International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-024-00447-5

ARTICLE

Gaze-Based Prediction of Students' Math Difficulties - A Time Dynamic Machine Learning Approach to Enable Early Individual Assistance

Kathrin Kennel[1] · Shoya Ishimaru[2] · Stefan Küchemann[3] · Steffen Steinert[1,3] · Jochen Kuhn[3] · Stefan Ruzika[1]

Accepted: 12 November 2024

Abstract
Graphical differentiation substantially promotes the understanding of basic concepts of calculus. At the same time, such tasks are challenging for many students. To be able to support students in complex tasks during the solution process, it is important to recognise if students are having difficulties and, if so, to identify precisely what the difficulty is. However, previous research was only able to identify correct and incorrect solution processes but not the specific difficulties of students. In our study, the eye tracking data of 143 students reveal how students' math tasks in context of graphically differentiation can be predicted. Retrospective-think-aloud (RTA) protocols were used to identify students' difficulties and three machine learning algorithms, k-nearest neighbours (KNN), Random Forest and support vector machine (SVM), provide an accurate prediction for this multiclass problem. The prediction results improve with increasing processing time and achieve already good values long before the solution process of the tasks is completed. Our results show that certain student difficulties can be detected very early during the task solution process. Although this approach has been demonstrated for a sub-area of calculus, it is transferable to other fields of the STEM domain and therefore has much wider scope.

Keywords Intelligent tutoring systems · Gaze-based adaptation · Learning difficulty detection

Extended author information available on the last page of the article

Published online: 22 January 2025

Introduction

Adaptive educational systems (AES) collect data about a learner, analyse it, and react, for instance, by providing the most suitable support for individual learning. The possibilities of adaptation include for example learning material sequencing, intelligent analysis of the solutions, interactive support for problem solving, and an adaptive presentation (Brusilovsky, 1998). The core of such a system is the diagnostic capability which could be increased by adding eye tracking data. Such data provide temporal and spatial information about how the learning content was visually processed. Based on the eye-mind hypothesis (Just & Carpenter, 1980), this can make perception and cognitive processes visible. Student difficulties might be identified so that a misconception-driven feedback is possible (Gusukuma et al., 2018). The great advantage of this sensor-based data collection is that support is already possible during the solution process. Nevertheless, there are very few approaches to integrate eye tracking in adaptive educational systems with the aim of recognizing content-related student difficulties.

The precondition for an adaptive system that is supposed to provide good feedback on the basis of gaze data is that it is possible to predict the response correctness on the basis of the eye tracking data. In this way, a decision can be made about which learner needs help. However, it is not only desirable to decide who needs assistance, but also to provide the best possible way of assistance. Therefore, the kind of student difficulty must also be classified. If this classification could already take place during the solution process, an early intervention would be possible. Marwan et al. (2020) examined the impact of adaptive immediate feedback (AIF) within programming environments and found evidence that AIF may improve student learning. Students reported in interviews that they found the system fun and helpful and that they felt more focused and engaged.

Related work and Theoretical Considerations

Graphical Differentiation

The derivative is one of the central concepts in calculus and is fundamental to many other fields, disciplines, and applications. For adequate understanding it is essential to switch between representations and to deal with visual representations in particular (Arcavi, 2003). Graphical differentiation usually requires formula-free derivation of the graph of the function to the corresponding graph of the derivative function. To do this, learners need to be aware of the connections between a function and its derivative function. The process can be assigned to the qualitative analysis, which according to Hussmann and Prediger (2010) makes a significant contribution to the development of conceptual understanding in calculus lessons. Graphical differentiation also involves the quantitative task of graphically determining the derivative at certain points. For a linear function, this task reduces to determining the slope, e.g., by looking at a slope triangle. For non-linear functions, a tangent must be applied to the graph beforehand. The various basic ideas of local rate of change, tangent slope,

local linearity, and amplification factor (Greefrath et al., 2016) help to build up a workable idea of the concept of derivative. The task considered in this paper takes up the basic idea of "tangent slope".

The process of graphical derivation demands a great deal of conceptual understanding from students (Klinger, 2017) and often leads to difficulties (e.g. Aspinwall et al., 1997, Ubuz, 2007, or Asiala et al., 1997). A well-known student difficulty in the context of linear graphs and functions that often occurs, is the so-called "slope-height-confusion" (e.g. Leinhardt et al., 1990 or Clement, 1985). For example, if the slope of a graph at a certain point is to be determined graphically, students sometimes just read off the y coordinate (the height of the graph at the point of interest). It is also often assumed that a graph that monotonically decreases in an interval also has a decreasing slope in the same interval (Ivanjek et al., 2016). If the abscissa and ordinate of the coordinate system are scaled differently, this can also lead to troubles. Students usually assume a uniform scaling and often orient themselves on the auxiliary lines in the coordinate system. Cho and Nagle (2017) refer to this student difficulty in their work as "block slope". Another common difficulty, reported by the authors, is that many students confuse rise over run and run over rise in the formula for slope. This mixing-up of the numerator and denominator when forming the slope triangle indicates a weak conceptual understanding. This student difficulty is referred to in the following as "inverted slope" and would be studied in the context of graphical differentiation.

Eye Tracking data as an Indicator of Cognitive Processes during Problem Solving

Gaze data contain temporal and spatial information about content to which gaze is directed. Quantitatively, these data can be used to determine perceptual measures such as total-visit-duration on learning material, time-to-first-fixation on specific areas of the content, or gaze transitions between related contents. In this context, the eye-mind hypothesis states that those contents are processed to which the gaze is directed. This suggests the interpretation that these perceptual measures can also be associated with cognitive processes. Gaze data are also associated with processes relevant to content processing during learning. In a review of 52 articles, Alemdag and Cagiltay (2018) highlighted that, for example, time to first fixation in a relevant area is associated with cognitive selection processes, total fixation time is associated with organization processes, and jumps between relevant areas are associated with cognitive content integration processes. At the same time, however, the authors found that the link between gaze data and cognitive processes is ambiguous. For example, a person might look longer at a graph because he or she struggles extracting the relevant information or because he or she is trying to relate his or her knowledge to the graph. Therefore, in order to clearly associate perceptual measures and cognitive processes, interviews (so-called think-alouds) need to be conducted with learners in addition. During retrospective think-alouds or stimulated recalls, immediately after gaze data recording, the subjects are played their gaze data again and asked about their cognitive processes. This method has the advantage that the problem-solving process is not affected by reporting their thoughts and it is directly linked to the eye-tracking data.

Apart from this, there are a number of factors that influence gaze data and thus, of course, the link between perceptual measures and cognitive processes. These include the content itself, metacognitive features such as response confidence, emotions and cognitive load (Alemdag & Cagiltay, 2018). To identify these influencing factors, unconscious processes can also be extracted and evaluated from the gaze data. For example, evidence for the three facets of cognitive load has also been identified in gaze data (Zu et al., 2020). Accordingly, increased pupil diameter suggests increased extrinsic load. However, the interpretation of pupil diameter is controversial and dependent on a number of factors.

Several studies found differences in the eye tracking data of experts and novices (Gegenfurtner et al., 2011). Fixation duration and the number of fixations on task-relevant areas are considered to be indicators of expertise. According to the information-reduction hypothesis (Haider & Frensch, 1999), experts should have less fixations of shorter duration on task-redundant areas and more fixations of longer duration on task-relevant areas. In a recent review that compares experts' and non-experts' visual processing of linear graphs during problem-solving and learning, Ruf et al. found that experts also perform more integrative eye-movements within a graph (Ruf et al., 2023). In another literature review on eye tracking during problem-solving and learning with linear graphs, the authors found in contrast that a high number of transitions between different answer options and between the answer options and the graph is related to a lower domain knowledge (Küchemann et al., 2022).

Moreover, it was shown that there are relationships between misconceptions and viewing patterns (Madsen et al., 2012). In their study, Madsen et al. found, that the attention of incorrect problem solvers was guided by misconceptions, as they spent more time fixating on areas, which were directly related to a misconception. Based on prior work in the context of linear graphs, we try to transfer these results to the context of graphical differentiation in this study.

Machine Learning Algorithms as a Predictor of Students' Difficulties

In addition to assessing problem-solving strategies, eye tracking is utilized for estimating learners' knowledge-levels and cognitive/affective states with machine learning. For instance, English skill level of non-native English speakers and sentences that are subjectively difficult for readers can be estimated by meaning eye movements on texts (Augereau et al., 2016; Okoso et al., 2015). In these studies, durations of each fixation and the number of regressions (backward-saccades) were reported as representative features. Yamada et al. (2017) investigated solving behaviors on multiple-choice questions and estimated self-confidence on the decision. Jacob et al. (2018) classified the level of interest on reading materials into four levels by using an eye tracker. Bixler and D'Mello (2016) proposed a gaze-based automatic mind-wandering (task-related or task-unrelated thoughts) detection algorithm. Recent deep-learning technologies has enabled estimating gaze points from appearances of eye images, which is not enough precise for saccade-based studies but applicable for AOI-based analysis (Zhang 2019).

