Predicting Visual Search Task Success from Eye Gaze Data as a Basis for User-Adaptive Information Visualization Systems

MORITZ SPILLER[*], INKA - Innovation Laboratory for Image Guided Therapy, Health Campus Immunology Infectiology and Inflammation ($GC-I^3$), Otto-von-Guericke-University, Germany
YING-HSANG LIU[†], University of Southern Denmark, Denmark
MD ZAKIR HOSSAIN, The Australian National University, Australia
TOM GEDEON, The Australian National University, Australia
JULIA GEISSLER, Otto von Guericke University, Germany
ANDREAS NÜRNBERGER, Otto von Guericke University, Germany

Information visualizations are an efficient means to support the users in understanding large amounts of complex, interconnected data; user comprehension, however, depends on individual factors such as their cognitive abilities. The research literature provides evidence that user-adaptive information visualizations positively impact the users' performance in visualization tasks. This study attempts to contribute towards the development of a computational model to predict the users' success in visual search tasks from eye gaze data and thereby drive such user-adaptive systems. State-of-the-art deep learning models for time series classification have been trained on sequential eye gaze data obtained from 40 study participants' interaction with a circular and an organizational graph. The results suggest that such models yield higher accuracy than a baseline classifier and previously used models for this purpose. In particular, a Multivariate Long Short Term Memory Fully Convolutional Network (MLSTM-FCN) shows encouraging performance for its use in on-line user-adaptive systems. Given this finding, such a computational model can infer the users' need for support during interaction with a graph and trigger appropriate interventions in user-adaptive information visualization systems. This facilitates the design of such systems since further interaction data like mouse clicks is not required.

CCS Concepts: • **Human-centered computing** → **User studies**; *Information visualization*; • **Computing methodologies** → *Machine learning approaches*.

Additional Key Words and Phrases: Eye tracking, User-adaptation, Time series classification, Individual differences

[*]Also with The Australian National University.
[†]Also with The Australian National University.

Authors' addresses: Moritz Spiller, INKA - Innovation Laboratory for Image Guided Therapy, Health Campus Immunology Infectiology and Inflammation ($GC-I^3$), Otto-von-Guericke-University, Leipziger Straße 44, Magdeburg, Germany, moritz.spiller@med.ovgu.de; Ying-Hsang Liu, University of Southern Denmark, Kolding, Denmark, yingliu@sdu.dk; Md Zakir Hossain, The Australian National University, Canberra, Australia, zakir.hossain@anu.edu.au; Tom Gedeon, The Australian National University, Canberra, Australia, tom.gedeon@anu.edu.au; Julia Geissler, Otto von Guericke University, Magdeburg, Germany, julia.geissler@ovgu.de; Andreas Nürnberger, Otto von Guericke University, Magdeburg, Germany, andreas.nuernberger@ovgu.de.

Manuscript submitted to ACM

ACM Reference Format:
Moritz Spiller, Ying-Hsang Liu, Md Zakir Hossain, Tom Gedeon, Julia Geissler, and Andreas Nürnberger. 2021. Predicting Visual Search Task Success from Eye Gaze Data as a Basis for User-Adaptive Information Visualization Systems. *ACM Trans. Interact. Intell. Syst.* 1, 1, Article 1 (January 2021), 25 pages. https://doi.org/10.1145/3446638

1 INTRODUCTION

Information visualizations are an efficient means to support the users in understanding large amounts of complex, interconnected data and are commonly used in visual analytics [38], e-learning [34], and visual information management platforms. However, the visualization of complex structures and cross-linked data, like affiliations between particular entities or networks of academic citations, on a computer's 2D-screen is a challenging task. Optimally presenting such data is crucial to provide an intuitive way to interact with them effectively and efficiently. However, which kind of visualization the users perceive as optimal depends on individual factors. Consequently, each user will perceive another visualization as optimal to impart a specific type of information, motivating research on visualization interfaces that can adapt to any individual user.

Research on individual differences for user-adaptive interface design reveals that the user characteristics affect visual search behavior [62, 65, 74, 79]. The users' cognitive abilities, such as perceptual speed and verbal working memory, are correlated with eye tracking measures [67, 75]. Users with lower levels of cognitive abilities have been found to perform worse in terms of completion time and accuracy while interacting with information visualizations [12, 73, 77]. The findings suggest that user characteristics, such as cognitive abilities, prior domain knowledge, and experience as well as user preferences need to be considered in the design of optimal visualization interfaces and highlight the urgent need for user-adaptive information visualization systems, which can infer the usefulness of visualization from interaction data and intervene accordingly. For example, recent research has attempted to infer the users' individual differences such as cognitive abilities or confusion from their eye movements to recognize the need for intervention in user-adaptive information visualization systems [45, 67, 69, 70].

Yet, there are relatively few user-adaptive interface studies that focus on the actual user comprehension of information when interacting with visualization interfaces in the research literature. This study will focus on the comprehension of circular and organizational graph types. These graph types have not been addressed in previous studies and differ strongly from generic information visualizations like bar and radar charts. In consequence, research results on the latter do not apply to circular and organizational graph types. The bar and radar charts used in previous studies illustrate the differences and relations between certain quantities, while the graphs used in this study contain textual information and highlight relations between the depicted entities. Since bar and radar charts are commonly known visualizations, the participants already possess an existing problem-solving strategy. These generic information visualizations may not be suitable for evaluating actual comprehension of complex analytic tasks partly because domain knowledge influences graph comprehension [41].

Comprehension of these graphs and visualizations is concerned with the users' interpretation of quantitative information, which depends on the users' cognitive abilities, such as attention span and short-term visual working memory [59]. A model of knowledge-based graph comprehension by Freedman and Shah [22] assumes that prior knowledge and skills interact with a graph's display characteristics in the course of graph comprehension. Graph comprehension processes emerge from integrated, sequential sub-processes like encoding graphical descriptions, information search, and reasoning [9, 43].

However, inferring the users' graph comprehension based on automatically acquired interaction data such as eye gaze seems a complex task, and very little research has been done in this area. On the other hand, a visual search task's success ultimately depends, at least to some extent, on the degree to which the user understands the data representation and the interpretation of complex information it contains. This relation motivates this study's approach to use the users' predicted success on a visual search task as an indicator for graph comprehension.

Existing research on user-adaptive visualization interfaces has mostly used static machine learning classifiers, which yielded average performance. Motivated by the success of deep learning in other research areas [37, 78], this study will make use of state-of-the-art deep learning time series classification algorithms for modeling visual search behavior and predict the success of a visual search task.

Previously, state-of-the-art performance in TSC (time series classification) was reached by HIVE-COTE, an ensemble of 37 classifiers that is computationally expensive to train and tune [2]. Motivated by the success of deep learning models in, e.g., computer vision and natural language processing, such models have also been adapted to TSC [36, 37, 78]. Research has shown that deep learning models like ResNet (Residual Network) reached state-of-the-art performance while being less computationally expensive than HIVE-COTE [19]. Since the data recorded by the eye tracker is, in fact, time series data and might include valuable information on the time scale [18], applying TSC models to eye tracking data can be expected to yield sufficient performance to drive on-line user-adaptive information visualization systems.

Previous computational models predicting the users' graph comprehension have yielded average performance. This study aims to exceed these results by providing models that perform good enough to drive on-line user-adaptive visualization systems. Considering the findings of the influence of time constraint on user's reading speed and search behavior [17, 46] and the average performance of simple machine learning models [e.g. 52, 53, 70], we imposed the time constraint on experimental conditions by experiment design.

This research aims to contribute to the objective of user-adaptive information visualization systems. In that process, the main objective is to explore if a computational model can infer the users' need for support based on a given task's predicted success. To achieve that, eye tracking data obtained while the user interacts with two graph types are analyzed to provide a basis for adaptation of the visualization.

To explore the feasibility of real-time prediction while the users interact with a visualization, the classifiers are trained on sequences of different lengths. This approach will help analyze a computational model's ability to be integrated into an on-line user-adaptive information visualization system.

Specifically, the following two research questions are addressed:

(1) To what extent can a computational model predict the users' visual search task success to detect their need for support when interacting with an information visualization from sequential eye gaze data?
(2) Is it feasible to infer the users' task success within the first ten seconds of interaction?

In summary, this study's key findings are:

(1) MLSTM-FCN (Multivariate Long Short Term Memory Fully Convolutional Network) is the best performing classifier for the used time series eye gaze data, based on the F1-scores. ResNet (Residual Network) and FCN (Fully Convolutional Network) do not perform significantly better than the baseline classifier, LR (Logistic Regression). Task success can be predicted with an accuracy above 0.8.
(2) Although MLSTM-FCN outperforms all other classifiers, its inference time increases linearly to more than one second on a ten-second sequence. ResNet and FCN should still be considered when near real-time inference is required.

This study contributes towards the goal of user-adaptive information visualizations by evaluating the applicability of a computational model for inferring if an individual user correctly acquires information imparted by a graph from eye gaze data. This study will focus on the comprehension of circular and organizational graph types. Furthermore, the feasibility of predictions within the first ten seconds of interaction is explored. Additionally, the performance of three deep learning classifiers on eye gaze data is assessed.

In this article's remainder, the related work is discussed (Section 2), and the used research methods (Section 3) are introduced. After presenting the results (Section 4) and discussing them (Section 5), conclusions are drawn, and future work is stated (Section 6).

2 RELATED WORK

Inferring the usefulness of a visualization for a particular user is an essential feature for user-adaptive interfaces. The users' visual search task success is a major indicator of the usefulness of visualizations. In the following, related work in graph comprehension and its link to eye tracking is discussed. Subsequently, research on user-adaptive interfaces is reviewed, and the gaps addressed by this study are highlighted.

2.1 Graph Comprehension

Many individual factors influence graph comprehension. These properties include perceptual and cognitive abilities [1, 59], prior domain knowledge, and experience in interaction with graph formats [22, 41]. These abilities and skills come into play over the two fundamental phases of graph comprehension: visual search and reasoning. The individual differences regarding perceptual and cognitive abilities, domain knowledge, and experience using graphs highlight that user-adaptive information visualization systems can facilitate graph comprehension in a wide variety of applications.

Early research of graph comprehension by Pinker [59] points out the relation between graph comprehension theory and perceptual and cognition theory. According to Pinker's model, graph comprehension is dependent on the users' attention, capacity in short-term visual working memory, and encoding faculties and emerges from four sub-processes, where the user assembles a conceptual message in order to solve the task at hand. Before message assembly, the user searches the graph schema for the required information. Based on Pinker's model Freedman and Shah [22] propose a model of knowledge-based graph comprehension. Their model assumes that prior knowledge and skills interact with a graph's display characteristics.

Studies suggest that graph comprehension emerges from integrated, sequential sub-processes like encoding graphical descriptions and retrieve information from them [9, 43]. Specifically, Körner et al. [43] proposed that the search and the reasoning processes in hierarchical graph comprehension are typically conducted in a sequence. Körner's findings reveal that the reasoning process has to be finished before a sequence is used to infer graph comprehension. Comprehension is then accomplished by extracting information from these relations [43]. Carpenter and Shah [9] found that even for simple tasks, graph comprehension is a cognitively demanding process. They further state that "graph comprehension might be more accurate and more complete if the graph's format were changed" [9, p. 75], highlighting the need for user-adaptive information visualization systems, which can infer the users' need for a different perspective on the data, i.e., another visualization.

