Comparison and AI-based prediction of graph comprehension skills based on the visual strategies of first-year physics and medicine students

Verena Ruf,[1,*] Yavuz Dinc,[1] Stefan Küchemann,[1] Markus Berndt,[2] Steffen Steinert,[1,3] Daniela Kugelmann,[4] Jonathan Bortfeldt,[5] Jörg Schreiber,[5] Martin R. Fischer,[2] and Jochen Kuhn[1]

[1]*Chair of Physics Education Research, Faculty of Physics, Ludwig-Maximilians-Universität München, Munich, Germany*
[2]*Institute of Medical Education, Ludwig-Maximilians-Universität University Hospital, Ludwig-Maximilians-Universität München, Munich, Germany*
[3]*Department of Electrical and Computer Engineering, University of Kaiserslautern-Landau, Kaiserslautern, Germany*
[4]*Institue of Anatomy, Faculty of Medicine, Ludwig-Maximilians-Universität München, Munich, Germany*
[5]*Chair of Medical Physics, Faculty of Physics, Ludwig-Maximilians-Universität München, Garching, Germany*

(Received 18 October 2023; accepted 3 September 2024; published 4 November 2024)

Graphical representations of data are common in many disciplines. Previous research has found that physics students appear to have better graph comprehension skills than students from social science disciplines, regardless of the task context. However, the graph comprehension skills of physics students have not yet been compared with (veterinary) medicine students, both of which are disciplines that require multiple science, technology, engineering, and mathematics (STEM) courses. This study extends previous research on this subject by exploring whether physics majors possess superior graph comprehension skills due to their study discipline. Here, participants solved 24 graph comprehension tasks across various subjects, including mathematics, physics, and medicine; these tests were conducted at the beginning and end of their first semester. Graph comprehension gain was calculated based on the percentage of correct and incorrect answers in the pretest and the post-test. In addition to these comparisons, we replicated previous research that successfully distinguished correct and incorrect solvers based on their visual behavior by using a novel machine-learning method tailored to small datasets. Through this replication of statistical analyses, we aim to ensure the reliability of adaptive learning systems in the future, regardless of data size, using the same machine-learning method. Physics and medical students were found to exhibit relatively similar graph comprehension gain; this is in contrast to previous research comparing physics and non-STEM students. Our results also revealed that both physics and medical students use similar visual strategies to solve these tasks. However, correct and incorrect solvers could be distinguished via machine-learning methods regardless of their discipline. Our research suggests that visual behavior is a good predictor of graph comprehension skills.

DOI: 10.1103/PhysRevPhysEducRes.20.020138

I. INTRODUCTION

Graphical representations are common in daily life. They are also often used in educational contexts, most notably in science, technology, engineering, and mathematics (STEM) education. Graphical representations are frequently found in learning environments and their benefits to learners are well documented both theoretically and empirically [1]. Graphical representations of data (i.e., graphs) are important, particularly in disciplines where the use of data is common, such as the aforementioned STEM disciplines. As dealing with information is one of the key competencies of the 21st century [2], the ability to handle information is a crucial learning point in students' education. However, many students can encounter difficulties when dealing with graphical representations of data, such as interpreting the represented data [3–5].

Previous research has shown that physics students have a definitive advantage when solving problems associated with graphical information compared to students from social sciences disciplines [6–8]. As these studies compared

[*]Contact author: v.ruf@physik.uni-muenchen.de

physics majors with students from non-STEM disciplines, the specific reasons for the advantages exhibited by physics majors are unclear. For example, physics students could be better at solving graphical problems due to their chosen discipline or other factors (e.g., because they study a STEM discipline that requires the more frequent handling of data compared to social science disciplines). Other factors, such as higher mathematical skills, could also improve graph comprehension skills.

To extend previous research on this subject, we conducted a pretest-post-test study comparing physics majors to a group of non-STEM students that consistently engage with several STEM courses: veterinary medicine and medicine students. Both physics and (veterinary) medicine students are voluntarily enrolled in disciplines with many courses that engage with data and require a high level of mathematical knowledge. In this study, we investigate whether the discipline itself—i.e., whether the difference between physics majors, which consist mainly of mathematics and physics courses, and (veterinary) medicine majors, which contain physics, biology, and chemistry courses—plays a key role in the difference in students' graph comprehension skills.