In a previous work on students' gaze data during problem-solving with graphs, Küchemann et al. (2021) compared the prediction probability of different eye-track-

ing metrics, namely the total visit duration and the transitions between AOIs, to classify the visual problem-solving strategy. They found that a combination of the total visit duration on AOIs and the number of transitions between AOIs reaches the highest prediction probability in most of the cases. Furthermore, to optimize the machine-learning algorithm, which is used to classify the visual problem-solving strategy, Küchemann et al. evaluated three different machine learning algorithms and found that a support vector machine (SVM) performs best in comparison to a random forest and multilayer perceptron, if the selected features discriminate well between correct and incorrect answers (Küchemann, 2020). This finding was confirmed by Dzsotjan et al. who observed that an SVM predicts the learning gain in an embodied learning environment with graphs based on eye-tracking data best (Dzsotjan et al., 2021). Moreover, Becker et al. found that eye tracking data during the solution of a graph task can even be used to predict students' performance during a subsequent graph task (Becker et al., 2022).

The use of eye Tracking in Adaptive Learning Systems

The following section provides an overview of approaches in which eye tracking has been integrated into adaptive learning systems. One of the first projects in which eye tracking was integrated into an e-learning system is AdeLE (Adaptive E-Learning through Eye Tracking). Barrios et al. (2004) report that the intention in that project was to observe users' learning behaviour in real time by monitoring characteristics such as objects and areas of focus, time spent on objects, frequency of visits, and sequences in which content is consumed. The goal of the approach was to be able to detect patterns which indicate disorientation or other suboptimal learning strategies. Conati and Merten (2007) investigated the applicability of eye tracking data to provide information on user meta-cognition. Text 2.0 is a framework with the aim of supporting a reader. On the basis of gaze data, online decisions are made as to whether, for example, further information should be displayed (Biedert et al. 2010). Gaze Tutor is an intelligent tutoring system developed by D'Mello et al. (2012), in which eye tracking data is used to identify and react to students' boredom and disengagement. D'Mello et al. (2017) also developed a technology which is able to detect mind wandering during computerized reading and to intervene by posing just-in-time questions and encouraging re-reading as needed. Following research aimed at supporting the design of novel user-adaptive visualization systems. Eye tracking has been leveraged to predict user's confusion during interaction with ValueCharts (Lallé et al., 2016), as well as to predict user's cognitive abilities during bar and radar graph processing (Steichen et al., 2013 and Gingerich et al., 2015) or while reading magazine-style narrative visualizations (Barral et al. 2020). Lallé et al. (2019) investigated the potential of gaze-driven adaptive interventions to support the processing of such textual documents with embedded visualizations and Barral et al. (2021) further investigated the effects on comprehension, perceived usefulness and distraction. Schmidt et al. (2014) developed a web-based eLearning environment called ALM (Adaptive Learning Module) which adapts the learning content based on a real time analysis of eye tracking data. Building on this, Scheiter et al. (2019) developed an adaptive system that provides personalised support based on gaze data. To

investigate the integration of text and pictures, fixation times as well as number of transitions between both representations were captured. After adapting the thresholds for adaptive responses, the authors found that the system tended to support students with stronger cognitive characteristics but hindered students with weaker ones. Taub and Azevedo (2019) recorded eye movements during learning with Meta Tutor, an intelligent tutoring system that teaches students about the human circulatory system. The authors used eye tracking and log-file data to investigate the impact of prior knowledge on fixations on learning-related AOIs and on sequences of cognitive and metacognitive self-regulated learning processes. Using data from the same system, Jaques et al. (2014) investigated how to predict learning-relevant emotions from eye tracking data. Li et al. (2021) investigated students' cognitive engagement while diagnosing virtual patients with BioWorld, an intelligent tutoring system designed to help medical students practice clinical reasoning skills. They trained supervised machine learning algorithms to predict on the basis of facial behaviours, e.g. eye gaze, whether students were cognitively engaged in problem-solving. In a recent study Emerson et al. (2023) used datasets collected using the Crystal Island game-based learning environment for microbiology education. Available multimodal data channels from the datasets were leveraged to simultaneously predict student post-test performance and interest.

Many of the systems described above use gaze data to diagnose learner motivation or attention, for example. Only a few systems make diagnoses about the presence of difficulties during the solution process and, to the best of our knowledge, there is no system that makes diagnosis about the problem-solving strategy and that is able to predict specific content-related student difficulties on the basis of gaze data.

Research Questions

In this work we applied three machine learning algorithms to investigate whether students' math difficulties, which can occur during solving graphical differentiation tasks, can be identified with eye tracking data. Compared to previous work, the attempt is not only to recognise that someone needs help, but to identify exactly what help is needed. Furthermore, the online decision-making situation is considered to see whether it is also possible to support the learner already during the solution process. We investigated the following research questions:

- Can student difficulties be predicted based on eye tracking data recorded while solving a task?
- Can student difficulties already be predicted during the problem-solving process by analysing eye tracking data in real-time?

Method

Participants and Study Design

The study, which consisted of an eye tracking part and a retrospective-think-aloud (RTA) part, was conducted at five senior high schools in Rhineland-Palatinate in Germany. The sample consisted of 148 upper-secondary students of ages 16 to 19 (68 female, 78 male, 2 non-binary) with basic and advanced mathematics courses. The participation was voluntary, not extrinsically motivated and took place either during a free period or during a lesson (with the teacher's permission). The data collection lasted half an hour and took place in a room of the school that was not used for any other purpose during the survey period. The students took part in the survey one after the other. The study was carried out by a team of three people, with only one person interacting with the participant at a time. In order to guarantee consistency, the researchers followed previously defined guidelines.

The participants received a short briefing and provided anonymous information about their age, gender, mathematics grade, motivation etc. on a sheet. Afterwards they sat down in front of a computer screen and a 9-point calibration was carried out. They solved 9 items in total, which thematically all dealt with graphical differentiation. This topic was previously covered in class. The items (see top-left of Fig. 1 for an example) were presented to the participants on a 22-inch monitor. The resolution of the computer screen was 1920×1080 pixels with a refresh rate of 75 Hz. The eye movements were recorded with a stationary eye tracking system (Tobii Pro X3-120, binocular) with a frequency of 120 Hz and an accuracy of 0.2 to 0.9 degrees on average (Tobii, 2015). The average distance between screen and participant was 65 cm. The students were instructed to solve the task first and to write down the answer only after pressing the space bar. By pressing the space bar, the eye tracking recording for the corresponding item was stopped. The solution was subsequently written down after the recording so that the students usually did not look away from the screen. During the data collection the students did not receive feedback and it was not possible to return to a previous task. After each solution was given, the response confidence was inquired by asking participants to select one of four options: "very confident", "confident", "uncertain", or "guessed". Participants were instructed to tick "uncertain" if they were pursuing a weak idea and to tick "guessed" only if they were answering purely at random.

After solving the 9 items, participants were asked on a voluntary basis to explain their solution strategy. The participants were given brief instructions in advance to explain the procedure. A video of their own gaze data while solving the task was played. While watching this video, the participants were asked to describe their approach to solving the problem by referring to their recorded gaze data. In order to enable a better description of thought processes, the participants could pause the video at any time by pressing the space bar. The instructor only intervened in the procedure to remind the participants to think aloud if necessary or to ask a question if the explanation was incomprehensible. Otherwise, only affirmative signals such as "aha", "I see" or "ok" were used. The explanations were recorded with the help of