The studies introduced above consistently found that graph comprehension depends on various individual factors such as perceptual and cognitive abilities, domain knowledge, and experience. It was also found that adapting the visualization to the user may improve graph comprehension. This study addresses the need for user-adaptive information visualization systems.

Furthermore, previous research repeatedly states that visual search is an important phase in graph comprehension, motivating the use of eye gaze data to infer a user's graph comprehension. The link between eye movements and search is described in the following.

2.2 Graph Comprehension and Eye Tracking

Some empirical studies described in the previous section 2.1 have used the eye tracking technique to gather data during various user experiments since eye tracking is a valuable metric to infer the users' intrinsic cognitive processes. The graphs used in this study include textual information organized as nodes to highlight their relations. Since comprehension of any graph relies on visual information search and individual abilities, these factors are important to consider. As demonstrated in the following, eye tracking can provide valuable insight into both intrinsic processes, motivating the use of eye tracking in this study.

2.2.1 Eye Gaze Metrics in Visual Information Search. Eye gaze data has been used in a variety of applications in regards to information search. Eye tracking data provides insights into the users' search activity, the relevance of search results, and the complexity of the search array (the area where the graph's elements are located) [63]. These insights also provide valuable information about the success of the visual search phase in graphs.

In information retrieval research, eye gaze data has been used to infer the users' search activity [66] and to discriminate the relevance of search results by analyzing the users' pupillary responses [56]. Research shows that 500 to 4000 ms (milliseconds) after a stimulus, differences in the users' pupil size concerning the relevance of a search result can be detected [56]. The eye gaze reveals the users' visual search strategy [23], and there are correlations between eye gaze patterns and user satisfaction [80]. Eye gaze data can also predict the users' intentions and goals during manipulation tasks in graphical user interfaces [13].

Further studies suggest that the complexity of the search array [63], as well as the difficulty of the search task [63, 79], influence the patterns of eye movements. Additionally, fixations and saccades provide insights into how much information is being processed during the search phase; however, how much information is acquired during a fixation depends on the particular search task [63].

2.2.2 Eye Gaze Metrics to Infer the Users' Individual Abilities. The users' individual cognitive abilities and prior knowledge influence their graph comprehension, task performance and preferred visualization type.

Research on intelligent user interfaces shows that users' perceptual speed and verbal working memory influence their eye gaze patterns in general and concerning the task difficulty and visualization type [74]. Cognitive abilities impact the users' task performance and satisfaction regarding a particular visualization type [58, 73]. The research literature also provides evidence that cognitive abilities like perceptual speed, visual and verbal working memory can be inferred from eye gaze data [67, 70]. Eye movement patterns are used to infer the users' domain knowledge during search [10], which also has an impact on graph comprehension (See Section 2.1).

Graph comprehension requires individual abilities and skills that are also relevant to text comprehension [22] and eye movements are a reliable indicator of the text's difficulty; the more difficult a text is, the less likely it is that the text is understood [64].

2.3 User-adaptive Information Visualization Systems

User-adaptive systems have been evaluated and applied in personalized search [68] or e-learning [34] to support the individual user during the related tasks or applications. In contrast, user-adaptive information visualization systems have only recently received increased research interest [e.g. 13, 44, 58, 76].

Early approaches to user-adaptive information visualization systems required active user interaction, experts' intervention, and have limited generalizability. These limitations make these systems impractical for real-time user-adaptive systems and, in particular, information visualizations. However, the related studies [25, 26] provide evidence that user-adaptive systems have a positive impact on the users' performance. For instance, the system proposed by Grawemeyer [26] can infer the users' visualization expertise and preferences during previous tasks by monitoring visualization selection. The findings suggest that user-adaptive visualizations improve their performance in terms of accuracy and task completion time.

Another approach by Gotz and Wen [25] makes use of interaction data in the form of mouse clicks to monitor the users' behavior in real-time. The findings indicate that interaction data can be used to recognize inefficient usage patterns and detect the need to intervene by recommending an alternative information visualization, which is a similar approach to the system proposed in Grawemeyer [26]. However, the usage patterns labeled as suboptimal and the alternative visualizations have been worked out by experts before the actual experiment, and the system requires interaction data in the form of mouse clicks or entered text. Since not all visualization interfaces are designed to provide the possibility for non-visual interactions, such approaches' generalizability is limited.

Another thread of research uses eye gaze data to make real-time predictions for driving adaptive systems [70]. The study demonstrates that the users' eye gaze pattern is an informative feature to infer task type and characteristics, user performance, and cognitive abilities. The results of the classification experiments, using LR (Logistic Regression), an SVM (Support Vector Machine), Decision Tree, and an ANN, show that the classification performance of all used classifiers has been significantly higher than the performance of a baseline classifier, which always selects the most likely class. LR has consistently been the best performing classifier, corroborating the findings of previous studies [7, 35] that LR performs well with eye gaze data.

Their study focuses on inferring the users' cognitive abilities and developing a user-adaptive visualization system based on the results. In contrast, this study aims to infer the need for intervention by the system based on the predicted outcome, i.e., if the user will correctly acquire the information imparted by the graph. Furthermore, the classification experiments using simple machine learning algorithms yield results between 55 % and 60 %, which might not be sufficient for a real-time system.

In summary, existing research has demonstrated the feasibility of driving user-adaptive information visualization systems based on eye gaze data. A more direct approach based on the user comprehension is a step forward to building a computational model to drive such a system. Sophisticated machine learning classifiers for time series classification can be expected to yield higher performance models, capable of driving on-line user-adaptive information visualization systems. Importantly, for some classification tasks, the users' eye gaze pattern contains the most relevant information at the beginning of the interaction [70].

Conclusively, this study addresses the gaps in the research literature, exemplified by user comprehension of circular and organizational graphs:

(1) In contrast to previous studies that focused on common visualization types, this research explores graph comprehension of complex, interconnected circular and organizational graphs using eye gaze data. These graph

types contain textual information organized as nodes to display their relations, which has not been specifically considered in the literature.

(2) In previous studies, the users have performed the experiments without restrictive time constraints. This study imposes a time constraint on the user to simulate conditions as in an realistic scenario like, e.g., e-learning.

(3) This study aims to develop a computational model that can infer if the user has actually correctly acquired information from their eye movements during interaction with the graph. Previous studies have developed computational models based on the users' cognitive abilities like perceptual speed and visual working memory from eye gaze data, but it is unclear how the user successfully acquires information.

(4) In this study, state-of-the-art deep learning models for time series classification are used to infer user comprehension from eye gaze data. Existing research utilized statistical features extracted from segmented eye gaze data to train basic machine learning classifiers. This approach yielded average performance, which might not be sufficient to guide visualization adaptation. State-of-the-art time series classification algorithms [33, 37] can be expected to surpass these results and reach the levels of accuracy required to drive real-time user-adaptive visualization systems.

3 METHODS

3.1 User Experiment

3.1.1 Experimental Setup. To gather eye gaze data for this study, a user experiment was conducted as part of a larger industry-linked research project. The experiment was designed to collect the participants' eye gaze data during interaction with two different graph types. The circular graph, depicted in Figure 1, highlights the complex structures of associations between the different entities, which can be centered into inner and outer rings. The organizational graph in Figure 2 ensures the links among the entities and is useful to determine major interests on a single entity. These two graph types are chosen since they are frequently used visualizations in industry partner's software for displaying different types of information, and this client was interested in comparing their effect on the interacting users. Guchev et al. [28] considered node-link-group and node-link diagrams, using a circular layout, to study the design of data visualizations displaying the relationships of the data amongst each other. The authors state that the readability of graphs for interactive visual exploration requires further research, providing additional motivation for this study. The experiment design enhances the study's ecological validity by choosing two graph types used in industrial applications and considering the relevance of experimental tasks [39].

After a brief training introducing the study, the participants were asked six similar questions for both graphs to find people's names, addresses, and connections between people. Specifically, the questions were: 1) Where does [*a person*] live? 2) Who lives at [*a particular address*]? 3) Who lives at [*a particular address*]? 4) How many directors does [*a particular company*] have? 5) Which state do most of the directors of [*a particular company*] live in and 6) What connects [*one company*] to [*another company*]?. The questions were presented on the top left corner of the screen with an empty input box to its right.

Since this study is within-subject design, the questions' presentation order was randomized by a Latin Square design [42]. Different datasets for the two types of graphs were used to minimize the possible learning effect of presenting the circular graph first in conditions arrangement. The graph was visible while answering the questions.

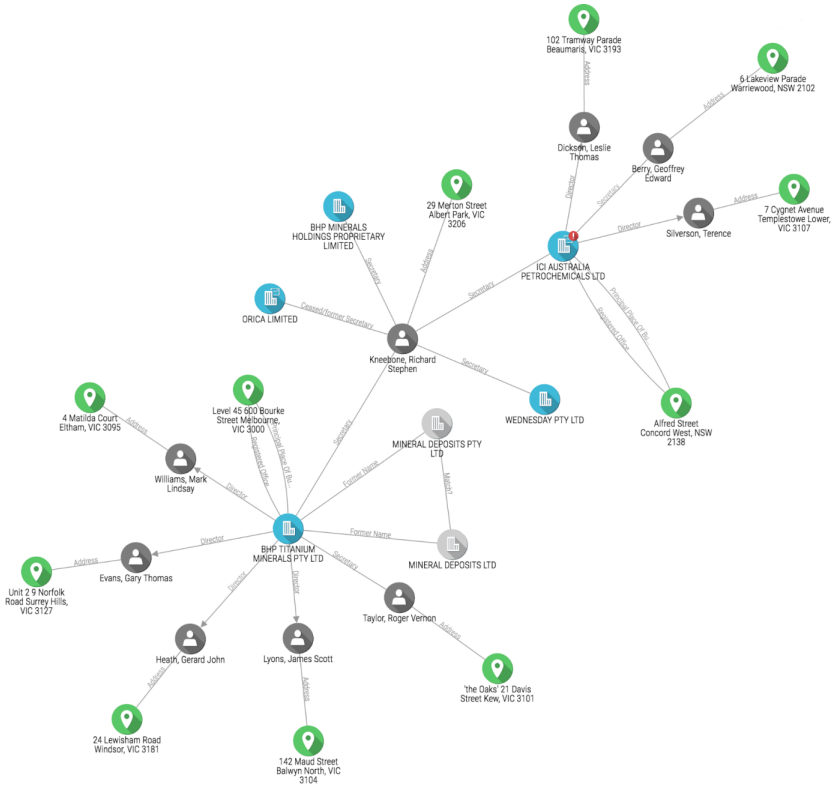

Fig. 1. The circular graph used in the user experiment highlights the complex structures of associations between the different entities.

The static system design allows gaining more control over the possible confounding variables that might have been introduced if interactive visualization systems were used in this study [cf. 6, 79]. The participants were able to scroll up and down the screen. A time limit of 45 seconds had been imposed for each session. When the participants finished the task earlier, the system provided the option of proceeding to the next question. When the time limit was reached, the system displayed the next question. Each participant finished a total of twelve search sessions (6×2, questions $\times$ graph types), with no break between sessions. Font size, lighting conditions, colors, and symbols were kept constant throughout the whole experiment. The university ethics committee approved the experiment protocol.