Previous research also compared the eye movements of students from physics majors with those of students of other disciplines as they were solving graphical tasks from both their respective disciplines [6–8]. The results revealed that correct and incorrect solvers could be distinguished based on their eye movements using statistical analyses [7].

Eye movements can also be valuable inputs for adaptive learning systems [9]. These systems will probably become more common, especially with the advent of web-based eye tracking [10]. Adaptive learning systems use machine-learning methods to analyze data instead of statistical analyses and can provide immediate feedback and scaffolding to students during the learning and problem-solving processes [11]. Process measures, such as eye-tracking data, are ideally suited as input for machine-learning models because they not only act as indicators of a student's performance [12] but also provide insights into their problem-solving strategies [13]. However, machine-learning methods must be able to accurately and reliably replicate the results of statistical analyses to ensure the reliability of adaptive learning systems. As a first step, this study attempts to previous results by using machine-learning techniques to predict students' performance based on their eye movements.

This manuscript is organized as follows: we first introduce the relevant literature (Sec. II) before describing our methodology (Sec. III). We then present our results on the differences in the participants' graph comprehension gain as well as an analysis of the eye movement data collected during the study (Sec. IV). We proceed to discuss the results, possible implications for educational practice, and the study's limitations (Sec. V). Finally, we present our conclusions (Sec. VI).

II. LITERATURE REVIEW

A. Learning with graphs

Many STEM disciplines use graphical representations of data within their learning environments, such as velocity-time diagrams in physics courses. Previous research suggests that such representations facilitate data processing and comparison [14]. Consequently, graph comprehension is a common research topic on which multiple reviews have been published [3–5,15]. In the literature, two terms referring to the same skill are often used interchangeably: graph interpretation (e.g., [3]) and graph comprehension (e.g., [16]). This skill describes one's ability to "obtain meaning from graphs" [4] (p. 190), which either can be constructed by oneself or provided by others. Graph comprehension is a crucial ability for many students [2,4] and often requires domain knowledge, knowledge about the context of the graph, graphical skills (e.g., knowledge about the axes), and explanatory skills [17]. Shah and Hoeffner [15] described three general steps of graph comprehension: (i) identifying the relevant visual features, (ii) relating those features to the "conceptual relations that are represented by those features" [15] (p. 53), and (iii) connecting this information to the appropriate domain concepts. Prior knowledge can also influence a viewer's interpretations of a graph [4,15].

Interpreting graphs is a challenging activity for students [4]. The most common difficulties faced by students include interpreting a graph as a picture (graph-as-picture errors) [5,18] or understanding functions (e.g., translating between a function graph and the corresponding equation) [5]. For example, 10th-grade students, who were asked to create a position-time graph, did not recognize graph-as-picture errors even when revising the line of their graph [18]. In physics, additional difficulties include identifying the relevant features of a graph or interpreting the data depicted in a graph [19]. Graph comprehension difficulties are also common in medical fields: for example, participants had trouble identifying conflicting information in graphs that presented medical information, such as treatment effects [20]. Furthermore, relating mathematical concepts with their contextual meaning is often not intuitive [21].

The context and characteristics of a discipline can influence graph comprehension [16]. Due to its importance across many disciplines, it is important to ensure that students can interpret graphical representations independently of their context. *Representational competence* refers to a student's knowledge of how content is presented in visual representations. It involves visual understanding, including the ability to identify relevant variables and information, as well as the ability to connect these representations to their respective concepts [22]. The ability to use representations and understand their meaning independently of the context is a part of *meta-representational competence* [23]—this is a key aspect of representational competence [22] and is a valuable skill in the context of graph comprehension [24]. However, students gain content knowledge at the same time as they acquire representational competence [22]; this prior knowledge and experience can contribute to learning. This is often assumed to be the case when utilizing previously learned procedures in a new situation [25], such as when students use the same method to calculate the slope of a function in both math and physics problems. Ideally, students should be able to apply their prior knowledge of problem-solving routines across multiple disciplines, such as calculating the slope of a line graph in both math and physics using the correct units. Once again, this requires representational competence, such as the ability to correctly identify which variables should be present on each axis, their units, and domain knowledge about the relationship between those variables. Cognitive processes during problem solving, such as identifying relevant features in a graph and the relationship between those features, can be inferred from eye movements [26].