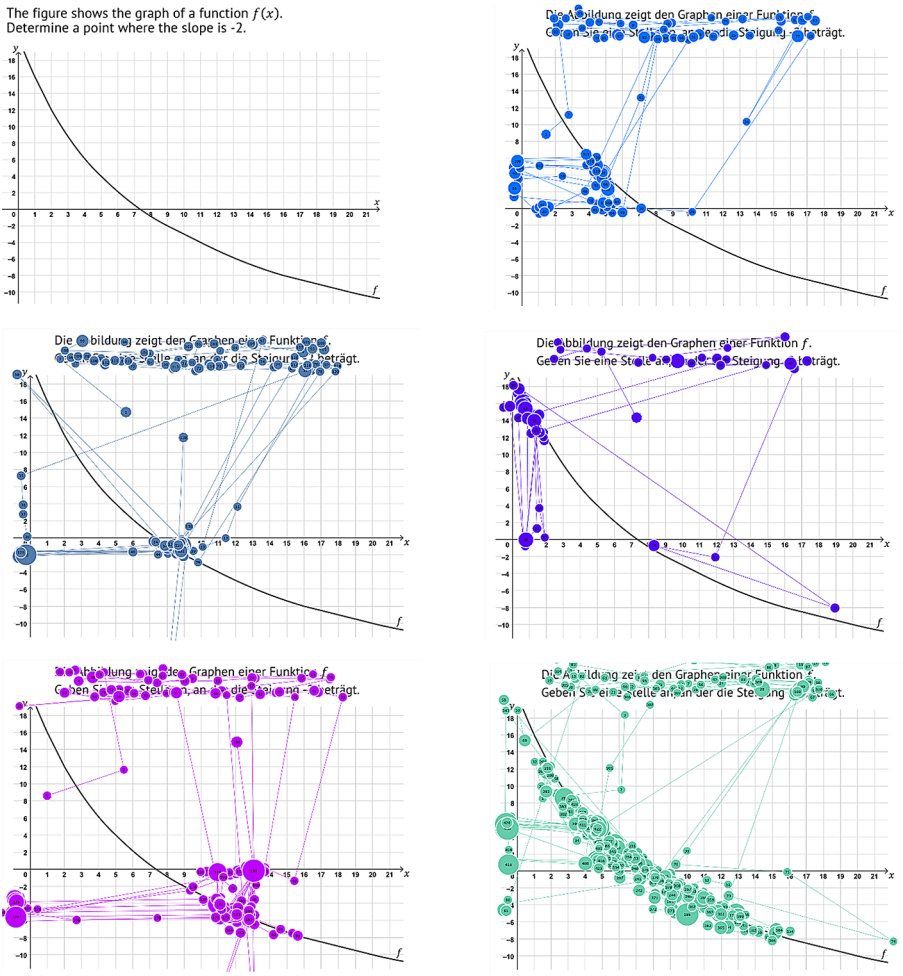

Fig. 1 Item with task text translated into English (top left) and five gaze plots of participants, who were assigned to the following groups on the basis of their RTA interviews: correct (top right), slope-height error (mid left), block-slope error (mid right), inverted-slope error (bottom left), guess (bottom right)

an external microphone. This collection of RTA protocols later serves to assign the participants to one of the coding categories (van Someren et al., 1994).

In this work, we analyse one of the items in depth. It addresses the concept of the derivative (see Fig. 1). The task is reversed in comparison to the standard task "Determine the slope of the function". Here, the slope is given and one should determine a point at which the function attains this slope.

Data Analysis

The data collection and the definition of AOIs was carried out with the software Tobii Pro Studio. Due to poor recording quality (given by a low percentage of weighted

gaze samples) or to technical problems during the eye tracking part, five participants were subsequently excluded from the analysis. Thus, a sample size of $n = 143$ is available. To identify fixations and saccades, an I-VT algorithm was applied with thresholds of 8500°/s2 for acceleration and 30°/s for velocity (Salvucci & Goldberg, 2000).

The data analysis was carried out with the programming language "R". Responses where participants indicated that they had guessed were marked as incorrect. Participants were binarily divided into the groups correct and incorrect. In order to allow a more precise prediction, participants with incorrect answer were additionally divided into different student difficulty classes. For this purpose, a coding guide was established and two coders independently assigned one of the following five labels to each participant on the basis of the RTA protocol: correct (C), slope-height error (SH), block-slope error (BS), inverted-slope error (INV), guess (GUE). Table 1 shows some examples of student explanations and assigned coding. If more than one student difficulty occurred at the same time, the student difficulty that was considered by the coder to be more fatal or fundamental for the wrong procedure was coded. Participants who followed a strategy that could not be categorized to any of the groups and that occurred only in individual cases were categorized to the coding category guess. Subsequent to the coding, the interrater reliability (Cohens kappa) was determined on the whole dataset and could be classified to be on a very high level of agreement ($\kappa = 0{,}97$).

Figure 1 shows typical gaze plots from each of the five coding categories. The post-hoc categorization of the remaining participants (without RTA interview) was based on the written solutions. This procedure was validated as follows: For the 100 participants who had already been assigned to coding categories by means of the interviews, an assignment was also made on the basis of the written solutions. These two allocation procedures were compared with each other and showed a high level of agreement with $\kappa = 0{,}85$ for Cohens kappa, which is considered almost perfect according to Landis and Koch (1977). Although the written answers show a high agreement with the RTAs, the coding is much more reliable if it is based on the RTAs, as the written solution could also be the result of totally different thought processes of the learner, rather than being related to the presumed difficulty. The post-hoc categorization is therefore based on the interviews whenever possible.

For the statistical data analysis, non-parametric tests were used, as the eye tracking measures were not normally distributed. The Mann-Whitney U test was utilized to investigate the differences in the eye tracking measures between the response accuracy levels (correct vs. incorrect). Similarly, the Kruskal-Wallis test was used to examine differences between the more finely divided five groups. The Dunn test for pairwise comparison was used as a post-hoc test (p-values adjusted by Benjamini-Hochberg). Each effect was considered statistically significant when the p-value was below the 5% threshold ($p < 0.05$). To judge the magnitude of the phenomenon, we also report the effect size measures.

Position of AOIs and used eye Tracking Metrics for Prediction

The following eye tracking metrics are used as features for the prediction: For each of the categories C, SH, BS and INV, we consider the number of fixations (FC) in a

Table 1 Examples of student explanations (translated into English)

Coding	Student Explanation	Comment
C	"Slope −2 would be - so I thought - one to the right and two down. Then I went along the graph and looked to see where that occurred. (…) I wrote it down at 5 - ah, exactly one box to the right and one down is what I noticed, because it's not on the y-axis in a one step."	typical explanation for coding category C
SH	"(…) I was sure that there had to be a negative gradient under the x-axis. And then all I had to do was read it off and I came up with 9 relatively quickly."	typical explanation for coding category SH
BS	"(…) I first briefly thought about what the gradient is, i.e. how it works with the −2, and then I thought that if it goes one over and two down, then it is a gradient of -2. And then I walked along the graph and found a point directly up there between x 1 and x 2, where it goes one over and 2 down, and then I took the point."	typical explanation for coding category BS
INV	"(…) Uh, then I thought to myself, okay, I remember from physics and maths that you always had to count the boxes for the gradient and at 13 it was easy to see exactly because, as I said, it was right at the corner. And then I went down to 13 and looked, okay, if I go two to the left and one box up, then I would have the gradient −2 and I wrote it down like that."	typical explanation for coding category INV
GUE	"(…) I looked at it for a long time to see if there was a slope of -2 somewhere. (…) Um, I chose a point and then counted the boxes upwards. Or, in this case, downwards. (…) For example, I counted here at eleven - and there I went down two boxes - and there I was at this point. (…)"	typical explanation for coding category guess Participant is searching for a solution that represents the number −2 in some way.
GUE	"(…) So the graph falls, i.e. the, uh, the derivative would, uh, be in the negative range at the beginning and would then approach the x-axis, because the, uh, because the graph falls more at the beginning and then no longer, so it flattens out. And this flattening movement would then be the approximation to the x-axis again in the derivation and I imagined that as an arc and was then approximately at six."	Participant was initially categorized to the coding category "estimation". Since this strategy only occurred very rarely (2x), these participants were categorized post-hoc to the coding category guess. The written answer was correct.

corresponding area of interest (AOI) group. Figure 2 (left) shows the position of these AOI groups, which were selected mainly for reasons of plausibility. In each case, the area on the graph was grouped together with the corresponding sections on the x- and y-axis. The blue AOI group displays the relevant areas. These areas were determined by an expert rating, in which 8 mathematicians marked the areas that they considered important for solving the task. It was previously shown that the students who choose the correct answer have a significantly higher number of fixations in that relevant area (Kennel et al., 2022), so that this measure will most likely be a good predictor. The heat map on the left in Fig. 3 visualises the fixations of students who solved the task correctly. Here, it is also apparent that the attention of those participants is concentrated on the blue AOI group. Similarly, the heat map in the right of Fig. 3

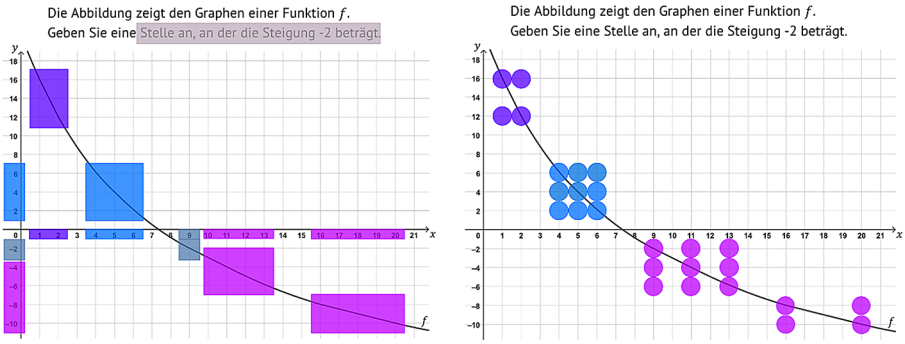

Fig. 2 AOIs for relevant area (blue) and areas corresponding to the block-slope error (dark purple), to the slope-height error (grey) and to the inverted-slope error (purple) [left]; AOIs for transition analysis [right]. English translation of the task: "The figure shows the graph of a function f. Determine a point where the slope is -2"