3.1.2 Participants. After excluding faulty recordings, the data from 40 participants (23 female, 17 male) with an average age of 22.15 ± 5.3 years remained in the dataset. Faulty recordings were identified by manually checking the acquired data for long sequences of zeros or outliers. All the subjects participated in the study without remuneration. The participants had normal or corrected to normal vision and signed consent forms before their participation. Basic demographic data such as age, gender, or mother tongue have been gathered. The participants indicated minor familiarity with circular or organizational graph types like they are used in this study. Further responses regarding familiarity with information visualization systems and other user perception data were not collected since this study focused on visual search behavior when interacting with the chosen two graph types.

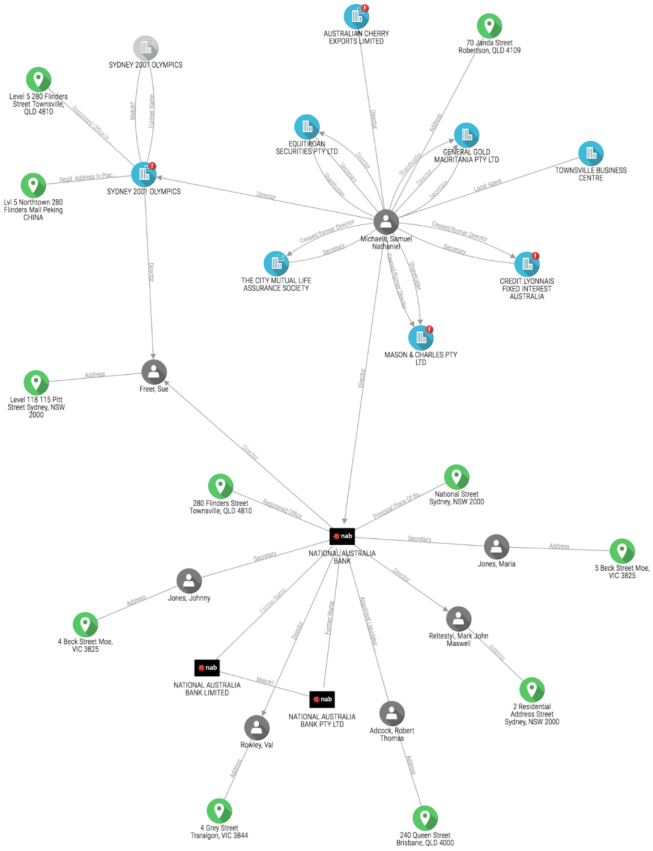

Fig. 2. The organizational graph ensures the links among the entities and is useful to determine major interests on a single entity.

3.1.3 Apparatus. The experiment was conducted on a 15.6" Dell laptop with a resolution of 1366 × 768 pixels. The laptop's display's maximum peak illuminance is 570 nits at a contrast ratio (peak/min) of 540 nits. Those illuminance parameters have been reported by the laptop's OS (Operating System). The luminance in the room was kept constant for all participants.

During the experiment, the users' eye movements were recorded with the EyeTribe eye tracker[1] at a sampling rate of 60 Hz. The participants were able to interact with the graph via a mouse and a keyboard. The experiment was started after calibrating the eye tracker. The calibration process was reiterated until the system reported optimal calibration accuracy. Participants were requested to minimize their upper body movements during the experiment to reduce undesired artifacts in the signals.

3.1.4 User Performance & Dataset. During the user experiment, the eye movements of 40 participants have been tracked while they answered six questions for both the circular and the organizational graph. In summary, 240 questions have been asked to the participants per graph type.

[1]https://theeyetribe.com/

On the circular graph, the participants answered 187 out of 240 questions correctly, while on the organizational graph, the participants answered 186 out of 240 questions correctly. Therefore, it can be assumed that the participants perceived the two graph types similarly difficult.

On average, a participant answered 77.71 ± 19.67 % of the questions correctly. Seven of them answered all questions correctly. Since the questions asked to the participants appear rather simple, it might be surprising that 20 % have been answered incorrectly. It was found that participants' took ~ 25 seconds to answer, if their answer was correct, while they took ~ 35 seconds if they answered incorrectly. This indicates that the time constraint that was, in contrast to other studies, imposed on the users influenced the share of correctly answered questions.

Consequently, both datasets are unbalanced for the classification of a correctly or incorrectly answered question, where ~ 78% of the data samples belong to the positive and ~ 22% of the samples belong to the negative class.

3.2 Data Analysis

3.2.1 Pre-processing. Eye tracking data is typically messy [50]. To facilitate optimal and interpretable results, the data are pre-processed to remove missing values, noise, and unrelated features and correct individual differences in pupil size.

The Steffen interpolation is applied to remove missing values, which frequently occur due to recording artifacts from the eye tracker or eye blinks. In contrast to the commonly used cubic spline interpolation [27, 49, 55], the Steffen interpolation does not produce any local extrema and is, therefore, more suitable for interpolating eye gaze data [30].

Subsequently, a periodic 10-point Hann moving window average is applied to filter out the noise and unrelated features from the eye tracking data [31, 32].

Subtractive baseline correction is applied to correct individual differences in pupil size [50]. To keep the experimental conditions as realistic as possible, no separate screen to measure the pupillary baseline was incorporated. Instead, the average of the first ten data samples is used as the baseline value as suggested in the literature [50]. This is equal to the average pupil size during the first 166 ms of the recording, which is well below the latency of pupillary responses to a stimulus, as found in relevant psychological studies [5, 20]. If more samples would be used for computing the baseline value, the pupil size could already be affected by the experiment's external stimulus and lead to a distorted baseline value.

3.2.2 Feature Calculation. An eye tracker provides information about the viewer's eye gaze in terms of fixations and saccades. Additionally, information about the participants' pupil size is recorded. Goldberg and Helfman [24] described basic eye tracking measures based on fixations and saccades for comprehensible eye gaze data processing.

The fixation-based measures are commonly used in eye gaze data analysis [8, 35, 47, 70] as they are an indicator of how much data is actually processed by the user [63]. Saccadic features are also included since they are useful to reveal trends in the users' attention patterns [24, 35, 70]. Pupil dilation is included since it has been found to provide insight into decision-making [71] and reveals the relevance of search results [56]. Additionally, saccadic velocity is used to provide insight into task difficulty and variation in mental workload [14, 15].

Based on these measures, basic statistics (mean, standard deviation, sum) are calculated for each time step as $v_t = f(b_0, b_t)$ where v is the feature value at time step t and is calculated as a function f (sum, mean, standard deviation) of the basic metric (e.g. pupil size) values from the beginning of the data b_0 to the current time step b_t. All calculated features are listed in Table 1.

The time series obtained by calculating these features are illustrated in Figure 3.

Gaze Measure	Feature
Fixations	Fixation rate: *fixations per millisecond* Number of Fixations Fixation duration: *sum, mean and standard deviation*
Saccades	Saccade length: *sum, mean and standard deviation* Relative saccade angle: *sum, mean and standard deviation* Absolute saccade angle: *sum, mean and standard deviation* Saccadic Velocity: *Saccade length over time (pixel/milliseconds)*
Pupil Diameter	Left eye: *absolute value, mean and standard deviation* Right eye: *absolute value, mean and standard deviation*

Table 1. The feature set contains 21 features, which are calculated for each time step in the data.

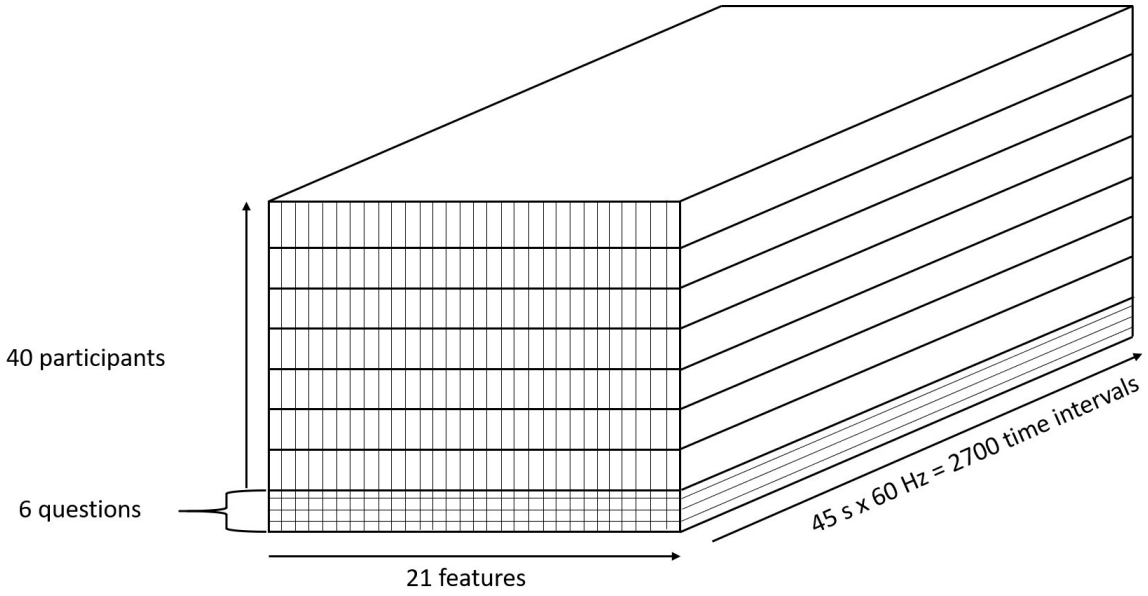

Fig. 3. Illustration of the time series obtained by calculating the aforementioned features for both graph types. 21 features have been calculated from each participant's eye gaze data while he or she answered a question. The limit for answering a question was 45 seconds. If a participant did not require that long, the time series was zero-padded. The participants' eye movements were tracked at a sampling frequency of 60 Hz. Overall, each graph type dataset contained 240 multivariate time series, which have been fed to the model along with a label to indicate if the corresponding question has been answered correctly (positive class) or not (negative class).

3.2.3 Baseline Classifier. LR (Logistic Regression) has been found to work well with eye gaze data by multiple studies [7, 35, 70] and is therefore chosen as the baseline classifier for this research. As it is not specifically designed to handle

time series data, the features used in the multivariate time series dataset (See also Table 1) have been calculated over the whole time segment, i.e., the length of the recorded data. The results of the baseline classifier are also cross-validated on ten folds. The data has been split on a "user-basis", i.e., the training set contained the data of 36 participants while the test set contained the data of the remaining four participants.

3.2.4 Time Series Classification. Three deep learning classifiers, MLSTM-FCN, ResNet, and FCN, have been trained for this study. These classifiers have been chosen since they have shown high performance on time series classification problems.

The first model, an MLSTM-FCN with attention mechanism, has been proposed by [37] and outperformed the respective state-of-the-art model on 23 out of 35 publicly available datasets in the study. Additionally, the univariate variant of this model [36] has been successfully used in activity recognition [48] and aggressive driving detection [54]. An attention mechanism further improves the LSTM's ability to learn long term-dependencies by contextualizing currently and previously observed data and was originally proposed for neural machine translation of text [3]. Attention conditions a context vector τ on the target sequence y. The context vector contains weights which are adjusted according to the correlation between the elements in the sequence. This happens while the network is trained. When inferring unseen data, the context vector τ is used to compute how strongly the elements of the unseen data correlate with the elements in τ. This value is then used to approximate the target value of the new data. [3]

Wang et al. [78] have proposed variants of ResNet and FCN, which have been successful in computer vision, adapted to TSC. A review of various deep learning models for TSC found that ResNet and FCN are the superior performing classifiers after being applied to twelve publicly available multivariate time series datasets [33]. Both classifiers have been successfully used in further research. ResNet has shown good performance in activity recognition [48], whereas FCN has been considered in another review of deep learning for TSC and ECG (electrocardiogram) classification [61].

