
$F1_{T2}$	0.506	0.565	0.637	0.987
$Accuracy_{T3}$	0.481	0.708	0.533	0.966
$Precision_{T3}$	0.361	0.772	0.411	0.971
$Recall_{T3}$	0.856	0.817	0.899	0.978
$F1_{T3}$	0.514	0.801	0.571	0.977
Note: All values are from predicting the test data.				

From Table 4, we can observe a few interesting findings. First, RF performs consistently better than NN. This is also visible from Figure 5 that shows the average values of the performance metrics for NN and RF. Second, models using participant level information perform clearly better than those predicted from fixation features only. If we consider 0.7 accuracy as a threshold for good performance, then all RF models are accurate and all NN models (apart from NN+PART for Treatment 2) are not.

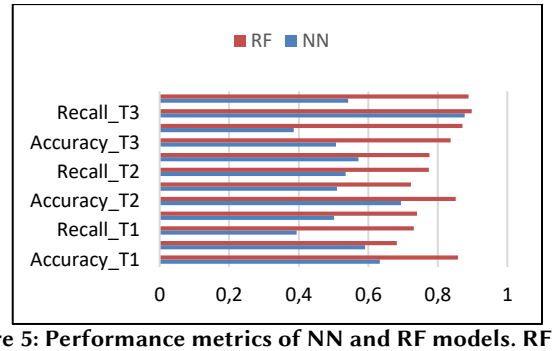

Figure 5: Performance metrics of NN and RF models. RF is performing consistently better.

Including participant information tends to increase the accuracy as well as other performance metrics for NN, the effect ranging from low to moderate improvements. However, for RF the inclusion of participant information results in a massive boost of performance, making this model clearly outperforming others. To understand why, we conducted a feature importance analysis, shown in Figure 6.

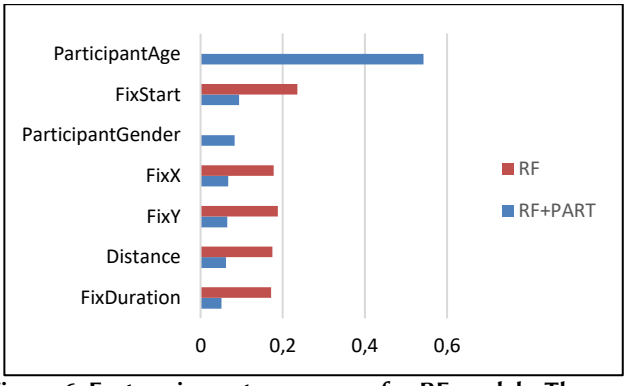

Figure 6: Feature importance scores for RF models. The range is between 0 and 1.

From Figure 6, we see that age is the most influential feature when participant information is considered. When using only fixation information, fixation start time is the most important, indicating that the model captures something from the sequence of the fixations. Fixation coordinates and duration seem to be equally important.

Third, all models tend to predict Treatment 3 better than other treatments. This effect can be explained by the nature of the experiment. Recall from Figure 3 that Treatment 3 included imagery that was not consistent with the other information in the persona profile (the images not depicting the persona but similar people). To explore the effect of treatment on confusion further, we conducted a statistical analysis using Cochran's Q test [55], which is a repeated-measures ANOVA for dichotomous variables. The result showed a significant effect between treatment and confusion (Chi-Square=30, df=2, p=3.059e-07).

We then performed the McNemar's posthoc test for each pair of treatments to isolate the effect, with results presented in Table 5. We have a significant difference of confusion between T1 (T2) and T3 (p=0.001).

Table 5: McNemar's test with continuity correction for each pair of treatments.

	T1-T2	T1-T3	T2-T3
Chi-Squared	NA	13.067	13.067
df	1	1	1
p-value	NA	0.00060	0.00060

Consequently, the higher prevalence of confusion in Treatment 3 is likely to have resulted in the observed higher performance. Overall, it seems that the lack of signal in treatments with less confusion cannot be fully compensated with data augmentation. This is an area for future research.