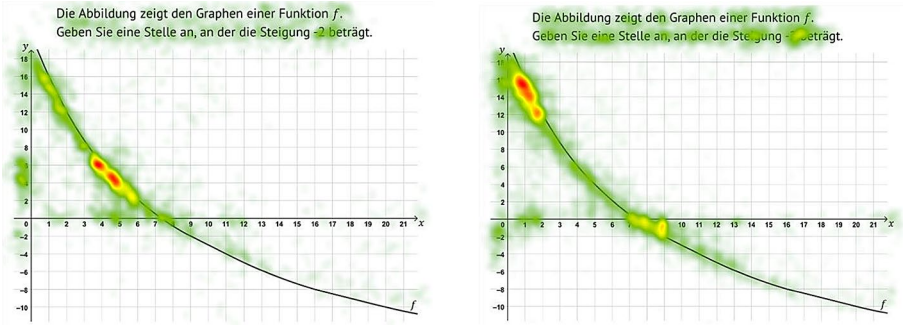

Fig. 3 Heat maps of fixation counts of students who solved the item correctly (left) and incorrectly (right). The gaze data of these groups clearly differ. Students in the second group have a higher number of fixations in areas that can be attributed to the block-slope error and the slope-height error. English translation of the task: "The figure shows the graph of a function f. Determine a point where the slope is -2"

visualises the fixations of all students who did not solve the task correctly. Here, attention is concentrated on two areas attributed to the slope-height error as well as to the block-slope error. In Fig. 2, the corresponding areas are coloured in grey and dark purple, respectively. For the inverted-slope error, the AOI is set accordingly at the bottom right. In addition, a second graph area is added to the group, which is also related to this error, but in combination with the disregard of the scaling of the y-axis.

Due to the different processing and reading times, relative measures were also considered. Thus, for all of the four AOI groups described above, the number of fixations in the respective area in relation to all fixations on the graph area was included. For this purpose, another AOI which covered the whole graph area was set.

To solve the problem, two steps must be undertaken: first, a tangent line must be virtually applied to the graph and, second, the slope of this line must be determined. In order to transfer the first solution step into a feature, the absolute saccadic direction was considered. More precisely, we count the number of saccades in the direction

of the tangent line (+/- 5°) for each participant. In order to translate the second step into an eye tracking metric, transitions are considered that indicate the observation of a slope triangle. AOIs were set on possible corner points of gradient triangles (Fig. 2). All transitions between the circles within a group (of 4 or 9 circles of the same colour) are counted. Additionally, transitions between the AOIs belonging to the slope-height error and the text are counted. Furthermore, fixations on the y-axis and the mean length of all saccades are evaluated.

The features were calculated for different time spans of increasing length: For the first 10 s, for the first 15 s, etc. up to the first 80 s. The data were thus truncated in each case so that we can examine how the machine learning algorithm would perform in the solution process with a real-time evaluation of the data. This post-hoc analysis allows to compare the prediction results over time.

Classification by Machine Learning Algorithms

For the classification of student difficulties, we applied three machine learning algorithms, which represent a diverse set of approaches to classification: k-nearest neighbors (KNN) is a simple instance-based learning method that relies on similarity measures, Random Forest is an ensemble learning method that combines multiple decision trees, and support vector machine (SVM) is a powerful algorithm for separating data points using hyperplanes in high-dimensional space. The utilization of these algorithms, which have been widely used and have demonstrated strong performance in many classification tasks, provides a comprehensive analysis. The data were first standardised to bring all features to a comparable scale. Giving each feature a mean value of zero and a variance of one prevents the objective function from being dominated by a feature with a very high variance. In order to predict student difficulties, we carried out a multiclass classification, classifying participants into one of five classes (C, SH, BS, INV, GUE), as well as five binary classifications, classifying the participants into one of two classes (into one of the coding categories or the respective rest). The multiple binary classifications were additionally conducted to investigate the possibility of individual intervention times for different difficulties. We examined how well an estimator performs over time by training and testing on different data sections.

Besides the widely used measure Accuracy (ACC), we also report the Matthew Correlation Coefficient (MCC) to compare the classifiers performance as well as the performance over time, as Accuracy does not cope well with unbalanced data (Jurman et al., 2012 / Chicco & Jurman, 2020). The definition of the MCC in the multiclass case was originally reported from Gorodkin (2004). It takes values between −1 and +1, whereby a coefficient of +1 represents a perfect prediction, 0 an average random prediction and −1 an inverse prediction. For the sake of completeness, precision, recall and f1-score are also reported in the appendix.

We used a repeated nested cross validation. The dataset was split into five outer folds using person independent stratified cross validation. Each training set was split into 5 folds again. Here, the parameters of each estimator were optimized by cross-validated grid-search over a parameter grid. The MCC is used to evaluate the performance of the cross-validated model on the test set. For the KNN algorithm, the

number of neighbours and the type of distance metric were optimised, for the random forest algorithm the number of trees in the forest was optimised and for the SVM the kernel function, the parameter C and the parameter gamma (except in the case of a linear kernel) were optimised. The best parameters found within each inner cross validation were then used to validate the model on the outer test set, 5 times, each time using a different fold as a test data set and using the remaining folds for training. This nested cross validation was repeated 10 times, each time with randomly chosen splits, resulting in a total of 50 iterations.

The described procedure was performed in the programming language python for every time span (first 10 s, first 15 s,…, first 80 s), as well as for the whole data. Due to the unbalanced classes we also tested the application of two different oversampling methods, which resulted in a new data size of 305. Both the random oversampling method, as well as the synthetic minority oversampling technique (SMOTE), resulted in significantly better values for all metrics in the multiclass case (see Tables 3 and 4 of Appendix A), so that overfitting probably occurred here. We therefore decided to stick to the original data without applying rebalancing of the data points and to apply the MCC in terms of a reliable evaluation.

Results

Note that whenever we talk about groups in the following chapters, we are referring to the post-hoc categorization described in Chap. 3.2.

Statistical Tests

The task was solved correctly by 34 of the students. Accordingly, the group of those who gave an incorrect answer or stated that they had guessed is clearly larger, with 109 participants. Table 2 shows the extent to which these two groups differ with regard to the chosen eye tracking metrics. The order of the eye tracking metrics is given by the ascending order of p-values of a Mann-Whitney-U-test. The first five metrics lead to significant differences between the two groups with medium and strong effects, respectively. In Fig. 4, a 3D scatterplot is shown with the three best features for the discrimination of participants with correct and wrong strategy.

Table 3 shows how often the different student difficulties occurred and how long it took the participants with a certain difficulty to complete the task on average. The participants who did not solve the task correctly, did most likely just use the boxes to form a gradient triangle (61 test persons). The slope-height error occurred 25 times and the inverted-slope error 13 times. 10 participants guessed. The Kruskal-Wallis test reveals that these five groups differ with respect to the 15 metrics. The test yields significant values for all 15 metrics.

The post hoc tests (Dunn-Bonferroni) demonstrate which groups differ significantly through pairwise comparisons. The Tables 4, 5, 6 and 7 show all pairwise comparisons with the group of students who answered correct (Table 3), with the slope-height error group (Table 5), with the block-slope error group (Table 6) and with the inverted-slope error group (Table 7). The features represented in each table

Table 2 The 15 features and their statistical comparison (p-value and effect size r) between students who gave the incorrect and correct answer and between the five groups C, SH, BS, INV and GUE