The common hyperparameter of the three evaluated models are summarized in Table 2. The models have been used with the same hyperparameter and implementation as in the original studies. The experiments are conducted on a single NVIDIA GeForce GTX 1070. To counter the class imbalance in the dataset, class weights inspired by King and Zeng [40] are assigned during the training of all models. To avoid overfitting, 10-fold cross-validation has been applied, and the training has been stopped early if the validation loss did not improve for 100 epochs[16, 21].

The used code will be published after the publication of this study to enhance the reproducibility of results.

Hyperparameter	Value
Initial Learning Rate	0.001
Optimizer	Adam
Batch Size	128
Max. Epochs	1000

Table 2. Common hyperparameter of the three used models. The initial learning rate is reduced after training loss did not improve for 50 epochs. Training is stopped if the validation loss did not improve for 100 epochs.

To ensure generalizable results, the models have been evaluated regarding efficiency and effectiveness. The data is split into training and test sets using 10-fold cross-validation, resulting in a training set containing 36 participants' data and a test set containing data of four participants. As the negative class only accounts for 22.29 % of the data samples, the F1-score is chosen over accuracy to evaluate the models' effectiveness. The F1-score is calculated as a weighted harmonic mean of precision (P) and recall (R) as suggested in [57]: $F_1 = \frac{2PR}{P+R}$. The models have been trained to classify

if the user will answer a particular question correctly (positive class) or not (negative class). Which performance in terms of F1-score is sufficient, largely depends on the specific application of a user-adaptive information visualization system. Since this study aims to provide a basis for developing such systems instead of developing one, no statement regarding which F1-score is considered a success can be made. The models' efficiency is evaluated based on the models' inference time per sample in milliseconds (ms).

4 RESULTS

The dataset is split by graph type, as it can be expected that eye movements vary between different graphs due to the different search array [63]. At first, the experiments are conducted by using the three classifiers described in Section 3.2.4. Secondly, the data is truncated to sequence lengths from one to ten seconds to evaluate if task success can be predicted within the first ten seconds of user interaction with the graph. Current empirical evidence suggests that 10% of a user's interaction can provide early prediction of user cognitive abilities in visualization tasks [11]. An online user-adaptive system that requires more than ten seconds to infer the users' need for support appears impractical and, thus, sequences that are longer than ten seconds are not evaluated.

All models are trained to predict if the users' answers to a particular question would be correct or incorrect (binary classification).

The results are compared with the baseline classifier, and a two-sample Wilcoxon test under the non-normality assumption has been applied to check for significant differences in the classifier's performance regarding F1-score.

4.1 Full-length Sequences

At first, the models have been trained on the full-length time series with 2700 (45 seconds x 60 Hz) time steps. If a participant has finished earlier than the maximum time limit of 45 seconds, the time series is padded with zeros [33, 37].

4.1.1 Effectiveness on the Circular & the Organizational Graph. MLSTM-FCN substantially outperforms the other individual-independent models on the circular graph. The differences are highly significant ($p-value < 0.01$). It reaches an F1-score of 0.83, while its standard deviation is relatively small and comparable to the standard deviation of ResNet, which reaches an F1-score of 0.64. This is slightly better than the F1-scores of FCN and LR as the baseline classifier, which both reach an F1-score of 0.63. The differences between ResNet, FCN and LR are not significant ($p-value > 0.05$).

On the organizational graph, MLSTM-FCN significantly outperforms ResNet and the baseline classifier, but the advantage compared to FCN is not significant anymore. ResNet and FCN also significantly outperform the baseline classifier, which reaches an F1-score of 0.58.

All four classifiers show substantial standard deviations in their F1-scores, indicating that the distribution of the data samples into training and test set impacts the classifiers' performance.

The F1-scores obtained by the classifiers on both graph types are depicted in Figure 4.

4.1.2 Efficiency. The efficiency of the models is measured by its inference time per sample. In that process, the time required while applying the model to the test set is recorded. This value is divided by the number of samples in the test set to obtain the inference time per sample.

The used classifiers show similar inference times on both graph types. Since LR is a statistical classifier, it is not surprising that it requires very short inference times. While the inference times of ResNet ($\sim$ 4 ms) and FCN ($\sim$ 16 ms) are also short, MLSTM-FCN requires the most inference time with about 105 ms per sample.

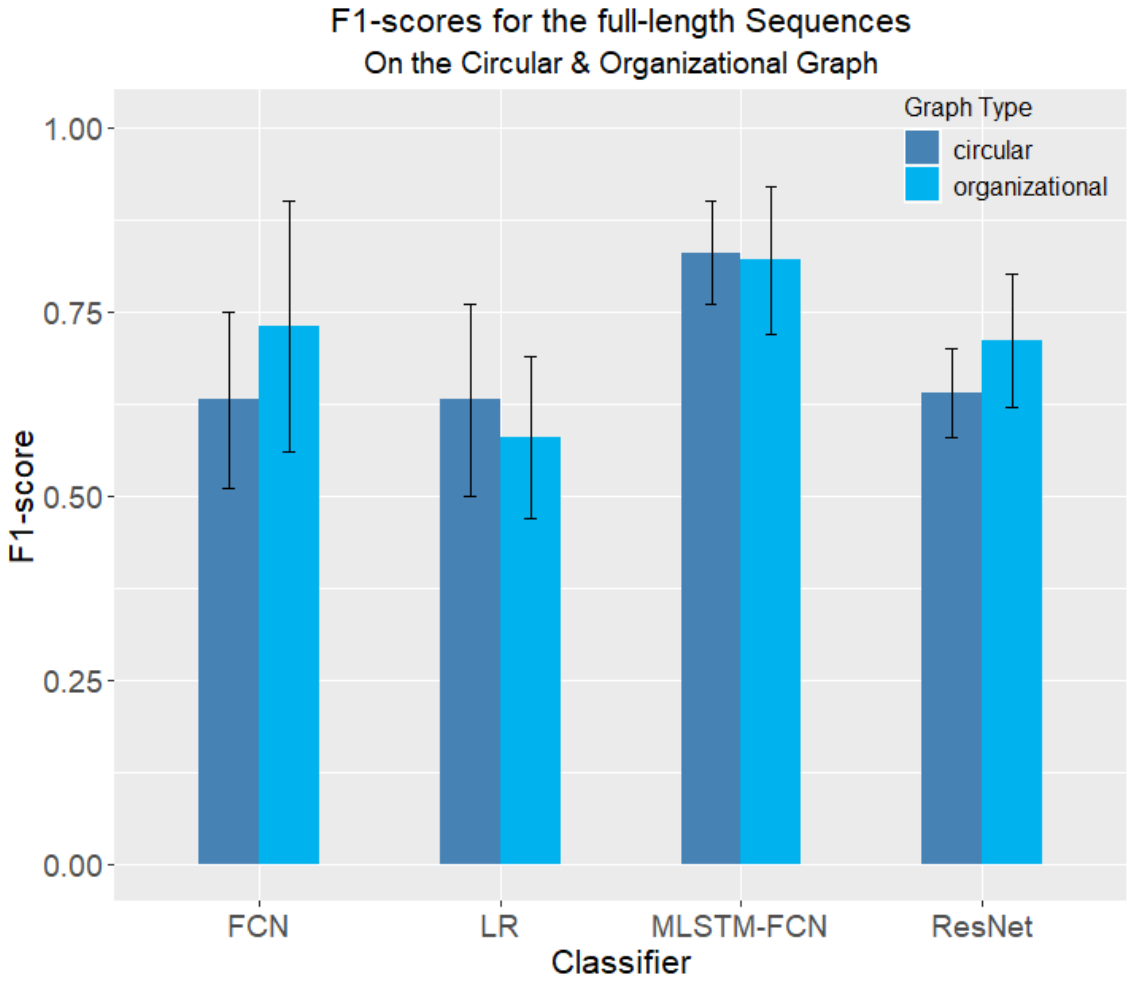

Fig. 4. Illustration of the TSC models' F1-scores on the 45 seconds long sequences for the circular and organizational graph. MLSTM-FCN shows similar performance on both graph types, ResNet and FCN perform better on the organizational graph, where all three deep learning models outperform the baseline classifier on the organizational graph. On the circular graph, ResNet and FCN show similar performance as the baseline classifier. For detailed results refer to Table 3

.

4.2 Time Intervals

Similar to other studies [4, 29, 70], the full-length sequences are truncated to sequence lengths of one to ten seconds to evaluate the classifiers' performance on short time intervals and explore the feasibility of real-time interventions of user-adaptive visualization systems.

4.2.1 Effectiveness on the Circular Graph. The F1-score on the time intervals decreased for all models. However, the extent of the performance decreased differs among the classifiers. MLSTM-FCN is again the superior performing classifier on the circular graph, significantly outperforming ResNet, FCN, and the baseline classifier. This corresponds

Classifier	Graph	F_1	AUROC	Inf. Time	P	R
MLSTM-FCN	C	0.83 (0.07)	0.89 (0.07)	104.58 (49.17)	0.83 (0.07)	0.83 (0.07)
	O	0.82 (0.10)	0.83 (0.13)	103.75 (47.92)	0.82 (0.10)	0.82 (0.10)
ResNet	C	0.64 (0.06)	0.62 (0.05)	15.83 (0.00)	0.67 (0.11)	0.62 (0.05)
	O	0.71 (0.09)	0.71 (0.16)	15.83 (0.00)	0.73 (0.10)	0.71 (0.16)
FCN	C	0.63 (0.12)	0.62 (0.09)	4.17 (0.00)	0.62 (0.18)	0.62 (0.09)
	O	0.73 (0.17)	0.74 (0.17)	4.17 (0.00)	0.73 (0.19)	0.71 (0.17)
LR	C	0.63 (0.13)	0.73 (0.16)	0.83 (0.00)	0.67 (0.18)	0.63 (0.11)
	O	0.58 (0.11)	0.67 (0.15)	0.83 (0.00)	0.65 (0.18)	0.59 (0.08)

Table 3. F_1-score (F_1), Area Under the Receiver Operating Characteristics (AUROC), Inference Time (Inf. Time), Precision (P) and Recall (R) along with their standard deviations in parenthesis of the classifiers' performance obtained after 10-fold cross-validation for the circular (C) and the organizational (O) graph. The inference time is stated in milliseconds per sample. MLSTM-FCN is the superior performing classifier in terms of F_1-score. All deep learning classifiers at least match the F1-scores of the baseline classifier. The FCN is the deep learning classifier with the least inference time; however, it can not outperform the statistical baseline classifier.

to a decrease of 12 % compared to its performance in the full-length sequences on the circular graph. ResNet and FCN suffer from a more significant performance decrease of 23 % and 25 %, respectively. While ResNet and FCN also significantly outperform the baseline classifier, the differences between ResNet and FCN are not significant.

The F1-scores obtained by the classifiers on the circular graph are depicted in Figure 5.

4.2.2 Effectiveness on the Organizational Graph. In contrast to the models trained on the full-length sequences, there are no substantial differences between the two different graph types' performances. MLSTM-FCN suffers from a 9 % performance decrease on the organizational graph but still outperforms all other classifiers significantly. ResNet and FCN, which have shown significantly diverging performance on the two graphs on the full-length sequences, suffer from a major decrease of 31 % and 37 % respectively, reaching average F1-scores of 0.49 and 0.46. These values are equal (ResNet) or minimally smaller (FCN) than their performance on the circular graph. While MLSTM-FCN and ResNet still significantly outperform the baseline classifier, the performance advantage of FCN compared to the baseline is not significant.