B. Eye movements

Eye movement analyses, such as the length of fixations, are often used in educational research [26,27] because they can indicate attentional processes (eye-mind hypothesis, [28]). Education studies frequently analyze the differences between experts and nonexperts in the context of STEM education; for example, in terms of domain knowledge, e.g., physics education research [29] or graph comprehension skills [30]. In particular, information processing can differ between experts and nonexperts. There are several theories as to why this might be the case: these include the theory of long-term working memory [31], the model of holistic image processing [32], and the information-reduction hypothesis [33]. All theories imply that experts engage in more efficient information processing, which can be investigated via their eye movements. According to the theory of long-term working memory, experts store larger chunks of information in working memory compared to nonexperts [34] and they access these chunks via retrieval cues [31]. Consequently, experts' fixation durations should be shorter due to the reduced time needed to retrieve information [35]. In contrast, the holistic model of image perception approaches the issue from a perceptual level: experts process images globally [32], which leads to more efficient searches [35]. This is assumed to lead to quicker fixations on relevant information [35,36]. Finally, the information-reduction hypothesis states that experts can focus on relevant information by ignoring information irrelevant to the current task [33]; this suggests that experts should fixate on relevant information longer [35].

Such differences in expertise were also found in the visual strategies used during graph comprehension [30,35,36]. Experts paid more attention to relevant information than nonexperts during learning and problem-solving involving graphs [30]. This is consistent with the information-reduction hypothesis [33]. Previous research compared participants with varying levels of expertise as they solved science-related problems that used conventional representations of data: they found that participants of differing expertise levels used different visual strategies [37]. Okan *et al.* [20] observed that participants with high graph literacy looked at relevant graphical information for longer periods when making their interpretations. Differences in expertise may also be influenced by specific disciplines. In a study on 131 high-school students, Becker *et al.* [38] found that the participants had problems applying mathematical knowledge to a kinematic context, especially for line graphs with negative gradients. Similar results were reported by Ceuppens *et al.* [39]. Furthermore, students tended to look along the line of the function when attempting math problems while instead choosing to pay closer attention to the axes when attempting kinematics problems [38]. In previous studies with higher-education participants, physics students solved graphical problems better than both economics students [6,7] and psychology students [8], regardless of the subject. Susac *et al.* [8] measured and compared the performance and viewing times of psychology ($N = 45$) and physics ($N = 45$) students as they attempted to solve isomorphic physics and finance problems. An analysis of students' strategies indicated that physics students relied on equations for a solution, while psychology students used more "common-sense strategies" [8] (p. 1) (i.e., rise over run) that were more likely to lead to errors, such as confusing points and intervals [8]. Klein *et al.* [7] replicated these results and found that physics students ($N = 29$) performed better than economics students ($N = 40$) on the same isomorphic test items even though the finance problems were, from a content perspective, more closely related to the field of study of the economics students. They also found that students who solved the tasks correctly focused longer on concept-specific areas than those who did not [7]. This was consistent with Susac *et al.* [8] who found that physics students analyzed the actual graph more carefully compared to psychology students. Students were also asked to report their confidence in their answers—these scores revealed that physics students were better judges of the correctness of their answers compared to economics students, although the scores did not differ significantly between groups [7]. These results were replicated in a postreplication study employing a pretest-posttest design [6]. Both physics ($N = 20$) and economics ($N = 21$) students solved the same problems as participants in the previous studies both at the beginning and the end of their first semester. Importantly, the participants were a matched sample to the participants in the pretest study conducted by Klein *et al.* [7]. In the post-test, it was once again observed that physics students performed better than economics students. However, "students from both domains showed a similar increase in the overall test score" [6] (p. 8).

In addition, there were no differences in visual behavior between the correct and incorrect solvers from different disciplines, although physics students tended to be overconfident in their answers during the post-test [6].

C. Machine learning using eye movements

Previous studies used statistical methods to compare correct and incorrect solvers' eye movements and visual processing. For example, Klein *et al.* [7] used ANOVAs to compare the eye movements of physics and economics students. Statistical tests, such as regression analyses (i.e., ANOVAs), can be used to predict the influence of independent variables, such as eye movements, on the dependent variable, such as problem-solving performance [40]. Whereas statistical tests are a well-established method to analyze data for hypothesis testing [41], adaptive learning systems often use large datasets with many independent variables and possibly complex correlations [42]. Therefore, such adaptive systems frequently employ machine-learning techniques, because they outperform statistical methods [42]. This is probably due to the advantages of machine learning when analyzing complex nonlinear relations between variables. Such considerations need to be kept in mind, considering the increasing number of personalized and adaptive learning systems [43].