Feature	Mann Whitney U test		Kruskal Wallis test	
	p-value	*r*	*p*-value	Chi^2
relative fixation count in relevant areas [relFC_RA]	<0.001	0.71	<0.001	82.83
fixation count in relevant areas [FC_RA]	<0.001	0.68	<0.001	78.39
transitions between circles in relevant area [Transitions_RA]	<0.001	0.66	<0.001	72.70
saccades in direction of tangent line [Sacc_Tangent]	<0.001	0.45	<0.001	53.95
fixation count in the AOI of the ordinate [FC_O]	<0.001	0.31	<0.001	21.12
relative fixation count in areas corresponding to the slope-height error [relFC_SH]	0.002	0.25	<0.001	79.63
relative fixation count in areas corresponding to the inverted-slope error [relFC_INV]	0.007	0.23	<0.001	37.17
relative fixation count in areas corresponding to the block-slope error [relFC_BS]	0.016	0.20	<0.001	105.82
mean saccade length [Sacc_Length]	0.016	0.20	<0.001	33.69
fixation count in areas corresponding to the slope-height error [FC_SH]	0.019	0.20	<0.001	67.69
transitions between text and areas corresponding to the slope-height error [Transitions_SH]	0.075	0.15	<0.001	71.68
fixation count in areas corresponding to the inverted-slope error [FC_INV]	0.126	0.13	<0.001	24.77
transitions between circles in area corresponding to the block-slope error [Transitions_BS]	0.210	0.10	<0.001	59.27
fixation count in areas corresponding to the block-slope error [FC_BS]	0.456	0.06	<0.001	89.04
transitions between circles in area corresponding to the inverted-slope error [Transitions_INV]	0.902	0.01	<0.001	50.48

are those which separate the respective group well from the others. It becomes clear that the metrics are well suited for separating the individual groups (see small p-values). All 15 features are listed across the 4 tables. Each table contains the features that were primarily intended to recognize this group. For example, in Table 7, the features belonging primarily to the INV category, number of fixations on the INV area and transitions between possible vertices of the slope triangle in the INV area, are considered and the results of the pairwise comparisons between participants in the INV

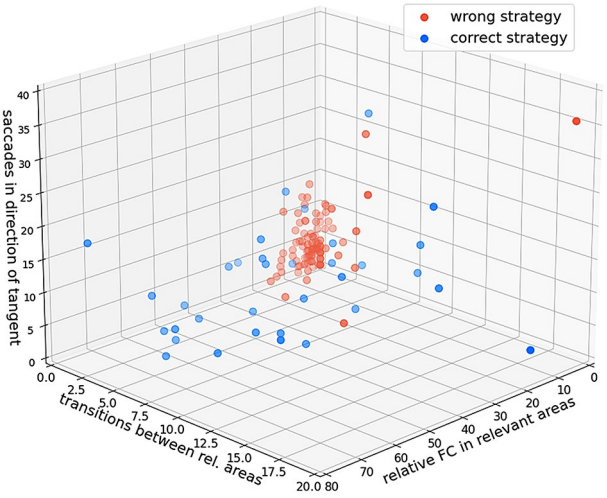

Fig. 4 3D Scatterplot of all participants (grouped on the basis of RTA-protocols in participants with correct and wrong strategy) with corresponding values of the three metrics: relative FC in relevant areas, transitions between relevant areas and saccades in the direction of the tangent line

Table 3 p-values and effect sizes of the Dunn-Bonferroni test; pairwise comparison with group C

Feature	C - SH		C - BS		C -INV		C – GUE	
	p	*r*	*p*	*r*	*p*	*r*	*p*	*r*
relFC_RA	<0.001	0.56	<0.001	0.69	<0.001	0.47	0.077	0.16
FC_RA	<0.001	0.59	<0.001	0.63	<0.001	0.45	0.181	0.13
Transitions_RA	<0.001	0.62	<0.001	0.60	<0.001	0.39	0.057	0.18
Sacc_Tangent	<0.001	0.59	<0.001	0.34	0.002	0.28	0.526	0.05
FC_O	0.218	0.12	0.001	0.31	0.001	0.32	0.241	0.12

See Table 4 for explanation of abbreviations

Table 4 Group frequencies and mean processing times in seconds

Group	C	SH	BS	INV	GUE
Frequency	34	25	61	13	10
Mean Processing Time	55	28	39	36	64

Note. C = correct; SH = slope-height error; BS = block-slope error; INV = inverted-slope error; GUE = guess

Table 5 p-values and effect sizes of the Dunn-Bonferroni test; pairwise comparison with group SH

Feature	SH – C		SH – BS		SH – INV		SH - GUE	
	p	*r*	*p*	*r*	*p*	*r*	*p*	*r*
relFC_SH	<0.001	0.59	<0.001	0.67	0.095	0.16	0.001	0.29
FC_SH	<0.001	0.51	<0.001	0.62	0.177	0.12	0.026	0.20
Transitions_SH	<0.001	0.53	<0.001	0.68	0.013	0.23	0.001	0.30
Sacc_Length	<0.001	0.43	<0.001	0.39	<0.001	0.37	0.015	0.23

Table 6 p-values and effect sizes of the Dunn-Bonferroni test; pairwise comparison with group BS

Feature	BS – C		BS – SH		BS – INV		BS - GUE	
	p	*r*	*p*	*r*	*p*	*r*	*p*	*r*
relFC_BS	<0.001	0.51	<0.001	0.73	<0.001	0.53	<0.001	0.34
FC_BS	<0.001	0.36	<0.001	0.70	<0.001	0.50	0.004	0.26
Transitions_BS	<0.001	0.35	<0.001	0.53	<0.001	0.41	<0.001	0.33

Table 7 p-values and effect sizes of the Dunn-Bonferroni test; pairwise comparison with group INV

Feature	INV – C		INV – SH		INV – BS		INV - GUE	
	p	*r*	*p*	*r*	*p*	*r*	*p*	*r*
relFC_INV	<0.001	0.49	<0.001	0.34	<0.001	0.46	0.002	0.28
FC_INV	<0.001	0.38	<0.001	0.34	<0.001	0.38	0.038	0.20
Transitions_INV	<0.001	0.47	<0.001	0.48	<0.001	0.59	<0.001	0.33

category with all other groups are presented by means of the p-value and the effect size. The pairwise comparisons in the Tables 3, 5, 6 and 7 reveal significant differences with strong and medium effects in the majority of cases. Only the comparisons between the group that answered correctly and the group that guessed in Table 3 and two values in Table 5 (column SH–INV) do not lead to significant differences. The first case can be explained by the fact that some of those who guessed gave the correct answer even though the explanation was wrong or missing (see also last student explanation in Table 1). An explanation for the second case could be that the areas for the two student difficulties slope-height and inverted-slope are not perfectly separable and have a small overlap at $x=9$.

Figure 5 visualises all participants with respect to the relative fixation count in relevant areas and areas attributed to the slope-height, as well to the block-slope error. It becomes apparent that the groups correct, slope-height error and block-slope error can be recognized quite well with the help of these 3 features alone.

If the prediction is done at the end of processing, the processing time could also be included in the features. Both tests yield significant differences between the groups (Mann-Whitney U-test: $p<0.001$, $r=0.29$; Kruskal-Wallis test: $p<0.001$, $Chi^2=22.13$). However, if the prediction should be carried out already during the solution process, this feature cannot be included, as the processing time is only known at the end.

Classification

The algorithms KNN, Random Forest and SVM were each applied to 16 different data sets: 1 time on the entire data and 15 times on temporally truncated data. Figure 6 illustrates the evaluation of the prediction quality of the multiclass classification using the MCC (for the evaluation by means of ACC, Precision, Recall and F1-Score see Table A1 and A2 in Appendix A). The evaluation of the prediction quality each occurred after all 50 runs (5-fold cross validation was repeated 10 times) of the models. As this is a multiclass classification with 5 classes, the random probability that a class will be predicted correctly is 20%. It is apparent that the prediction quality improves with increasing information, i.e. a larger data section. The MCC already

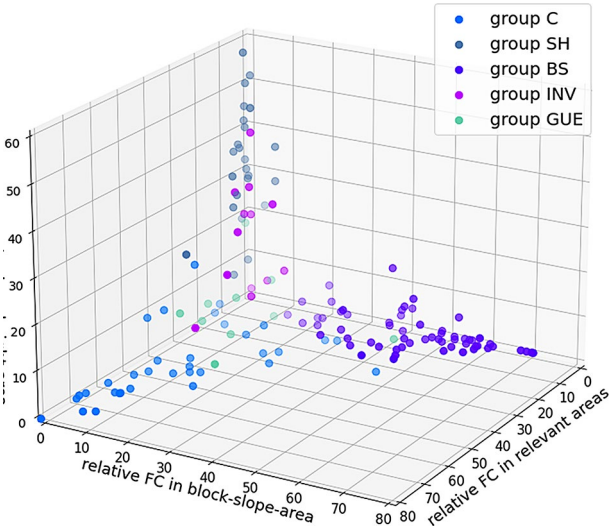

Fig. 5 3D Scatterplot of all participants with respect to the relative FC in relevant areas and in areas corresponding to the slope-height and the block-slope error

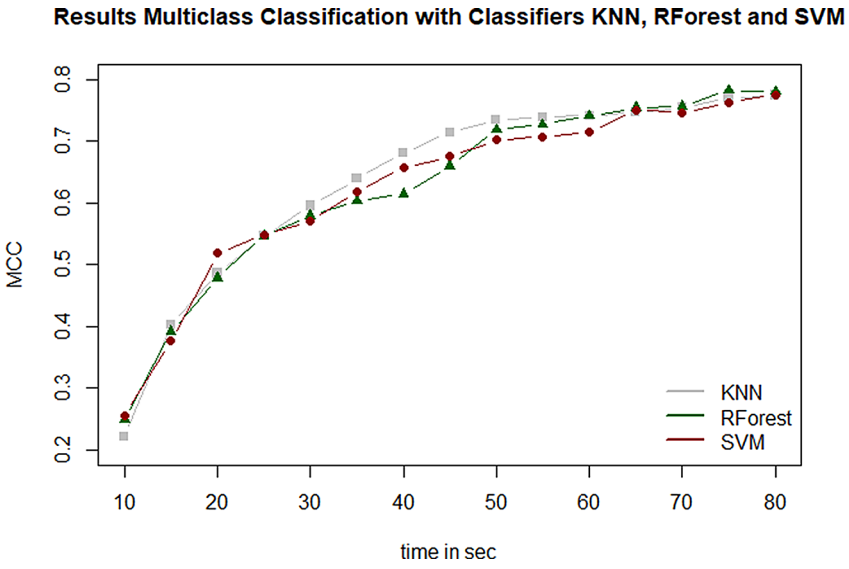

Fig. 6 MCC of the Multiclass Classification with Classifiers KNN, Random Forest and SVM, evaluated after 10, 15,…, 80 s

exceeds a value of 0.6 after 35 s for all 3 classifiers and 0.7 is exceeded after 50 s. Since the MCC takes values between −1 and +1, these scores already indicate a high agreement between prediction and observation.