The F1-scores obtained by the classifiers on the organizational graph are depicted in Figure 6.

4.2.3 Efficiency. Compared to their inference times per sample on the full-length sequences, the inference times of ResNet and FCN remain stable on the time intervals. They also do not correlate with the sequence length. In contrast, the inference time of MLSTM-FCN shows a linear increase of about 184 ± 12 ms additional inference time per second of additional sequence length. While the inference time of 95 ms for a one-second-long sequence is comparable with the average inference time on the full-length sequences (105 ms), its inference time reaches approximately 1.3 seconds for a ten seconds long sequence.

The inference times are depicted in Figure 7.

4.3 Summary of Results

MLSTM-FCN is consistently yielding the highest F1-scores and significantly outperforms the ResNet, FCN, and the baseline classifier. ResNet and FCN show similar F1-scores; however, their performance is not substantially better than

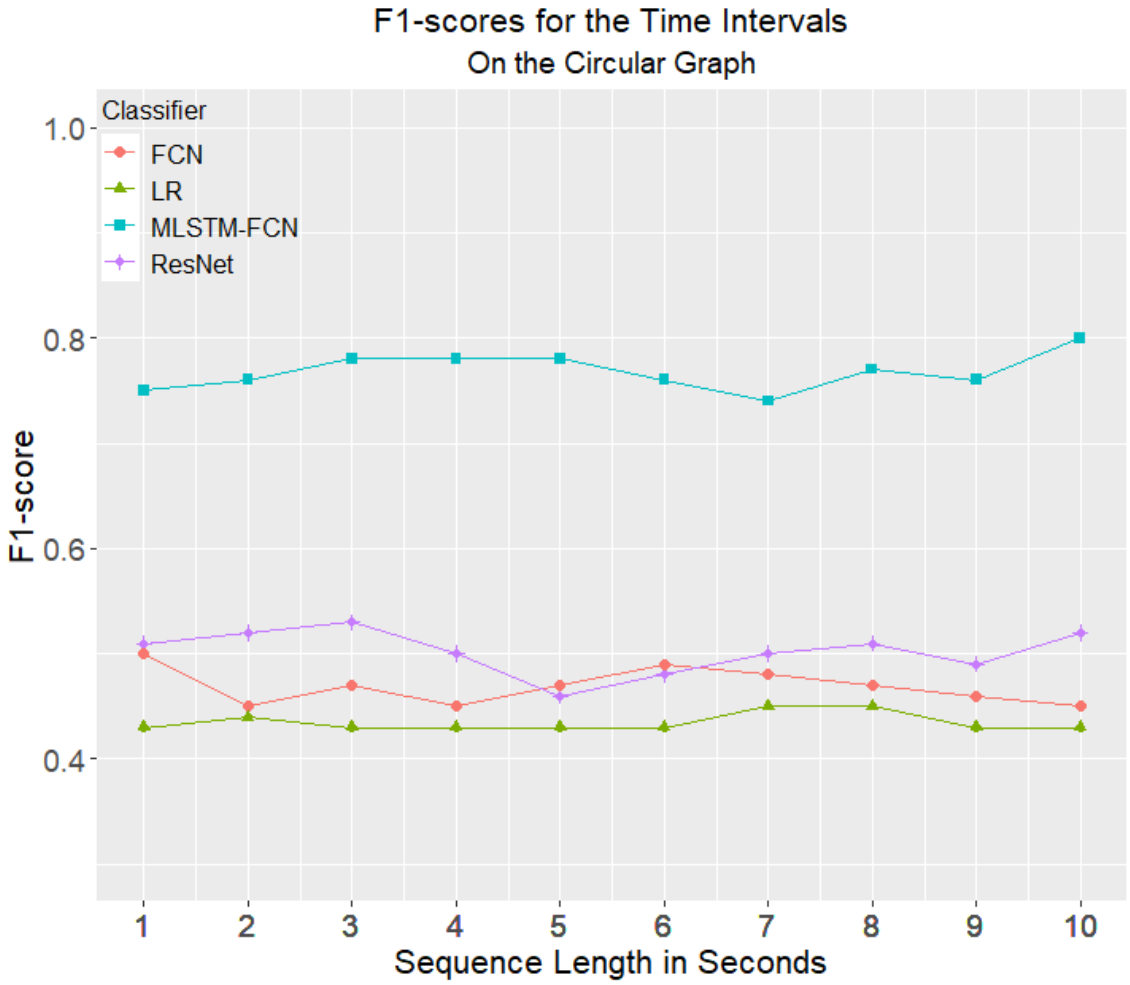

Fig. 5. MLSTM-FCN achieves the highest F1-score on the time intervals per sequence length on the circular graph. ResNet and FCN cannot outperform the baseline classifier.

the baseline classifier on three of four experiments. On the time intervals of the organizational graph and the full-length sequences of the circular graph, the baseline classifier even matches the performance of ResNet and FCN.

Regarding inference time, MLSTM-FCN has a clear disadvantage compared to the other classifiers. While its inference time on the full-length sequences is still acceptable in human-computer interaction, its inference time on time intervals increases linearly to more than one second on a ten-second long sequence. While ResNet shows already good inference times, FCN reaches the best inference time per sample of all deep learning classifiers with about 4 ms per sample. Unsurprisingly, none of the deep learning models can outperform the statistical baseline classifier in inference time. ResNet and FCN show stable inference times on full-length sequences as well as on time intervals.

No consistent observation can be made regarding any performance differences between the two graph types. On the full-length sequences of the organizational graph, the performance of ResNet and FCN improves substantially compared

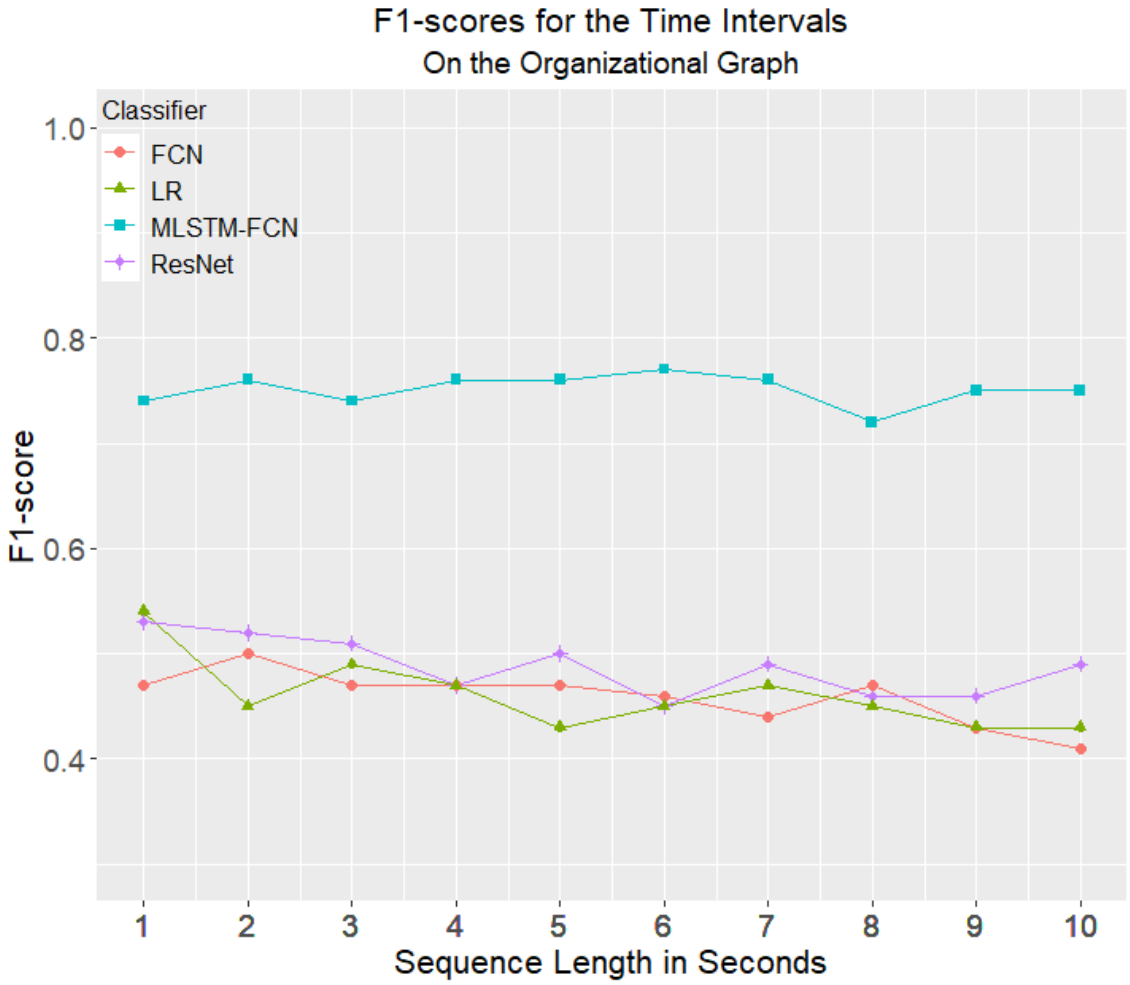

Fig. 6. MLSTM-FCN is again the superior performing classifier on the time intervals of the organizational graph. ResNet slightly outperforms the baseline classifier, while shows similar performance as LR.

to the circular graph. In contrast to that, the performance of ResNet and FCN on time intervals is not affected by the graph type. MLSTM-FCN shows no performance differences between the two graph types.

In summary, MLSTM-FCN is the classifier to choose when the highest accuracy possible is the primary concern, but ResNet and FCN should still be considered when near real-time inference is required. Usually, deep learning models for TSC outperform the statistical baseline classifier for the used eye gaze data.

4.4 Limitations of the Study

Any user evaluation has its limitations, and this study is no exception. It might be argued that the use of a low-cost eye tracker with lower sampling rates is insufficient for scientific purposes. However, there is evidence that the EyeTribe's accuracy and sampling frequency are sufficient for performing fixation, pupillometry, and saccade analyses [60, 72].