Statistical tests, such as ANOVAs—that were previously used to analyze eye movements [7], assume that data are based on underlying stochastic models [44]. In contrast, algorithmic models employing machine-learning methods assume that the underlying data distribution is unknown [44]. Machine-learning methods can use training data to generate generalizable and scalable models, which can be applied to unknown datasets of different sizes [45]. Such models are more robust to changes if the distribution of new incoming data differs strongly from the original data on which the model is trained. Machine learning can also be used to investigate complex (e.g., nonlinear or unknown) correlations [45], allowing users to find new patterns in the data without making previous assumptions regarding the data's distribution. An overview comparing statistical and machine-learning methods is presented in Table I. Combining the two approaches—statistical modeling and machine learning—in a hybrid approach can have various advantages, such as investigating causal effects while leveraging the benefits of machine learning by identifying the most predictive variables [41]. In this way, causality can be examined by comparing models trained with multiple variables and based on various assumptions; additionally, the predictive power of scientific theories can be tested [41]. This approach is suitable due to the increasing use of machine-learning techniques for complex analyses with real-time predictions on large datasets and the benefits of statistical methods employed to test hypotheses. Currently, datasets in education are often small (i.e., 29 physics students and 40 economics students [7]), whereas we believe that large amounts of data will be available to future adaptive learning systems. At the moment, small datasets make applying machine-learning techniques more difficult due to the limited amount of training data, which increases the necessity of testing the suitability of machine learning in a hybrid approach—including the benefits of both statistical and machine-learning analyses. Using an optimized method specifically designed for small datasets (i.e., Ref. [46]) in such an approach can lead to even more reliable results.

Most machine-learning methods used in educational contexts are supervised machine-learning models [42]; this refers to algorithms that can predict the outcome of unknown data based on labeled training data [45]. Only a portion of the collected data is used for training; in general, only 80% of the data is used as training data, with the remaining 20% withheld as test data [45]. Machine-learning tools can be applied to evaluate eye-tracking data: Supervised machine-learning methods have previously been employed to predict performance based on eye-tracking metrics [12,49,50]. Küchemann *et al.* [12] successfully predicted the performance of 11th-grade high-school students' ($N = 115$) on the Test of Students' Understanding of Graphs in Kinematics [51] based on the total visit duration on areas of interest (AOIs) and the frequency of gaze switches between AOIs with a support vector machine (SVM). Mozaffari *et al.* [52] and Rebello *et al.* [53] achieved similar results when students attempted

TABLE I. An overview of statistical and machine-learning methods.

	Statistical methods	Machine-learning methods
Purpose	Hypothesis testing [41] Statistical inference [41]	Model complex relationships [44,45] Discover new patterns
Focus	Investigating the influence of relation between individual variables [44]	Making accurate predictions [41,44]
Measurement of variable importance	Exact coefficients [47]	Feature importance based on the trained model [45]
Proposed application	Theory-based research [41,44]	Adaptive learning systems [42,48]
Recommended use cases	Simple assumptions Complex assumptions	Large dataset with complex assumptions

physics-based tasks. In addition to duration metrics, machine-learning models can predict performance using other eye-tracking metrics, such as the number of saccades and pupil size [54]. From an educational perspective, this research is highly relevant to adaptive learning systems, which have become more relevant due to the increasing use of artificial intelligence in education [48]. For example, adaptive learning systems can support students during problem-solving tasks by prompting them if they do not look at relevant areas.

In this study, we compare several machine-learning algorithms, including an SVM. SVMs have already been successfully applied to learning analytics [55] and have also been used in studies attempting to predict task performance based on visual behavior [12,49,52,53]. SVMs can be used as binary classifiers, capable of separating two groups by defining a boundary that attempts to maximize the margin of error on either side [45]. They can also be used in real time [55,56]. Other classification algorithms include the k-nearest neighbor (KNN), which forms groups based on the distance between data points and random forest (RF) models, which separate data points into groups based on an ensemble of decision trees [45]. Dzsotjan *et al.* [49] compared the accuracy of various machine-learning algorithms trained to predict performance based on visual behavior in an interactive augmented reality environment finding that SVMs returned high F1 scores. This suggests that eye movements are suitable inputs that can be used to predict students' performance on graph comprehension tasks using machine-learning techniques such as SVMs.