For the classifications with data excerpts of the first 30 and 50 s, as well as for the full data set, the confusion matrices for each model are shown in the appendix

(Figs. 1, 2 and 3 in Appendix A). The matrices indicate that all of the three models are similar good in learning the individual classes. The KNN algorithm delivers higher scores than the other two algorithms between 30 and 60 s. By the chosen method, the algorithms were executed 50 times for each data section. In total, parameters were determined 800 times during the hyperparameter tuning in the inner cross validation to get the results. The frequency of hyperparameter combinations are reported in Appendix A. The most frequent parameter combinations are as follows: For the KNN, nine closest neighbours were considered most frequently, in relation to the Manhattan distance (see also Table A5). In the Random Forest, 100 trees were chosen most often (see also Table A6). The SVM algorithm was executed with RBF kernel in most cases. The RBF kernel was applied in 511 cases, whereas a linear kernel was selected only 289 times. Using the linear kernel, C = 0.1 was chosen in 83% of cases. With the RBF kernel, 1 was set most frequently for the parameter C (in combination with gamma = 0.1 or gamma = 0.01). The most frequent combination was C = 10 and gamma = 0.001 (see also Table A7).

Overall, all 3 classifiers achieve values of the MCC around 0.78, whereby the values after 50 s are already very good with values above 0.7 (for the KNN algorithm even a little earlier).

In order to evaluate how well a certain student difficulty can be found by a machine learning algorithm, 5 binary classifications were carried out in addition to the multiclass classification. The results of all three algorithms are illustrated in Fig. 7. The first five plots show the scores of the MCC for the evaluation of the binary classifications respectively between one of the classes and the others. In the last plot all scores of the KNN are presented together. The MCC is reported because it provides more meaningful values than the ACC, especially for the comparisons INV-rest and GUE-rest, due to the unbalanced data (see also Table B1 of Appendix B).

It is evident that all three algorithms provide good predictions for the splits SH-rest and BS-rest. For the first split, the MCC already exceeds a value of 0.7 after 15 s and after 30 s it is already above 0.8. For the second split, a value of 0.7 is reached after 35 s, 0.8 is achieved after 45 s and after 50 s the values remain constantly high between 0.85 and 0.9. The prediction quality of the binary classification with splitting C- rest is similar to the quality of the multiclass classification. After 35 s, the value of the MCC is just under 0.6 and after 50 s it is almost 0.7. While the values of the three algorithms barely differ in the described cases, they differ widely for the classification with INV-rest splitting. The KNN predicts best and exceeds a value of 0.6 after 50 s, while the Random Forest remains distinctly below a value of 0.5 even with the data section of 80 s. The fifth plot shows that the group of those who guessed cannot be recognised well. The MCC fluctuates around the value 0, indicating an average random prediction.

The application of the three machine learning methods results in similar values demonstrating the robustness and reliability of the results. The KNN provides the best prediction in the multiclass case as well as for the binary classifications (especially for the separation INV-rest and SH-rest). For the KNN, between 2 and 5 neighbours were chosen most frequently. Here, the Euclidean and the Manhattan distance were chosen approximately equally often (see also Table B2 in Appendix B).

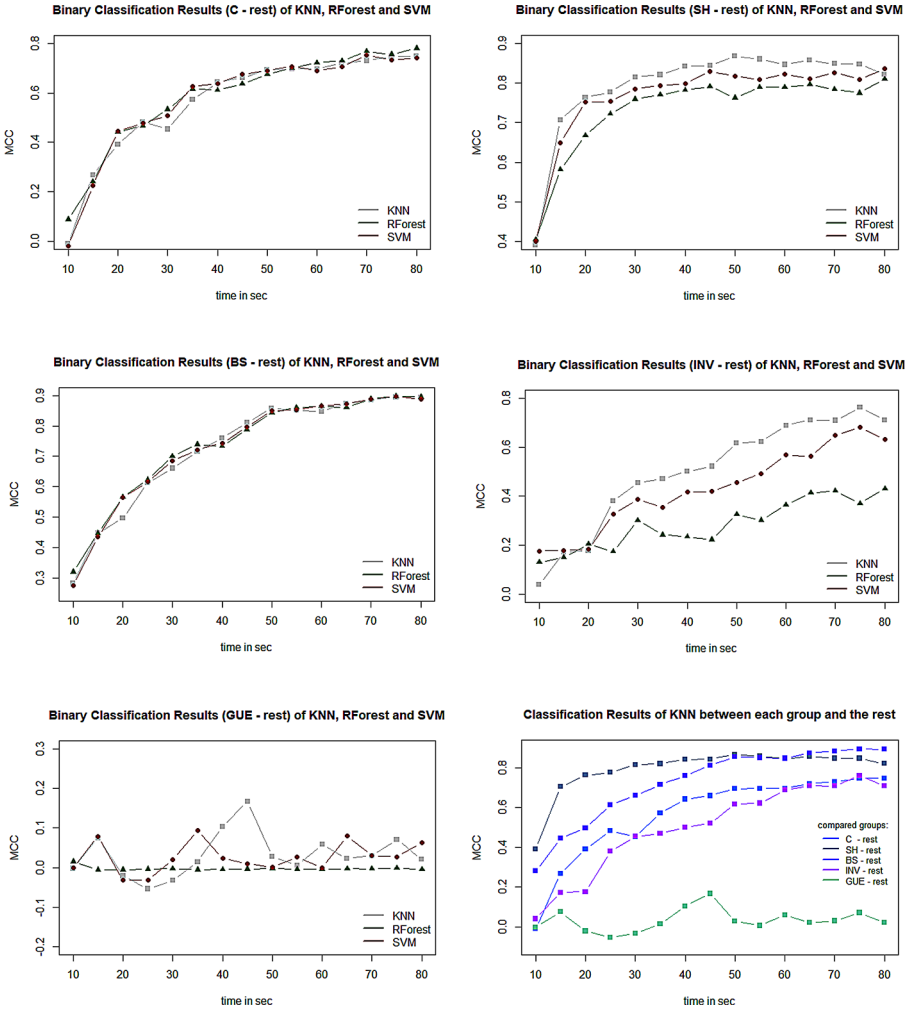

Fig. 7 MCC of the five Binary Classifications, each classifying participants into one of two classes (into one of the coding categories or the respective rest) with Classifiers KNN, Random Forest and SVM. The last plot on the bottom right combines the MCC Scores of all five binary classifications using the KNN algorithm

Discussion and Outlook

Considering prior work on prediction students' visual strategies in the context of linear graphs the classification results of the full data in the context of graphical differentiation reveal that it is possible to predict student difficulties that may occur based on a person's gaze data. This could significantly improve adaptive systems, as it would allow for truly individualised feedback that addresses the learner's misconception during a problem-solving or learning process. Looking at the prediction results of the temporally truncated data, which simulates the online decision-making situation, we were able to see that the quality of such a multiclass classification can

be very high even before the end of the solution process. This fact, together with the possibility to analyse eye tracking data in real time, can enable the AES to intervene already in the solution process, e.g., by highlighting important areas or by overlaying visual assistance.