To explore whether creating a more complex model could increase performance, we added more layers to NN and more trees to RF; however, only achieving minimal performance gain (in the range of 1-2%). Overall, RF performs better with the eye tracking data we are dealing with. Going for more complicated models than what we experimented does not seem feasible, and we believe the signal would have been detected by our existing architectures. Thus, the conclusion is that when the prevalence of confusion is high, it can be detected easier from the fixation and participant information than when the prevalence is low.

Finally, low recall scores seem to be an issue for all other models than RF with participant information. This means that the models are under-predicting confusion, i.e., predicting non-confusion even when there is confusion. Only for Treatment 3 the recall was consistently above 0.80; however, there the low precision indicates that the model is over-predicting confusion. Thus, we can conclude that the balanced model is RF with participant level information. We use this model to visualize confusion in the form of heatmaps (Figure 7).

To address our third research question about AOI level detection of confusion, it seems from Figure 7 that the model is somewhat successful in reproducing the confusion episodes of the participants. Namely, the Treatment 1 displays noticeably lower density of confusion, whereas Treatments 2 and 3 with the additional contextual images (recall Figure 2) seem to have confusion targeting specifically those AOI regions of the screen. The AOI regions are visible in Figure 3.

The fact that confusion is indeed centered on these regions of the screen is corroborated by the qualitative think-aloud records, in which the participants repeatedly express confusion toward the additional photos. For example,

- **Participant 29 (T3):** "*not understanding why three pictures are shown.*"
- **Participant 28 (T3):** "*There are different pictures, I don't understand.*"; and
- **Participant 19 (T2):** *"[looking at] pictures of her friends… if I were to pitch based on this profile, I'm confused.*"

5 DISCUSSION

5.1 Contribution

Prediction without information on user attributes, such as gender, age, and experience, is important because often these features are not available (e.g., in mobile, VR, or webcam eye-tracking). We demonstrate the efficiency of machine learning in predicting user's cognitive state from eye-tracking data alone, obtaining a reasonable accuracy without user attributes. Including the demographic features of the users improves the accuracy furthermore, with age being the most influential feature.

Besides demonstrating the applicability of machine learning in inferring user confusion from eye-tracking data, this research addresses the call of Blascheck et al. [2] for using eye-tracking and think-aloud data to query user's mental states beyond qualitative analysis. Think-aloud data is most typically analyzed qualitatively using manual coding, whereas here we operationalize it for quantitative analysis using CDA technique [45], and use it as a feature for machine learning.

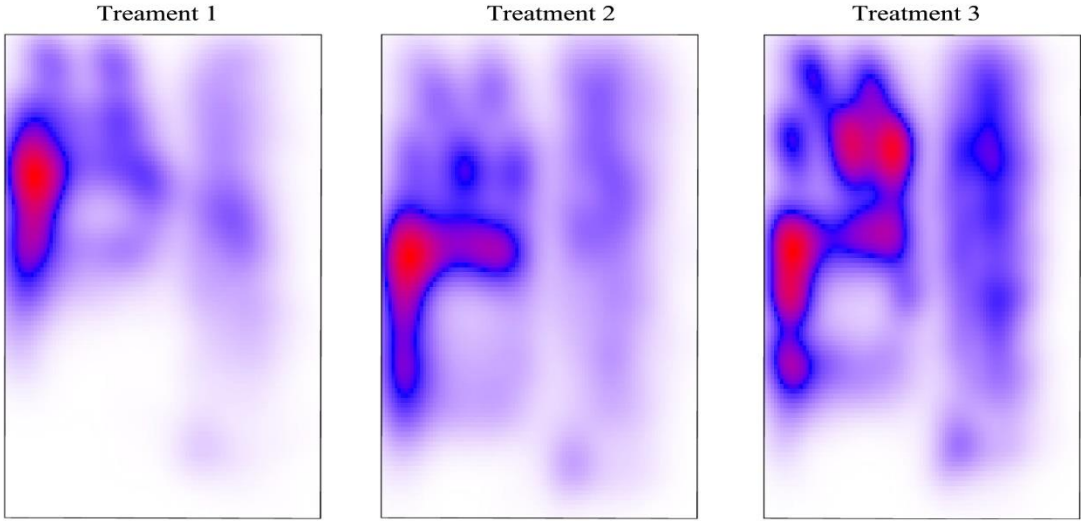

Figure 7: Heatmap of confusion predictions illustrating how the Random forest model (with participant information) "sees" confusion. Visualization is drawn using the following information from fixations the model predicted confused: Fixation x, Fixation y, Treatment. Red color indicates higher density of confusion.