D. Research questions

Previous studies have compared physics students with non-STEM students in terms of their performance at graph comprehension tasks and found that physics students outperformed non-STEM students regardless of context [6–8]. In this study, we extend this research comparing physics majors with non-STEM students [6–8] by comparing the graph comprehension skills of physics majors with non-STEM students who also take STEM-related courses. To the best of our knowledge, no studies have compared the visual behavior of physics majors (STEM students) with that of (veterinary) medicine students (non-STEM students) who take STEM courses, such as anatomy, chemistry, and physics. In particular, while mathematics is a key subject for physics majors, medical students do not take any math classes. Based on the results of previous studies, we pose the following research questions:

(1) Are there differences in graph comprehension gain of physics and non-physics majors?
(2) Can differences in students' performance be accurately identified via machine-learning models trained on students' visual behavior?

If physics students perform better due to their discipline rather than the amount or type of STEM courses taken, then physics students should outperform medicine students as was found in previous studies comparing physics students to social sciences students [6–8]. However, if STEM courses are relevant to graph comprehension capabilities in general, then both the physics and veterinary medicine/medical students in this study should exhibit equal graph comprehension.

To answer the second research question, we replicated the results of Klein *et al.* [7] and assessed the visual differences between correct and incorrect solvers with machine-learning methods. Eye movements can be used as input for adaptive learning systems [9] and as a predictor of performance (Sec. II C). However, machine-learning models must be both accurate and reliable in order for them to be successfully employed as personalized systems in educational environments. In other words, it must be possible to replicate the statistical findings of machine-learning methods, even for small datasets common in educational contexts.

III. METHODS

We recruited first-semester students as participants in our pretest-post-test study. The participants' demographics, study design, test materials, apparatus used in the study, and machine-learning techniques used in this study are described in the following sections.

A. Participants

Twenty-four first-semester students from the Ludwig-Maximilians-Universität München (LMU Munich) voluntarily participated in the study at the beginning and the end of the semester. The students were enrolled in the physics ($N = 9$), physics education ($N = 3$), medicine ($N = 4$), or veterinary ($N = 8$) programs. A typical first semester in physics includes only STEM courses: experimental physics, mathematics, and physics labs. Both medical and veterinary students take the chemistry and anatomy STEM courses as well as an additional STEM course (medicine: biology, veterinary medicine: physics) and non-STEM course (medicine: medical terminology, veterinary medicine: zoology), depending on their vocation. In particular, mathematics is typically not part of the medical curriculum. Physics majors ($\mu = 19.25$ years) were the same age as the veterinary and medical students ($\mu = 19.75$ years); $t(15.177) = -0.62,\ p = 0.55$. Participants received €20 for their participation in both the pretest and the post-test. Most students were native German speakers; one student spoke German at the C1 level. The participants' mean A-level grade was 1.5 (i.e., the middle grade between A and B in the United States). There was no significant difference between the grades of physics majors ($\mu = 1.72$) and medical students ($\mu = 1.34$), $W = 44,\ p = 0.30$.