When implementing and integrating eye tracking into an adaptive system, the question of a suitable intervention time arises. If the time is chosen too early, there is a risk that the algorithm will assign many learners to the wrong class. This is especially problematic if assistance is offered too early, even though the learners would have found the correct solution on their own. If the intervention time is chosen too late, on the other hand, a large proportion is no longer reachable, as the average processing time was 43 s. This problem could be circumvented by providing an additional opportunity for assistance before the solution is given. One possibility to set an appropriate intervention time is to choose a point in time for which it is known that the prediction quality exceeds a certain fixed value, e.g. MCC > 0.7. In addition to choosing an appropriate intervention time, the decision on the type of assistance is also crucial for its effectiveness. There is a lot of studies indicating that support should be adapted to the prior knowledge (e.g. Richter et al., 2021). Although the intervention type is not the focus of this work, we can say that by identifying the specific student difficulties with the help of eye tracking, it becomes possible to accurately address these difficulties. In this way, a visual cue, for example, is only displayed to students who need specific support.

In our example, the KNN reaches a prediction quality of MCC > 0,7 after 45 s in the multiclass case. However, as can be seen from the average processing times in Table 4, a hint at this stage would come too late for many learners. Because of their strategy, those who made the slope-height error, for example, finished very quickly and only needed half a minute on average. The results of the binary classifications, which are illustrated in Fig. 7, show that some groups can be better identified than others. Thus, for the slope-height error group, assistance could be given much earlier. A classification based only on the first 15 s already results in a prediction quality of 0.7. For the block-slope error group, assistance, e.g. a visual highlight on the y-axis, could also appear much earlier, as the MCC exceeds 0,7 after 35 s. The prediction quality of the binary classification with splitting correct-rest is similar to the quality of the multiclass classification. Thus, the group of participants who solved the task correctly was also well recognised. This is very important, as it is essential to avoid giving help even though it is not needed. The prediction quality for the inverted-slope error or for rates is much lower. For the INV-rest distinction, the prediction quality only reaches a value of 0.7 for the MCC after about 60 s. It is likely that areas that are typical for this student difficulty are only considered at a relatively late stage in the solution process, e.g., because learners scan the graph with their eyes from left to right and first look at other areas. The 3 features that are supposed to separate the inverted-slope error group particularly well from the other groups consistently lead to significant differences (Table 7). Thus, it is possible that the prediction quality would improve with a larger data set. The results show that it is very difficult to judge from the gaze data whether someone is guessing or not. This might be due to the heterogeneity of this group: there are test persons who answered very quickly most likely in combination with low motivation and did not even seriously try to solve the task. On

the other hand, this group also includes persons who spent an above-average amount of time on the task and simply did not come to any conclusion. These test persons often looked at all areas very long and thoroughly. Another complicating factor is that the information about the confidence is based on the subjective judgement of the test persons. Furthermore, no special features were defined for the guess group to distinguish it from the others. Overall, it has been shown that the student difficulties can be classified well based on eye tracking data, either at an earlier or later point in time. It would be very interesting to explore which features are suitable for recognising uncertainties and to include them as features in the classification so that those who are guessing can also be detected.

Our results show that some student difficulties can be classified earlier than others, so that it is possible to set a specific intervention time for each student difficulty. A possible approach to finding suitable intervention times is thus to set a limit for the desired predictive quality and to determine a specific time for each student difficulty, accordingly to that limit. An open question remains how high the prediction quality should be in order to define the corresponding time as an intervention time. If we set the bound to MCC = 0.7, the following intervention times could be defined for the discussed item: slope-height assistance after 15 s, block-slope assistance after 30 s, inverted-slope assistance after around 60 s. Even if the learner has already finished the task, a prediction about a student difficulty can be helpful for the learner. In this case, the system can no longer help in the solution process, but can provide feedback with diagnosis about the student difficulty. It would also be possible to adapt the time to the learner, since there are great differences, for example, in the reading speed of the task text and also in the general approach to solve a task. Here, learning types would first have to be identified.

The approach has been demonstrated for a subfield of calculus. Although generalization is challenging due to smaller data sets, the approach has broader implications, because the method described is transferable to other items. Only the features need to be adapted. If it is not possible to have experts to determine relevant areas and areas belonging to student difficulties, the AOIs for the features could also be detected automatically. For example, areas with large differences in the gaze data of the groups could be systematically identified by placing a uniform AOI grid over the task and carrying out statistical tests (see Fig. C1 in Appendix C) or using a machine learning algorithm (Rebello et al., 2018). Furthermore, one could consider AOI-independent features, such as the saccade length in this example. The saccade velocity or the observation of scanpaths, for example, could also be good predictors. The RTAs were used for coding in this study. It was investigated how well the learners' true difficulties (given by the RTAs) can be recognized purely on the basis of gaze data. For later implementations of an eye tracking based adaptive system, the use of interviews would at best no longer be necessary. However, the use of concurrent think-alouds (CTAs) could be beneficial here (despite the possible influence on the gaze) and should be investigated in further research.

The presented approach could have great potential for constructing adaptive learning systems. It becomes particularly interesting when it is applied to more difficult tasks that also involve a longer processing time. We are currently working on the implementation of a prototype.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40593-024-00447-5 .

Acknowledgements The project MAL-i (sub-project of U.EDU: Unified Education) on which this paper is based is part of the "Qualitätsoffensive Lehrerbildung", a joint initiative of the Federal Government and the Länder which aims to improve the quality of teacher training. The programme is funded by the Federal Ministry of Education and Research. The authors are responsible for the content of this publication.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education*, *125*, 413–428.

Arcavi, A. (2003). The role of visual representations in the learning of mathematics. *Educational Studies in Mathematics*, *52*(3), 215–241.

Asiala, M., Cottrill, J., Dubinsky, E., & Schwingendorf, K. E. (1997). The development of students' graphical understanding of the derivative. *The Journal of Mathematical Behavior*, *16*(4), 399–431.

Aspinwall, L., Shaw, K. L., & Presmeg, N. C. (1997). Uncontrollable mental imagery: Graphical connections between a function and its derivative. *Educational Studies in Mathematics*, *33*(3), 301–317.

Augereau, O., Fujiyoshi, H., & Kise, K. (2016). Towards an automated estimation of English skill via TOEIC score based on reading analysis. In Proceedings of the 23rd International Conference on Pattern Recognition. IEEE, 1285–1290.

Barral, O., Lallé, S., Iranpour, A., & Conati, C. (2021). Effect of adaptive guidance and visualization literacy on gaze attentive behaviors and sequential patterns on magazine-style narrative visualizations. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3–4), 1–46.

Barral, O., Lallé, S., Guz, G., Iranpour, A., & Conati, C. (2020, October). Eye-tracking to predict user cognitive abilities and performance for user-adaptive narrative visualizations. In Proceedings of the 2020 international conference on multimodal interaction (pp. 163–173).

Barrios, V. M. G., Gütl, C., Preis, A. M., Andrews, K., Pivec, M., Mödritscher, F., & Trummer, C. (2004). AdELE: A framework for adaptive e-learning through eye tracking. Proceedings of IKNOW, 609–616.

Becker, S., Küchemann, S., Klein, P., Lichtenberger, A., & Kuhn, J. (2022). Gaze patterns enhance response prediction: More than correct or incorrect. *Physical Review Physics Education Research*, *18*(2), 020107.

Biedert, R., Buscher, G., Schwarz, S., Möller, M., Dengel, A., & Lottermann, T. (2010, February). The text 2.0 framework: writing web-based gaze-controlled realtime applications quickly and easily. In Proceedings of the 2010 workshop on Eye gaze in intelligent human machine interaction (pp. 114–117).

Bixler, R., & D'Mello, S. (2016). Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*, *26*(1), 33–68.

Brusilovsky, P. (1998, August). Adaptive educational systems on the world-wide-web: A review of available technologies. In Proceedings of Workshop WWW-Based Tutoring at 4th International Conference on Intelligent Tutoring Systems (ITS'98), San Antonio, TX.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*, 1–13.

Cho, P., & Nagle, C. (2017). Procedural and conceptual difficulties with slope: An analysis of students' mistakes on routine tasks. *International Journal of Research in Education and Science*, *3*(1), 135–150.

Clement, J. (1985, July). Misconceptions in graphing. In Proceedings of the ninth international conference for the psychology of mathematics education (Vol. 1, pp. 369–375). Utrecht, The Netherlands: Utrecht University.

Conati, C., & Merten, C. (2007). Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation. *Knowledge-Based Systems*, *20*(6), 557–574.

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer Studies*, *70*(5), 377–398.

D'Mello, S. K., Mills, C., Bixler, R., & Bosch, N. (2017). *Zone out no more: Mitigating mind Wandering during Computerized Reading*. International Educational Data Mining Society.

Dzsotjan, D., Ludwig-Petsch, K., Mukhametov, S., Ishimaru, S., Küchemann, S., & Kuhn, J. (2021, September). The Predictive Power of Eye-Tracking Data in an Interactive AR Learning Environment. In Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (pp. 467–471).

Emerson, A., Min, W., Rowe, J., Azevedo, R., & Lester, J. (2023, March). Multimodal predictive student modeling with multi-task transfer learning. In LAK23: 13th International Learning Analytics and Knowledge Conference (pp. 333–344).