5.2 Limitations

There are some limitations in our approach that should be mentioned. First, the controlled experiment may hinder the ability to find differences between the participants, because the users may feel obliged to follow a different scanpath than they would do in an authentic environment [56]. Second, users may differ by their *expressiveness*, so that some more confidently express their true cognitive states.

Therefore, it is possible that the model detects a hidden underlying pattern of confusion in the fixation data that it applies to all participants, so the participants who remained silent about their confusion are predicted wrong even though the prediction is right. Such issues could be possible when applying machine learning to user perceptions.

Third, there is limited generalizability due to the sample size. However, this issue is debatable. The whole purpose of using a machine learning is scaling with the data, and so our approach can be deployed with other eye-tracking data, especially since we provide the source code for replication. Moreover, the features used in our model are available from the output of different eye-tracking hardware: as far we know, all commercial trackers provide fixation duration, accuracy, and position. Perceived confusion could be labeled. By asking directly, coding from think-aloud records (e.g., using CDA), or by inference (e.g., the pitch of voice during confusion episodes).

5.3 Future Research Avenues

More work with richer features and larger datasets is needed to learn the relationship between fixation patterns and confusion, as there are features that we did not have access to, that could be impactful. These include, e.g., the modality of a person's voice when he or she expresses confusion or the time-synchronization between confusion utterance and eye-tracking observations.

Therefore, more eye-tracking analyses with machine learning in the user study context are needed. We encourage other researchers to validate and further develop our architecture and to improve techniques for prediction of user confusion from eye-tracking data.

REFERENCES

[1] T. Jiang, Q. Guo, Y. Xu, and S. Fu, "A diary study of information encountering triggered by visual stimuli on micro-blogging services," *Information Processing & Management*, vol. 56, no. 1, pp. 29–42, 2019.

[2] T. Blascheck, M. John, S. Koch, L. Bruder, and T. Ertl, "Triangulating User Behavior Using Eye Movement, Interaction, and Think Aloud Data," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, New York, NY, USA, 2016, pp. 175–182.

[3] M. A. Just and P. A. Carpenter, "A theory of reading: From eye fixations to comprehension," *Psychological Review*, vol. 87, no. 4, pp. 329–354, 1980.

[4] J. H. Goldberg and X. P. Kotval, "Computer interface evaluation using eye movements: methods and constructs," *International Journal of Industrial Ergonomics*, vol. 24, no. 6, pp. 631–645, Oct. 1999.

[5] Z. Bylinskii, M. A. Borkin, N. W. Kim, H. Pfister, and A. Oliva, "Eye Fixation Metrics for Large Scale Evaluation and Comparison of Information Visualizations," in *Eye Tracking and Visualization*, 2015, pp. 235–255.

[6] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognitive Psychology*, vol. 8, no. 4, pp. 441–480, Oct. 1976.

[7] S. Eraslan, Y. Yesilada, and S. Harper, "Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison," *Journal of Eye Movement Research*, vol. 9, no. 1, Dec. 2015.

[8] S. Eraslan, Y. Yesilada, and S. Harper, "Scanpath Trend Analysis on Web Pages: Clustering Eye Tracking Scanpaths," *ACM Transactions on the Web*, vol. 10, no. 4, pp. 1–35, Nov. 2016.

[9] L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. Granka, and G. Gay, "The influence of task and gender on search and evaluation behavior using Google," *Information Processing & Management*, vol. 42, no. 4, pp. 1123–1131, Jul. 2006.

[10] J. H. Goldberg and J. I. Helfman, "Identifying Aggregate Scanning Strategies to Improve Usability Evaluations," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 54, no. 6, pp. 590–594, Sep. 2010.

[11] R. J. K. Jacob and K. S. Karn, "Commentary on Section 4 - Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises," in *The Mind's Eye*, J. Hyönä, R. Radach, and H. Deubel, Eds. Amsterdam: North-Holland, 2003, pp. 573–605.

[12] J. H. Goldberg and J. I. Helfman, "Scanpath Clustering and Aggregation," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, New York, NY, USA, 2010, pp. 227–234.