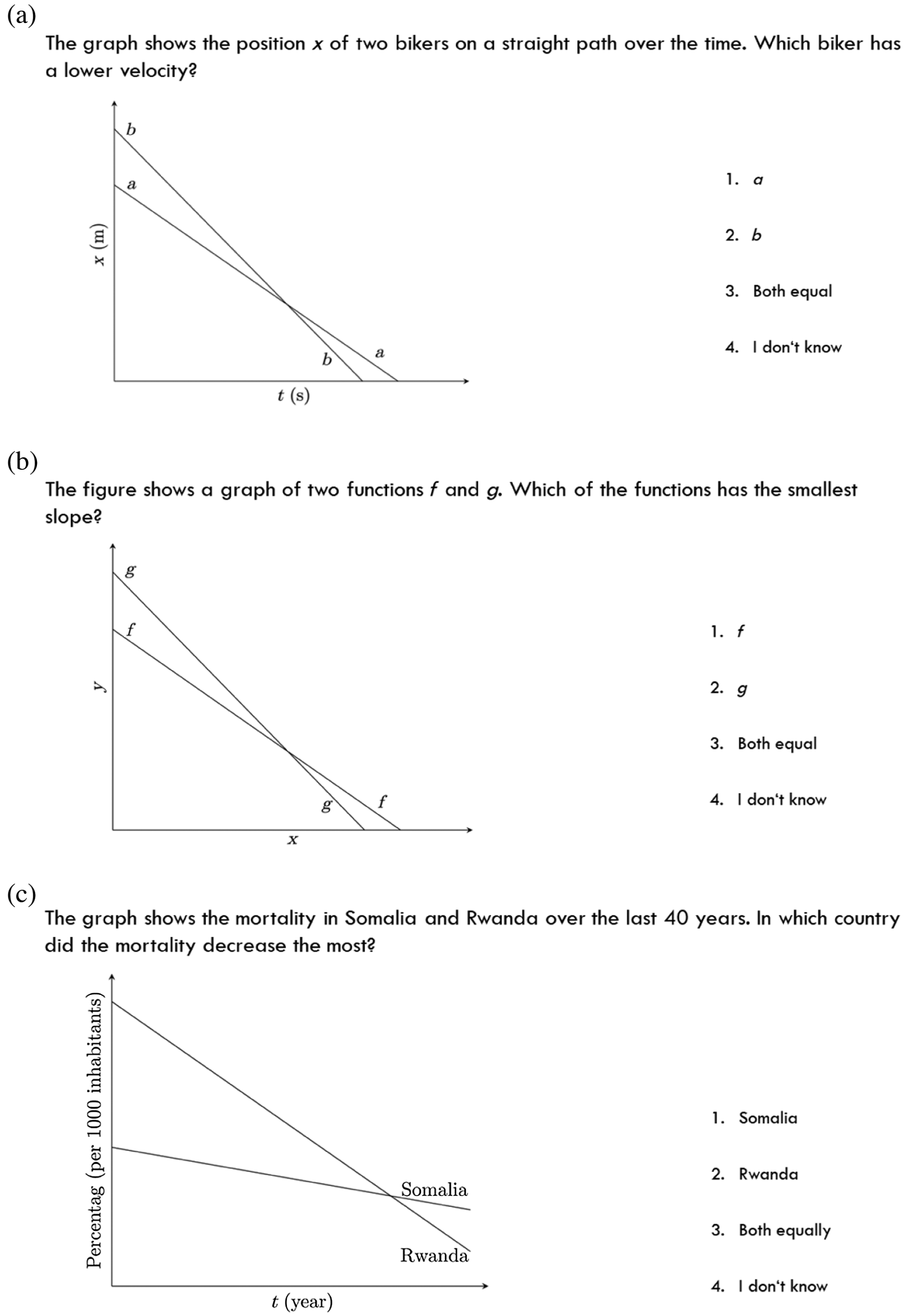

FIG. 1. Example math (a) and physics (b) items from Ceuppens *et al.* [39] as well as (c) the isomorphic medical item. The text is translated from the original German (Fig. 2).

B. Procedures and study design

The ethics committee of the mathematics, computer science, and statistics faculty found the pretest and the post-test to be ethically sound (EK-MIS-2022-122, EK-MIS-2023-143). Students from introductory physics and medicine courses at LMU Munich were recruited in the winter semester of the 2022/23 academic year. Participants answered the same questions at the beginning and end of the semester. The pretest took place in November 2022, while the post-test was conducted in February 2023.

Students signed a consent form and created a pseudonymized code they used to take the pretest and the post-test. The pretest also included a demographic questionnaire. Following this, the students attempted the given tasks.

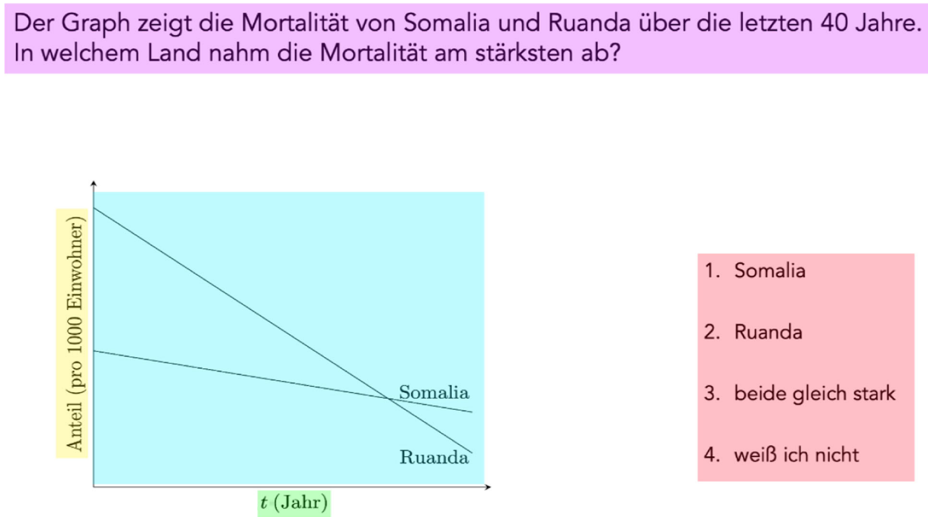

FIG. 2. The areas of interest (AOIs) on an exemplary test item. Selected AOIs include the text, answers, graph, labels, and axes (item presented in German, see Fig. 1(c) for the English translation).