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, *23*(4), 523–552.

Gingerich, M., & Conati, C. (2015, February). Constructing models of user and task characteristics from eye gaze data for user-adaptive information highlighting. In Proceedings of the AAAI conference on artificial intelligence (Vol. 29, No. 1).

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, *28*(5–6), 367–374.

Greefrath, G., Oldenburg, R., Siller, H. S., Ulm, V., & Weigand, H. G. (2016). *Didaktik Der Analysis*. Springer Berlin Heidelberg.

Gusukuma, L., Bart, A. C., Kafura, D., & Ernst, J. (2018, August). Misconception-driven feedback: Results from an experimental study. In *Proceedings of the 2018 ACM Conference on International Computing Education Research* (pp. 160–168).

Haider, H., & Frensch, P. A. (1999). Eye movement during skill acquisition: more evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 172.

Hussmann, S., & Prediger, S. (2010). Vorstellungsorientierte analysis–auch in Klassenarbeiten Und Zentralen Prüfungen. *Praxis Der Mathematik in Der Schule*, *52*(31), 35–38.

Ivanjek, L., Susac, A., Planinic, M., Andrasevic, A., & Milin-Sipus, Z. (2016). Student reasoning about graphs in different contexts. *Physical Review Physics Education Research*, *12*(1), 010106.

Jacob, S., Ishimaru, S., Bukhari, S. S., & Dengel, A. (2018, June). Gaze-based interest detection on newspaper articles. In Proceedings of the 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction (pp. 1–7).

Jaques, N., Conati, C., Harley, J. M., & Azevedo, R. (2014). Predicting affect from gaze data during interaction with an intelligent tutoring system. In Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5–9, 2014. Proceedings 12 (pp. 29–38). Springer International Publishing.

Jurman, G., Riccadonna, S., & Furlanello, C. (2012). A comparison of MCC and CEN error measures in multi-class prediction.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329.

Kennel, K., Becker, S., Klein, P., Küchemann, S., Kuhn, J., & Ruzika, S. (2022). Blickbewegungen Beim Grafischen Ableiten–Lassen Sich Fehler Durch Eye-Tracking-Daten vorhersagen und elaborieren? *Eye-Tracking in Der Mathematik-Und Naturwissenschaftsdidaktik: Forschung Und Praxis* (pp. 125–143). Springer Berlin Heidelberg.

Klinger, M. (2017). *Funktionales Denken Beim Übergang Von Der Funktionenlehre Zur Analysis: Entwicklung eines testinstruments und empirische Befunde Aus Der Gymnasialen Oberstufe*. Springer-.

Küchemann, S., Becker, S., Klein, P., & Kuhn, J. (2021). Gaze-Based Prediction of Students' Understanding of Physics Line-Graphs: An Eye-Tracking-Data Based Machine-Learning Approach. In Computer Supported Education: 12th International Conference, CSEDU 2020, Virtual Event, May 2–4, 2020, Revised Selected Papers 12 (pp. 450–467). Springer International Publishing.

Küchemann, S., Cullmann, N., Kovac, S., Becker, S., Klein, P., Kennel, K., & Kuhn, J. (2022). Blickverhalten Beim Lernen Und Problemlösen Mit Graphen–Ein Literaturüberblick bis 2020. *Eye-Tracking in Der Mathematik-Und Naturwissenschaftsdidaktik* (pp. 177–192). Springer Spektrum.

Küchemann, S., Klein, P., Becker, S., Kumari, N., & Kuhn, J. (2020, May). Classification of Students' Conceptual Understanding in STEM Education using Their Visual Attention Distributions: A Comparison of Three Machine-Learning Approaches. In CSEDU (1) (pp. 36–46).

Lallé, S., Toker, D., & Conati, C. (2019). Gaze-driven adaptive interventions for magazine-style narrative visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *27*(6), 2941–2952.

Lallé, S., Conati, C., & Carenini, G. (2016, July). Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data. In IJCAI (pp. 2529–2535).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics, 159–174.

Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research*, *60*(1), 1–64.

Li, S., Lajoie, S. P., Zheng, J., Wu, H., & Cheng, H. (2021). Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving. *Computers & Education*, *163*, 104114.

Madsen, A. M., Larson, A. M., Loschky, L. C., & Rebello, N. S. (2012). Differences in visual attention between those who correctly and incorrectly answer physics problems. *Physical Review Special Topics-Physics Education Research*, *8*(1), 010122.

Marwan, S., Gao, G., Fisk, S., Price, T. W., & Barnes, T. (2020, August). Adaptive immediate feedback can improve novice programming engagement and intention to persist in computer science. In Proceedings of the 2020 ACM conference on international computing education research (pp. 194–203).

Okoso, A., Toyama, T., Kunze, K., Folz, J., Liwicki, M., & Kise, K. (2015). Towards extraction of subjective reading incomprehension: Analysis of eye gaze features. In Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems: Extended Abstracts. ACM, 1325–1330.

Rebello, N. S., Nguyen, M. H., Wang, Y., Zu, T., Hutson, J., & Loschky, L. C. (2018). Machine learning predicts responses to conceptual tasks using eye movements. In Physics Education Research Conference 2018, PER Conference, Washington, DC (Vol. 10).

Richter, J., Wehrle, A., & Scheiter, K. (2021). How the poor get richer: Signaling guides attention and fosters learning from text-graph combinations for students with low, but not high prior knowledge. *Applied Cognitive Psychology*, *35*(3), 632–645.

Ruf, V., Horrer, A., Berndt, M., Hofer, S. I., Fischer, F., Fischer, M. R., & Küchemann, S. (2023). A literature review comparing experts' and non-experts' visual Processing of Graphs during problem-solving and learning. *Education Sciences*, *13*(2), 216.

Salvucci, D. D., & Goldberg, J. H. (2000, November). Identifying fixations and saccades in eye-tracking protocols. In Proceedings of the 2000 symposium on Eye tracking research & applications (pp. 71–78).

Scheiter, K., Schubert, C., Schüler, A., Schmidt, H., Zimmermann, G., Wassermann, B., & Eder, T. (2019). Adaptive multimedia: Using gaze-contingent instructional guidance to provide personalized processing support. *Computers & Education*, *139*, 31–47.

Schmidt, H., Wassermann, B., & Zimmermann, G. (2014). An adaptive and adaptable learning platform with realtime eye-tracking support: Lessons learned. *DeLFI 2014-Die 12*. e-Learning Fachtagung Informatik.

Steichen, B., Carenini, G., & Conati, C. (2013, March). User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In Proceedings of the 2013 international conference on Intelligent user interfaces (pp. 317–328).

Taub, M., & Azevedo, R. (2019). How does prior knowledge influence eye fixations and sequences of cognitive and metacognitive SRL processes during learning with an intelligent tutoring system? *International Journal of Artificial Intelligence in Education*, *29*, 1–28.

Tobii, A. B. (2015). Accuracy and precision test report: Tobii Pro X3-120 fw 1.7.1. https://www.tobiipro.com/siteassets/tobii-pro/accuracy-and-precision-tests/tobii-pro-x3-120-accuracy-and-precision-test-report.pdf

Ubuz, B. (2007). *Interpreting a graph and constructing its derivative graph: Stability and change in students' conceptions*. International Journal of a.

Van Someren, M., Barnard, Y. F., & Sandberg, J. (1994). *The think aloud method: A practical approach to modelling cognitive* (Vol. 11, pp. 29–41). Academic.

Yamada, K., Kise, K., & Augereau, O. (2017). Estimation of confidence based on eye gaze: an application to multiple-choice questions. In Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. ACM, 217–220.

Zhang, X., Sugano, Y., & Bulling, A. (2019). Evaluation of appearance-based methods and implications for gaze-based applications. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13).

Zu, T., Hutson, J., Loschky, L. C., & Rebello, N. S. (2020). Using eye movements to measure intrinsic, extraneous, and germane load in a multimedia learning environment. *Journal of Educational Psychology*, *112*(7), 1338.

Authors and Affiliations

Kathrin Kennel[1] · Shoya Ishimaru[2] · Stefan Küchemann[3] · Steffen Steinert[1,3] · Jochen Kuhn[3] · Stefan Ruzika[1]

✉ Kathrin Kennel
kathrin.kennel@math.rptu.de

Shoya Ishimaru
ishimaru@omu.ac.jp

Stefan Küchemann
s.kuechemann@physik.uni-muenchen.de

Steffen Steinert
s.steinert@physik.rptu.de

Jochen Kuhn
jochen.kuhn@physik.uni-muenchen.de

Stefan Ruzika
stefan.ruzika@math.rptu.de

[1] RPTU – Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Kaiserslautern, Germany

[2] OMU – Osaka Metropolitan University, Osaka, Japan

[3] LMU – Ludwig-Maximilians-Universität München, Munich, Germany