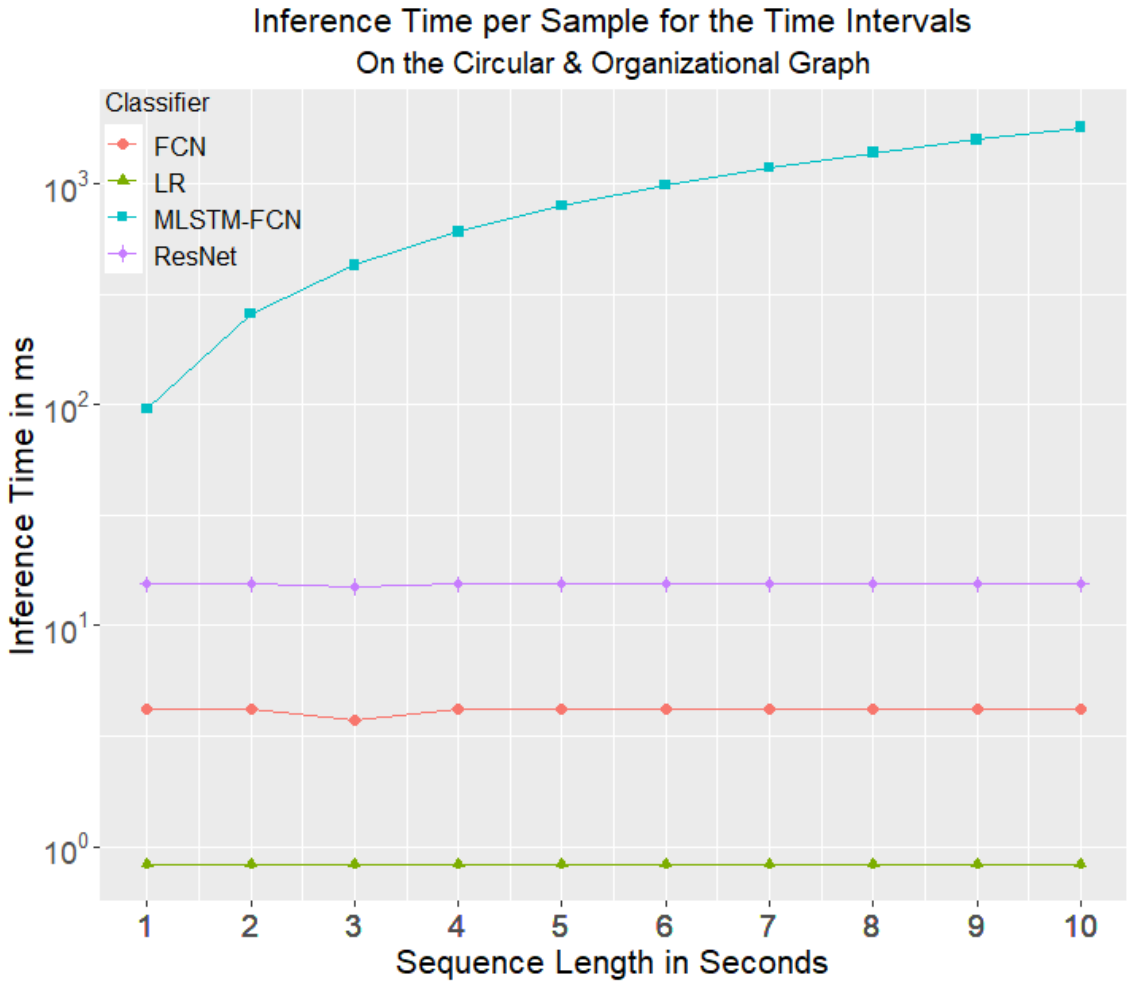

Fig. 7. While ResNet and FCN show a stable inference time per sample of about 16 ms and 4 ms respectively, the inference time of MLSTM-FCN increases approximately linear to more than 1000 ms on a ten second long sequence. No deep learning model can match the inference time of the baseline classifier.

In this study, the models, including their implementation, have been adopted from the original studies [37, 78], and have not been tuned to fit the eye tracking dataset. The reported performances may not reflect the full potential of the models. However, as the advantage of MLSTM-FCN is significant, the relative performance win over other deep learning models is likely to hold after hyperparameter tuning. Additionally, inference time may be impacted by the implementation of the models. In consequence, the way the authors of the original studies have implemented the models' may influence the performance regarding inference time presented in Section 4.1.2 and Section 4.2.3.

Since this study focuses on circular and organizational graphs, the generalizability of results to other graph types is limited. Since the classifiers have been trained to predict task success based on the users' answers to simple questions as an indicator of user comprehension, a comprehensive understanding of a graph may rely on the user's familiarity

with different types of information visualization systems and search tasks. Future research can explore the relationships among these variables for personalizing real-time user-adaptive interactions in visualization tasks.

5 DISCUSSION

In this study, three different deep learning classifiers are evaluated on users' eye gaze data when interacting with a circular and organizational graph. In particular, three deep learning classifiers for TSC are used to predict the participants' task success based on their answers to various questions about the graph's information.

5.1 To What Extent can Visual Search Task Success be Predicted from Sequential Eye Gaze Data?

This study results reveal that MLSTM-FCN, as the superior performing classifier, can infer task success from eye gaze data with an F1-score of 0.82 on the full-length sequences.

The results suggest that the reason for the superior performance of MLSTM-FCN is its LSTM block since its fully convolutional block is identical to the FCN. As observed in Fawaz et al. [19] and Wang et al. [78], ResNet and FCN perform similarly in terms of F1-score, which is substantially worse than the MLSTM-FCN's F1-scores.

All used classifiers yielded substantial standard deviations in their F1-scores (see Figure 4), indicating that the distribution of the data samples into training and test set impacts the performance of the classifiers. This could suggest that some participants' eye gaze patterns differed fundamentally from others' eye gaze patterns, making it difficult for the models to generalize well enough.

On the other hand, MLSTM-FCN requires a substantially higher inference time per sample than the other classifiers. While the LSTMs inference time on similar problems is sparsely reported in the literature, the results on the time intervals suggest that its inference time correlates with the sequence length (See also Figure 7).

MLSTM-FCN shows high inference times, requiring more than 100 ms per sample on full-length sequences and up to 1000 ms for time intervals. Even though an inference time of about 100 ms is still considered real-time in human-computer interaction [4, 51, 56], inference times of more than 1000 ms might be considered too long for modern on-line user-adaptive information visualization systems.

However, for a binary classification problem, an accuracy of 0.8 might not be considered good enough to drive on-line data visualization systems. One factor limiting the model's accuracy could be the different tendencies of the features over time. Figure 8 shows two plots of a participant's pupil size while giving a correct answer vs. two plots of the pupil size while giving an incorrect answer. Incorrect answer 1 and correct answer 1 show a similar big variance in pupil size, while incorrect answer 2 and correct answer 2 show less variance. Plotting on a sample basis has shown similar behavior in about 10 % of the cases and was also observed on other features like the number of fixations. These differences in time series belonging to the same class have two implications. At first, ideally 50 % of these cases should be included in each, the training and the test set. However, this was not ensured when splitting the data. Secondly, features showing the described behavior are less informative than others. Therefore, excluding them from the feature set or using them AOI-based (change of pupil size within a certain AOI) like in Steichen et al. [70] could lead to better performance of the classifier.

5.2 Is it Feasible to Infer the Users' Task Success within the First Ten Seconds of Interaction?

After evaluating the classifiers' performance on different sequence lengths from one to ten seconds, the results suggest that the sequence length used to infer task success has no substantial impact on the performance of the classifiers within

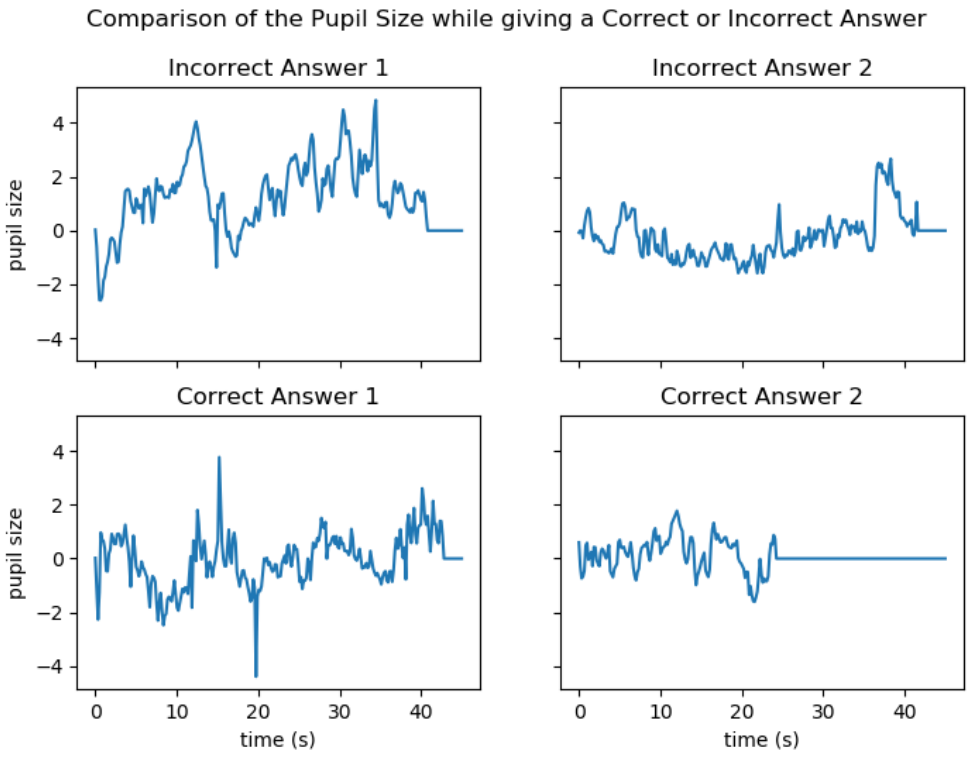

Fig. 8. Two plots of the pupil size while giving a correct answer vs. two plots of the pupil size while giving an incorrect answer. Note that subtractive baseline correction has been applied, i.e., the change in pupil size is illustrated in the plots. The eye tracker records the pupil size without a specific measure. Since the time series have been zero-padded to an equal length if the participant needed less than 45 seconds, the plot shows a constant pupil size of zero in the end.

that period. The F1-scores of all evaluated classifiers oscillate with a maximum standard deviation of 0.026 (ResNet, baseline classifier: 0.032) on the organizational and 0.02 (ResNet, baseline classifier: 0.008) on the circular graph.

These results indicate that most users acquire the essential information required to complete a given task successfully within the first second of interaction with the graph. Using longer sequences than one second does not result in significantly better classification performance. This suggests that the first second of interaction with a graph is critical for intervention by the system.

However, since the F1-scores for full-length sequences (45 s) are substantially better than on shorter sequences, it might be possible that some users require more than ten seconds to finalize the acquisition of the essential information. According to the model proposed by Freedman and Shah [22] and Körner et al. [43], the individual differences such as prior knowledge and experience in interaction with circular and organizational graphs can affect the users' graph comprehension. Some users may take longer to establish relations between the graph's elements and comprehend the information. Importantly, if the sequence used for classification is too short to comprise the whole process of the users' graph comprehension, the predictions could become inaccurate.

It is worth noting that Steichen et al. [70] found that classification accuracy is better on shorter time sequences at the beginning of the data for some classification tasks. However, there are several differences in this study. While their study aims to infer cognitive abilities, this study attempts to infer graph comprehension based on the predicted users' success on a visual search task. Although the use of longer sequences to predict cognitive abilities could lead to more noise, answers to a particular question might be determined at any time, and even change during the interaction.

Such a model can drive user-adaptive information visualization systems by providing a measure to decide if the user understands the information imparted by the graph or if intervention by the system is necessary. A recent study shows that users benefit from system-driven customization about the information content presented in an information visualization system, which depends on the user characteristics of visualization literacy and locus of control [44]. However, which interventions are undertaken is not directly addressed in this study. The interventions depend on the actual system and application, including (but not limited to) proposing alternative graph types, add reference lines to the graph or highlight information, presenting complementary information (e.g., text), and changing the way how the information is imparted (e.g., switch from a graph to a video) from the perspectives of user interface design. This study aims to build computational models to infer the users' task success, using varying time series eye gaze data.