C. Test material

The test material was based on instruments used in previous studies [8,39] and focused on assessing graph comprehension (Sec. II A). The test consisted of 24 items, composed of eight mathematics, physics, and medical items (*subjects*), respectively. Physics and medical items were posed in the context of their respective fields, while mathematical items were not presented in the form of a real-world scenario. Of the eight items in each of the three subjects, four were concerned with the area under the graph, while the other four were related to the slope of the graph (*concepts*). Example items are presented in Fig. 1. The complete test is available in Supplemental Material [57]. The participants' answers were rated as either correct (1) or incorrect (0).

D. Eye-tracking apparatus and measurements

Participants sat in front of a 24-inch computer screen with a resolution of 1920×1200 pixels and a refresh rate of 75 Hz. The distance to the screen was about 60 cm. A Tobii Nano eye tracker with a sampling frequency of 60 Hz and an ideal accuracy of 0.30° of the visual angle (according to the manufacturer) was used to collect eye movement data. The system allows for a high degree of movement; consequently, a chin rest was not necessary. More information on the eye tracker used in this study can be found in Tobii [58]. Fixations were detected using an I–VT algorithm [59].

Eye-tracking metrics were calculated based on specific stimulus regions that were identified as areas of interest (AOIs) [60]. We chose the areas around the text, the graph, and the answers as AOIs (Fig. 2). We also extracted the total viewing time (also known as "dwell time," see Ref. [6]) for these AOIs for comparison with the previous research [6–8].

E. Classification via machine learning

This study aimed to replicate the findings of Klein *et al.* [7]; i.e., it aimed to distinguish correct and incorrect solvers based on their eye movements—using machine-learning methods. In line with the hybrid modeling approach combining statistical and machine-learning methods [41], we used a two-step process: First, we conducted a statistical linear regression analysis, allowing us to account for the statistical influence of independent variables, such as study discipline and the type of AOI, on visual behavior (dependent variable). Additionally, possibly nonrelevant variables can be excluded to improve the efficiency of the machine-learning model [12,61]. All statistical analyses were conducted in R. The package "lm.beta" was used for regression analyses.

Second, this statistical analysis was extended using an analysis based on machine-learning methods to answer the second research question: Can differences in students' performance be accurately identified via machine-learning models trained on students' visual behavior? We aimed to replicate the results of Klein *et al.* [7] using machine-learning techniques and trained various machine-learning models to predict the participants' performance (independent variable) based on their total viewing time on the AOIs (dependent variable). We (a) compare several models (SVM, KNN, and RF; Sec. II C) to assess the models' performances and (b) compare a conventionally trained SVM with an optimized model [46]. We adopted a machine-learning method suited to small datasets that improves reliability and reduces bias [46]. We used fivefold cross-validation in conjunction with permutation tests and hyperparameter tuning to train the algorithm and evaluate its performance using the Matthews correlation coefficient (MCC). This method is particularly suited to small datasets [46]. We compared this optimized method with a simple

TABLE II. An overview of the results of the multiple linear regression used to assess the importance of the independent variables on viewing time, including study discipline (step 1), AOI (step 2), test date (pretest vs posttest; step 3), as well as the subject (mathematics vs physics vs medicine; step 4) and concept tested by each task (area vs slope; step 5), Note: [a]The adjusted $\boldsymbol{R^2}$ indicates how well the model explains the observed outcome. [b] SE B is the standard error of the unstandardized beta, which indicates whether the b-value differs significantly from 0. [c] β indicates how important a predicts is. For further information regarding these metrics, please refer to Ref. [47].