[13] J. Salminen, L. Nielsen, S.-G. Jung, J. An, H. Kwak, and B. J. Jansen, "'Is More Better?': Impact of Multiple Photos on Perception of Persona Profiles," in *Proceedings of ACM CHI Conference on Human Factors in Computing Systems (CHI2018)*, Montréal, Canada, 2018.

[14] B. Follet, O. Le Meur, and T. Baccino, "New Insights into Ambient and Focal Visual Fixations using an Automatic Classification Algorithm," *i-Perception*, vol. 2, no. 6, pp. 592–610, Aug. 2011.

[15] O. Alhadreti and P. Mayhew, "To Intervene or Not to Intervene: An Investigation of Three Think-aloud Protocols in Usability Testing," *J. Usability Studies*, vol. 12, no. 3, pp. 111–132, May 2017.

[16] K.-H. Tan, D. J. Kriegman, and N. Ahuja, "Appearance-based eye gaze estimation," in *Sixth IEEE Workshop on Applications of Computer Vision, 2002. (WACV 2002). Proceedings.*, 2002, pp. 191–195.

[17] E. Y. Kim and S. K. Kang, "Eye Tracking Using Neural Network and Mean-Shift," in *Computational Science and Its Applications - ICCSA 2006*, 2006, pp. 1200–1209.

[18] M. J. Coughlin, T. R. H. Cutmore, and T. J. Hine, "Automated eye tracking system calibration using artificial neural networks," *Computer Methods and Programs in Biomedicine*, vol. 76, no. 3, pp. 207–220, Dec. 2004.

[19] B. Wolfe and D. Eichmann, "A Neural Network Approach to Tracking Eye Position," *International Journal of Human–Computer Interaction*, vol. 9, no. 1, pp. 59–79, Mar. 1997.

[20] J. Zhu and J. Yang, "Subpixel eye gaze tracking," in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, 2002, pp. 124–129.

[21] L. Behera, I. Kar, and A. C. Elitzur, "A Recurrent Quantum Neural Network Model to Describe Eye Tracking of Moving Targets," *Found Phys Lett*, vol. 18, no. 4, pp. 357–370, Aug. 2005.

[22] M. Pomplun, B. Velichkovsky, and H. Ritter, "An artificial neural network for high precision eye movement tracking," in *KI-94: Advances in Artificial Intelligence*, 1994, pp. 63–69.

[23] K. Essig, M. Pomplun, and H. Ritter, "A neural network for 3D gaze recording with binocular eye trackers," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 21, no. 2, pp. 79–95, Apr. 2006.

[24] R. Zemblys, D. C. Niehorster, O. Komogortsev, and K. Holmqvist, "Using machine learning to detect events in eye-tracking data," *Behav Res*, pp. 1–22, Feb. 2017.

[25] T. Harada, H. Iwasaki, K. Mori, A. Yoshizawa, and F. Mizoguchi, "Evaluation model of cognitive distraction state based on eye-tracking data using neural networks," in *2013 IEEE 12th International Conference on Cognitive Informatics and Cognitive Computing*, 2013, pp. 428–434.

[26] G. Kuperberg and S. Heckers, "Schizophrenia and cognitive function," *Current Opinion in Neurobiology*, vol. 10, no. 2, pp. 205–210, Apr. 2000.

[27] A. Campana, R. Dud, O. Qambini, and S. Scarone, *An Artificial Neural Network That Uses Eye-Tracking Performance to Identify Patients With Schizophrenia Abstract*.

[28] F.-Y. Hsu, H.-M. Lee, T.-H. Chang, and Y.-T. Sung, "Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques," *Information Processing & Management*, vol. 54, no. 6, pp. 969–984, 2018.

[29] C. Ehmke and S. Wilson, "Identifying Web Usability Problems from Eye-tracking Data," in *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not As We Know It - Volume 1*, Swinton, UK, UK, 2007, pp. 119–128.

[30] O. Ferhat and F. Vilariño, "Low Cost Eye Tracking," *Intell. Neuroscience*, vol. 2016, pp. 17–, Mar. 2016.