5.3 Implications for User-adaptive Visualization Systems

The results of this study suggest that a computational model can infer task success from eye gaze data. Given this finding, such a computational model can also be used to infer the users' need for support during interaction with a graph and trigger appropriate interventions in user-adaptive information visualization systems. As revealed in Lallé and Conati [44] some users benefit from system-driven customization about the information content presented in an information visualization system due to individual differences. In contrast to previous studies of user interaction data such as mouse clicks and search behavior data [25, 26], such a system requires no other interaction data than eye gaze. This facilitates the design of user-adaptive visualization systems since further interaction data like mouse clicks is not required.

6 CONCLUSION & FUTURE WORK

Corroborating previous studies [68, 70], this study provides encouraging results to conclude that the inference of the users' task success from eye gaze data is possible. This study assesses task success based on the estimation if the user has correctly acquired the graph's information and shows that this approach is suitable to drive on-line user-adaptive information visualization systems.

The used deep learning classifiers perform worse on time intervals than on full-length sequences. MLSTM-FCN can infer task success with higher accuracy than a baseline classifier and previous models for this purpose. To consider the model's high inference time during system design, it is necessary to explore how the inference time of MLSTM-FCN evolves on sequences between 10 and 45 seconds. Since this research question was out of this study's scope, this needs to be done in future work.

The presented results have been obtained using general deep learning models for TSC and simple eye gaze metrics. The full potential of the models for driving on-line user-adaptive information visualization systems can be realized by adding AOI (areas of interest) features and specific consideration of user characteristics [67] and fine-tuning the models towards the objective of integrating such a model into user-adaptive systems.

7 ACKNOWLEDGEMENTS

This research was supported partially by the Australian Government through the Australian Research Council's Linkage Projects funding scheme (project LP140100995). The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council.

REFERENCES

[1] John R. Anderson and Gordon H. Bower. 1974. *Human Associative Memory.* Psychology Press, New York. https://doi.org/10.4324/9781315802886

[2] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. 2017. The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery* 31, 3 (2017), 606–660. https://doi.org/10.1007/s10618-016-0483-9

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR).* https://arxiv.org/pdf/1409.0473

[4] Oswald Barral, Manuel J A A Eugster, Tuukka Ruotsalo, Michiel M Spapé, Ilkka Kosunen, Niklas Ravaja, Samuel Kaski, and Giulio Jacucci. 2015. Exploring peripheral physiology as a predictor of perceived relevance in information retrieval. In *Proceedings of the International Conference on Intelligent User Interfaces - IUI '15.* ACM, New York, 389–399. https://doi.org/10.1145/2678025.2701389

[5] Oliver Bergamin and Randy H. Kardon. 2003. Latency of the pupil light reflex: Sample rate, stimulus intensity, and variation in normal subjects. *Investigative Ophthalmology and Visual Science* 44, 4 (2003), 1546–1554. https://doi.org/10.1167/iovs.02-0468

[6] Tanja Blascheck, Lindsay MacDonald Vermeulen, Jo Vermeulen, Charles Perin, Wesley Willett, Thomas Ertl, and Sheelagh Carpendale. 2019. Exploration strategies for discovery of interactivity in visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 2 (2019), 1407–1420. https://doi.org/10.1109/TVCG.2018.2802520

[7] Daria Bondareva, Cristina Conati, Reza Feyzi-Behnagh, Jason M. Harley, Roger Azevedo, and François Bouchet. 2013. Inferring learning from gaze data during interaction with an environment to support self-regulated learning. In *Artificial Intelligence in Education*, H.c. Lane, K. Yacef, J. Mostow, and Pavlik P. (Eds.). Springer, Berlin, 229–238. https://doi.org/10.1007/978-3-642-39112-5_24

[8] Matt Canham and Mary Hegarty. 2010. Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction* 20, 2 (2010), 155–166. https://doi.org/10.1016/j.learninstruc.2009.02.014

[9] Patricia A Carpenter and Priti Shah. 1998. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied* 4, 2 (1998), 75–100. https://doi.org/10.1037/1076-898X.4.2.75

[10] Michael J. Cole, Jacek Gwizdka, Chang Liu, Nicholas J. Belkin, and Xiangmin Zhang. 2013. Inferring user knowledge level from eye movement patterns. *Information Processing & Management* 49, 5 (2013), 1075–1091. https://doi.org/10.1016/j.ipm.2012.08.004

[11] Cristina Conati, Sébastien Lallé, Md Abed Rahman, and Dereck Toker. 2020. Comparing and Combining Interaction Data and Eye-tracking Data for the Real-time Prediction of User Cognitive Abilities in Visualization Tasks. *ACM Transactions on Interactive Intelligent Systems* 10, 2 (2020), 1–41. https://doi.org/10.1145/3301400

[12] Cristina Conati and Heather Maclaren. 2008. Exploring the role of individual differences in information visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '08).* ACM, New York, 199–206. https://doi.org/10.1145/1385569.1385602

[13] Çağla Çığ Karaman and Tevfik Metin Sezgin. 2018. Gaze-based predictive user interfaces: Visualizing user intentions in the presence of uncertainty. *International Journal of Human-Computer Studies* 111 (2018), 78–91. https://doi.org/10.1016/j.ijhcs.2017.11.005

[14] Leandro L. Di Stasi, Mauro Marchitto, Adoracíon Antolí, Thierry Baccino, and José J. Cañas. 2010. Approximation of on-line mental workload index in ATC simulated multitasks. *Journal of Air Transport Management* 16, 6 (2010), 330–333. https://doi.org/10.1016/j.jairtraman.2010.02.004

[15] Leandro L Di Stasi, Rebekka Renner, Peggy Staehr, Jens R Helmert, Boris M Velichkovsky, José J Cañas, Andrés Catena, and Sebastian Pannasch. 2010. Saccadic peak velocity sensitivity to variations in mental workload. *Aviation, Space, and Environmental Medicine* 81, 4 (2010), 413–417.

[16] Tom Dietterich. 1995. Overfitting and Undercomputing in Machine Learning. *ACM Comput. Surv.* 27, 3 (sep 1995), 326–327. https://doi.org/10.1145/212094.212114

[17] Geoffrey B. Duggan and Stephen J. Payne. 2009. Text Skimming: The Process and effectiveness of foraging through text under time pressure. *Journal of Experimental Psychology: Applied* 15, 3 (2009), 228–242. https://doi.org/10.1037/a0016995

[18] Philippe Esling and Carlos Agon. 2012. Time-series data mining. *Comput. Surveys* 45, 1 (2012), 1–34. https://doi.org/10.1145/2379776.2379788

[19] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. Muller. 2019. Deep Neural Network Ensembles for Time Series Classification, In 2019 International Joint Conference on Neural Networks (IJCNN). *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–6. https://doi.org/10.1109/IJCNN.2019.8852316

[20] Richard Feinberg and Edward Podolak. 1965. *Latency of pupillary reflex to light stimulation and its relationship to aging.* Charles C Thomas, Springfield, IL, 326–339.

[21] Matthias Feurer and Frank Hutter. 2019. Hyperparameter Optimization. In *Automated Machine Learning: Methods, Systems, Challenges*, Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren (Eds.). Springer International Publishing, Cham, 3–33. https://doi.org/10.1007/978-3-030-05318-5_1

[22] Eric G. Freedman and Priti Shah. 2002. Towards a model of knowledge-based graph comprehension. In *Diagrammatic Representation and Inference*, Mary Hegarty, Bernd Meyer, and N. Hari Narayanan (Eds.). Lecture Notes in Computer Science, Vol. 2317. Springer, Berlin, 18–30. https://doi.org/10.1007/3-540-46037-3

[23] Joseph Goldberg and Jonathan Helfman. 2011. Eye tracking for visualization evaluation: Reading values on linear versus radial graphs. *Information Visualization* 10, 3 (2011), 182–195. https://doi.org/10.1177/1473871611406623

[24] Joseph H Goldberg and Jonathan I. Helfman. 2010. Comparing information graphics. In *Proceedings of the Workshop on Beyond Time and Errors: Novel evaluation methods for information visualization (BELIV'10)*. ACM, New York, 71–78. https://doi.org/10.1145/2110192.2110203

[25] David Gotz and Zhen Wen. 2009. Behavior-driven Visualization Recommendation. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI '09)*. ACM, New York, 315–324. https://doi.org/10.1145/1502650.1502695

[26] Beate Grawemeyer. 2006. Evaluation of ERST: An external representation selection tutor. In *Diagrammatic Representation and Inference*, D. Barker-Plummer, R. Cox, and N. Swoboda (Eds.). Springer, Berlin, 154–167. https://doi.org/10.1007/11783183_21

[27] Günther Greiner and Kai Hormann. 1997. Interpolating and approximating scattered 3D-data with hierarchical tensor product B-splines. *Surface Fitting and Multiresolution Methods* 3 (1997), 163–172.

[28] Vladimir Guchev, Paolo Buono, and Cristina Gena. 2018. Towards Intelligible Graph Data Visualization Using Circular Layout. In *Proceedings of the International Conference on Advanced Visual Interfaces* (Castiglione della Pescaia, Grosseto, Italy) *(AVI '18)*. ACM, New York, Article 63, 3 pages. https://doi.org/10.1145/3206505.3206592

[29] Jacek Gwizdka, Rahilsadat Hosseini, Michael Cole, and Shouyi Wang. 2017. Temporal dynamics of eye-tracking and EEG during reading and relevance decisions. *Journal of the Association for Information Science and Technology* 68, 10 (2017), 2299–2312. https://doi.org/10.1002/asi.23904

[30] Roy S. Hessels, Diederick C. Niehorster, Chantal Kemner, and Ignace T.C. Hooge. 2017. Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). *Behavior Research Methods* 49, 5 (2017), 1802–1823. https://doi.org/10.3758/s13428-016-0822-1

[31] Md Zakir Hossain, Tom Gedeon, Sabrina Caldwell, Leana Copeland, Richard Jones, and Christopher Chow. 2018. Investigating differences in two visualisations from observer's fixations and saccades. In *Proceedings of the Australasian Computer Science Week Multiconference (ACSW '18)*. ACM, New York, 1–4. https://doi.org/10.1145/3167918.3167933

[32] Md Zakir Hossain, Tom Gedeon, and Ramesh Sankaranarayana. 2019. Using temporal features of observers' physiological measures to distinguish between genuine and fake smiles. *IEEE Transactions on Affective Computing* 11, 1 (2019), 163–173. https://doi.org/10.1109/TAFFC.2018.2878029

[33] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre Alain Muller. 2019. Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery* 33, 4 (2019), 917–963. https://doi.org/10.1007/s10618-019-00619-1

[34] Anthony Jameson. 2007. Adaptive interfaces and agents. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (2 ed.), Andrew Sears and Julie A. Jacko (Eds.). CRC Press, Boca Raton, FL, 433–458. https://doi.org/10.1201/9781410615862

[35] Samad Kardan and Cristina Conati. 2012. Exploring Gaze Data for Determining User Learning with an Interactive Simulation. In *User modeling, adaptation, and personalization*, Judith Masthoff, Bamshad Mobasher, Michel C. Desmarais, and Roger Nkambou (Eds.). Springer, Berlin, 126–138. https://doi.org/10.1007/978-3-642-31454-4_11

[36] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. 2017. LSTM fully convolutional networks for time series classification. *IEEE Access* 6 (2017), 1662–1669. https://doi.org/10.1109/ACCESS.2017.2779939

[37] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. 2019. Multivariate LSTM-FCNs for time series classification. *Neural Networks* 116 (2019), 237–245. https://doi.org/10.1016/j.neunet.2019.04.014

[38] Daniel Keim, Jorn Kohlhammer, Geoffrey Ellis, and Florian Mansman (Eds.). 2010. *Mastering the information age: Solving problems with visual analytics*. Eurographics Association, Goslar, Germany.

[39] Suzanne Kieffer and Université catholique de Louvain. 2017. ECOVAL: Ecological Validity of Cues and Representative Design in User Experience Evaluations. *Association for Information Systems transactions on human-computer interaction* 9, 2 (2017), 149–172. https://doi.org/10.17705/1thci.00093

[40] Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political Analysis* 9, 2 (2001), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

[41] Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review* 95, 2 (1988), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

[42] Roger Kirk. 2013. *Kirk, Roger E. (2013). Experimental design: Procedures for the behavioral sciences (4th ed.). Thousand Oaks, CA: Sage.*

[43] Christof Körner, Margit Höfler, Barbara Tröbinger, and Iain D. Gilchrist. 2014. Eye movements indicate the temporal organisation of information processing in graph comprehension. *Applied Cognitive Psychology* 28, 3 (2014), 360–373. https://doi.org/10.1002/acp.3006

[44] Sébastien Lallé and Cristina Conati. 2019. The role of user differences in customization: A case study in personalization for infovis-based content. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI '19)*. ACM, New York, 329–339. https://doi.org/10.1145/3301275.3302283

[45] Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. 2016. Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data. In *Proceedings of the International Joint Conference on Artificial Intelligence* (New York, USA) *(IJCAI '16)*. AAAI Press, 2529–2535.

[46] Chang Liu, Ying-Hsang Liu, Tom Gedeon, Yu Zhao, Yiming Wei, and Fan Yang. 2019. The effects of perceived chronic pressure and time constraint on information search behaviors and experience. *Information Processing & Management* 56, 5 (2019), 1667–1679. https://doi.org/10.1016/j.ipm.2019.04.004

[47] Tomasz D. Loboda and Peter Brusilovsky. 2010. User-adaptive explanatory program visualization: Evaluation and insights from eye movements. *User Modeling and User-Adapted Interaction* 20, 3 (2010), 191–226. https://doi.org/10.1007/s11257-010-9077-1

[48] Jianjie Lu and Kai-Yu Tong. 2019. Robust single accelerometer-based activity recognition using modified recurrence plot. *IEEE Sensors Journal* 19, 15 (2019), 6317–6324. https://doi.org/10.1109/JSEN.2019.2911204

[49] Sebastiaan Mathot, Edwin Dalmaijer, Jonathan Grainger, and S. Van der Stigchel. 2014. The pupillary light response reflects exogenous attention and inhibition of return. *Journal of Vision* 14, 14 (2014). https://doi.org/10.1167/14.14.7

[50] Sebastiaan Mathôt, Jasper Fabius, Elle Van Heusden, and Stefan Van der Stigchel. 2018. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods* 50, 1 (2018), 94–106. https://doi.org/10.3758/s13428-017-1007-2

[51] Robert B. Miller. 1968. Response time in man-computer conversational transactions. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I on - AFIPS '68 (Fall, part I)*. ACM, New York, 267–277. https://doi.org/10.1145/1476589.1476628

[52] Enrique Garcia Moreno-Esteva, Sonia White, Joanne Wood, and Alexander Black. 2017. Identifying key visual-cognitive processes in students' interpretation of graph representations using eye-tracking data and math/machine learning based data analysis. In *CERME 10*. Dublin, Ireland. https://hal.archives-ouvertes.fr/hal-01950548

[53] Enrique Garcia Moreno-Esteva, Sonia L. J. White, Joanne M. Wood, and Alex A. Black. 2018. Application of mathematical and machine learning techniques to analyse eye-tracking data enabling better understanding of children's visual-cognitive behaviours. *Frontline Learning Research* 6, 3 (2018), 72–84. https://doi.org/10.14786/flr.v6i3.365

[54] Youness Moukafih, Hakim Hafidi, and Mounir Ghogho. 2019. Aggressive Driving Detection Using Deep Learning-based Time Series Classification. In *2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE, Piscataway, NJ, 1–5. https://doi.org/10.1109/INISTA.2019.8778416

[55] Manuel Oliva and Andrey Anikin. 2018. Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Scientific Reports* 8 (2018), 4871. https://doi.org/10.1038/s41598-018-23265-x

[56] Flavio T.P. Oliveira, Anne Aula, and Daniel M. Russell. 2009. Discriminating the relevance of web search results with measures of pupil size. In *Proceedings of the SIGCHI Conference (CHI '09)*. ACM, New York, 2209–2212. https://doi.org/10.1145/1518701.1519038

[57] David L. Olson and Dursun Delen. 2008. *Advanced data mining techniques*. Springer, Berlin.

[58] Alvitta Ottley. 2020. Adaptive and personalized visualization. *Synthesis Lectures on Visualization* 7, 1 (2020), 1–117. https://doi.org/10.2200/S00973ED1V01Y201912VIS011

[59] Steven Pinker. 1990. A theory of graph comprehension. In *Artificial Intelligence and the Future of Testing*, R. Freedle (Ed.). Lawrence Erlbaum Associates, Hillsdale, NJ, 73–126.

[60] Stanislav Popelka, Zdeněk Stachoň, Čeněk Šašinka, and Jitka Doležalová. 2016. EyeTribe tracker data accuracy evaluation and its interconnection with hypothesis software for cartographic purposes. *Computational Intelligence and Neuroscience* 2016 (2016). https://doi.org/10.1155/2016/9172506

[61] B Pyakillya, N Kazachenko, and N Mikhailovsky. 2017. Deep Learning for ECG Classification. *Journal of Physics: Conference Series* 913 (2017), 012004. https://doi.org/10.1088/1742-6596/913/1/012004

[62] George E. Raptis, Christina Katsini, Marios Belk, Christos Fidas, George Samaras, and Nikolaos Avouris. 2017. Using Eye Gaze Data and Visual Activities to Infer Human Cognitive Styles: Method and Feasibility Studies. In *Proceedings of the Conference on User Modeling, Adaptation and Personalization* (Bratislava, Slovakia) *(UMAP '17)*. ACM, New York, 164–173. https://doi.org/10.1145/3079628.3079690

[63] Keith Rayner. 2009. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology* 62, 8 (2009), 1457–1506. https://doi.org/10.1080/17470210902816461

[64] Keith Rayner, Kathryn H. Chace, Timothy J. Slattery, and Jane Ashby. 2006. Eye Movements as Reflections of Comprehension Processes in Reading. *Scientific Studies of Reading* 10, 3 (2006), 241–255. https://doi.org/10.1207/s1532799xssr1003_3

[65] Joni Salminen, Ying-Hsang Liu, Sercan Şengün, João M. Santos, Soon-gyo Jung, and Bernard J. Jansen. 2020. The effect of numerical and textual information on visual engagement and perceptions of AI-driven persona interfaces. In *Proceedings of the International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. ACM, New York, 357–368. https://doi.org/10.1145/3377325.3377492

[66] Johannes Schwerdt, Michael Kotzyba, and Andreas Nurnberger. 2018. Inferring user's search activity using interaction logs and gaze data. In *2017 International Conference on Companion Technology (ICCT)*. 1–6. https://doi.org/10.1109/COMPANION.2017.8287075

[67] Julia Sheidin, Joel Lanir, Cristina Conati, Dereck Toker, and Tsvi Kuflik. 2020. The effect of user characteristics in time series visualizations. In *Proceedings of the International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. ACM, New York, 380–389. https://doi.org/10.1145/3377325.3377502

[68] Ben Steichen, Helen Ashman, and Vincent Wade. 2012. A comparative survey of personalised information retrieval and adaptive hypermedia techniques. *Information Processing & Management* 48, 4 (2012), 698–724. https://doi.org/10.1016/j.ipm.2011.12.004

[69] Ben Steichen, Giuseppe Carenini, and Cristina Conati. 2013. User-adaptive information visualization. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces - IUI '13*. ACM, New York, 317–328. https://doi.org/10.1145/2449396.2449439

[70] Ben Steichen, Cristina Conati, and Giuseppe Carenini. 2014. Inferring visualization task properties, user performance, and user cognitive abilities from eye gaze data. *ACM Transactions on Interactive Intelligent Systems* 4, 2 (2014), 1–29. https://doi.org/10.1145/2633043

[71] Christoph Strauch, Lukas Greiter, and Anke Huckauf. 2018. Pupil dilation but not microsaccade rate robustly reveals decision formation. *Scientific Reports* 8, 1 (2018), 13165. https://doi.org/10.1038/s41598-018-31551-x

[72] Johannes Titz, Agnes Scholz, and Peter Sedlmeier. 2018. Comparing eye trackers by correlating their eye-metric data. *Behavior Research Methods* 50, 5 (2018), 1853–1863. https://doi.org/10.3758/s13428-017-0954-y

[73] Dereck Toker, Cristina Conati, Giuseppe Carenini, and Mona Haraty. 2012. Towards adaptive information visualization: On the influence of user characteristics. In *User modeling, adaptation, and personalization*, Judith Masthoff, Bamshad Mobasher, Michel C. Desmarais, and Roger Nkambou (Eds.). Springer, Berlin, 274–285.

[74] Dereck Toker, Cristina Conati, Ben Steichen, and Giuseppe Carenini. 2013. Individual user characteristics and information visualization. In *Proceedings of the SIGCHI Conference (CHI '13)*. ACM, New York, 295–304. https://doi.org/10.1145/2470654.2470696

[75] Dereck Toker, Sébastien Lallé, and Cristina Conati. 2017. Pupillometry and head distance to the screen to predict skill acquisition during information visualization tasks. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces - IUI '17*. ACM, New York, 221–231. https://doi.org/10.1145/3025171.3025187

[76] Dereck Toker, Robert Moro, Jakub Simko, Maria Bielikova, and Cristina Conati. 2019. Impact of English reading comprehension abilities on processing magazine style narrative visualizations and implications for personalization. In *Proceedings of the ACM Conference on User Modeling, Adaptation and Personalization (UMAP '19)*. ACM, New York, 309–317. https://doi.org/10.1145/3320435.3320447

[77] M.C. Velez, D. Silver, and M. Tremaine. 2005. Understanding visualization through spatial ability differences. In *VIS 05. IEEE Visualization, 2005.* IEEE, 511–518. https://doi.org/10.1109/VISUAL.2005.1532836

[78] Z. Wang, W. Yan, and T. Oates. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International Joint Conference on Neural Networks (IJCNN)*. 1578–1585. https://doi.org/10.1109/IJCNN.2017.7966039

[79] Peter Wittek, Ying-Hsang Liu, Sándor Darányi, Tom Gedeon, and Ik Soo Lim. 2016. Risk and ambiguity in information seeking: Eye gaze patterns reveal contextual behavior in dealing with uncertainty. *Frontiers in Psychology* 7 (2016), 1790. https://doi.org/10.3389/fpsyg.2016.01790

[80] Yingying Wu, Yiqun Liu, Yen-Hsi Richard Tsai, and Shing-Tung Yau. 2019. Investigating the role of eye movements and physiological signals in search satisfaction prediction using geometric analysis. *Journal of the Association for Information Science and Technology* 70, 9 (2019), 981–999. https://doi.org/10.1002/asi.24240