[31] J. Salminen *et al.*, "Generating Cultural Personas from Social Data: A Perspective of Middle Eastern Users," in *Proceedings of The Fourth International Symposium on Social Networks Analysis, Management and Security (SNAMS-2017)*, Prague, Czech Republic, 2017.

[32] J. Salminen, S.-G. Jung, J. An, H. Kwak, and B. J. Jansen, "Findings of a User Study of Automatically Generated Personas," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2018, p. LBW097:1–LBW097:6.

[33] J. Salminen *et al.*, "From 2,772 segments to five personas: Summarizing a diverse online audience by generating culturally adapted personas," *First Monday*, vol. 23, no. 6, Jun. 2018.

[34] J. Salminen, B. J. Jansen, J. An, H. Kwak, and S.-G. Jung, "Automatic Persona Generation for Online Content Creators: Conceptual Rationale and a Research Agenda," in *Personas - User Focused Design*, L. Nielsen, Ed. London: Springer London, 2019, pp. 135–160.

[35] J. Salminen, S. Sengun, S.-G. Jung, and B. J. Jansen, "Design Issues in Automatically Generated Persona Profiles: A Qualitative Analysis from 38 Think-Aloud Transcripts," in *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR)*, Glasgow, UK, 2019.

[36] J. An, H. Kwak, J. Salminen, S. Jung, and B. J. Jansen, "Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data," *ACM Transactions on the Web (TWEB)*, vol. 12, no. 3, 2018.

[37] J. An, H. Kwak, S. Jung, J. Salminen, and B. J. Jansen, "Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data," *Social Network Analysis and Mining*, vol. 8, no. 1, 2018.

[38] J. An, H. Kwak, and B. J. Jansen, "Validating Social Media Data for Automatic Persona Generation," in *Proceedings of Second International Workshop on Online Social Networks Technologies (OSNT-2016), 13th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, Agadir, Morocco, 2016.

[39] S.-G. Jung, J. An, H. Kwak, M. Ahmad, L. Nielsen, and B. J. Jansen, "Persona Generation from Aggregated Social Media Data," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2017, pp. 1748–1755.

[40] L. Nielsen, K. S. Nielsen, J. Stage, and J. Billestrup, "Going Global with Personas," in *Human-Computer Interaction – INTERACT 2013*, 2013, pp. 350–357.

[41] J. Salminen, B. J. Jansen, J. An, H. Kwak, and S. Jung, "Are personas done? Evaluating their usefulness in the age of digital analytics," *Persona Studies*, vol. 4, no. 2, pp. 47–65, Nov. 2018.

[42] C. N. Chapman and R. P. Milham, "The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, no. 5, pp. 634–636, Oct. 2006.

[43] J. Salminen, S. Jung, J. An, H. Kwak, L. Nielsen, and B. J. Jansen, "Confusion and information triggered by photos in persona profiles," *International Journal of Human-Computer Studies*, vol. 129, pp. 1–14, Sep. 2019.

[44] J. Salminen, B. J. Jansen, J. An, S. Jung, L. Nielsen, and H. Kwak, "Fixation and Confusion – Investigating Eye-tracking Participants' Exposure to Information in Personas," in *Proceedings of The ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2018)*, New Jersey, USA, 2018.

[45] T. Tenbrink, "Cognitive Discourse Analysis: accessing cognitive representations and processes through language data," *Language and Cognition*, vol. 7, no. 1, pp. 98–137, Jul. 2014.

[46] L. Granka, M. Feusner, and L. Lorigo, "Eye Monitoring in Online Search," in *Passive Eye Monitoring*, R. I. Hammoud, Ed. Springer Berlin Heidelberg, 2008, pp. 347–372.

[47] A. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Phys. Dokl.*, vol. 10, pp. 707–710, 1966.

[48] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[49] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008, pp. 1322–1328.
[50] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.
[51] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.
[52] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, 2013, pp. 1310–1318.
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[54] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.
[55] M. W. Tate and S. M. Brown, "Note on the Cochran Q test," *Journal of the American Statistical Association*, vol. 65, no. 329, pp. 155–160, 1970.
[56] R. J. Fisher, "Social Desirability Bias and the Validity of Indirect Questioning," *Journal of Consumer Research*, vol. 20, no. 2, pp. 303–315, 1993.