	Adjusted R^2[a]	SE B[b]	β[c]	p[d]
Step 1	0.0002			0.15
Constant		4.06 (0.11)		<0.001***
Physics		0.23 (0.16)	0.02	0.15
Step 2	0.2			<0.0001***
Constant		2.71 (0.19)		<0.001***
Physics		0.22 (0.14)	0.02	0.11
AOI axes		0.84 (0.24)	0.05	0.01**
AOI graph		5.79 (0.24)	0.33	<0.001***
AOI text		4,62 (0.24)	0.27	<0.001***
AOI x-label		−1.32 (0.24)	−0.08	<0.001***
AOI y-label		−1.82 (0.24)	−0.10	<0.001***
Step 3	0.2			<0.001***
Constant		2.17 (0.2)		<0.001***
Physics		0.18 (0.14)	0.01	0.2
AOI axes		0.84 (0.24)	0.05	0.01**
AOI graph		4.62 (0.24)	0.33	<0.001***
AOI text		5.79 (0.24)	0.27	<0.001***
AOI x-label		−1.32 (0.24)	−0.08	<0.001***
AOI y-label		−1.82 (0.24)	−0.10	<0.001***
Test pre		1.09 (0.14)	0.08	<0.001***
Step 4	0.23			<0.001***
Constant		0.8 (0.2)		<0.001***
Physics		0.18 (0.14)	0.14	0.2
AOI axes		0.84 (0.24)	0.04	<0.001***
AOI graph		4.62 (0.24)	0.33	<0.001***
AOI text		5.79 (0.24)	0.27	<0.001***
AOI x-label		−1.32 (0.24)	−0.08	<0.001***
AOI y-label		−1.82 (0.14)	−0.10	<0.001***
Test pre		1.09 (0.14)	0.08	<0.001***
Subject medicine		2.61 (0.17)	0.19	<0.001***
Subject physics		1.48 (0.17)	0.11	<0.001***
Step 5	0.23			<0.001***
Constant		1.04 (0.23)		<0.001***
Physics		0.18 (0.14)	0.14	0.19
AOI axes		0.84 (0.24)	0.05	<0.001***
AOI graph		4.62 (0.24)	0.33	<0.001***
AOI text		5.79 (0.24)	0.27	<0.001***
AOI x-label		−1.32 (0.24)	−0.08	<0.001***
AOI y-label		−1.82 (0.24)	−0.10	<0.001***
Test pre		1.09 (0.14)	0.08	<0.001***
Subject medicine		2.61 (0.17)	0.19	<0.001***
Subject physics		1.49 (0.17)	0.11	<0.001***
Concept slope		−0.5 (0.14)	−0.04	<0.001***

SVM trained on 60% of the dataset and tested on the remaining 40% of the data. This split was used to ensure that there was enough data to calculate feature importance. We used an SVM to predict performance based on the participants' total viewing time on specific AOIs because previous studies have shown that this algorithm can be successfully applied to similar tasks [52,53]. In the context of binary classification, SVMs attempt to find the ideal hyperplane that separates groups of data points by maximizing the margin between the data and the hyperplane [45]; consequently, SVMs can be considered to be a type of regression analysis. Due to its ability to transferability to unknown data and identify patterns in complex data (Table II), we believe that implementing an SVM would be an ideal way of comparing the results of machine-learning regression techniques to statistical regression. This will allow us to compare the variable (statistics) and feature importance (machine learning) of both types of analyses.

IV. RESULTS

Here, we present the results of the participants' answers and the analysis of their visual behavior. A p value of less than 0.05 was used as the threshold for a statistically significant test statistic [47]. The assumptions of all statistical tests, such as normal distribution, were checked and nonparametric tests were used if these assumptions were violated.

A. Graph comprehension gain

The participants answered the same items in both the pretest and the post-test. A Wilcoxon signed-rank test revealed that students took significantly longer to solve the pretest ($\mu = 16.8$ min, $\sigma = 5.1$) than the post-test ($\mu = 13.2$ min, $\sigma = 3.5$); $V = 40$, $p = 0.001$. 916 items were answered correctly, while 236 items were answered incorrectly. The graph comprehension gain was assessed by subtracting the percentage of correct answers in the pretest from the percentage of correct answers in the post-test. Across the entire sample, a dependent t test showed no significant difference in the graph comprehension gain of students on the area under the curve and the slope tasks; $t(45.98) = 0.42$, $p = 0.68$. A Wilcoxon signed-rank test revealed that physics majors exhibited improved graph comprehension gain with respect to the tasks associated with the area under the curve ($\mu = 0.11$, $\sigma = 0.18$) compared to tasks associated with the slope of the graph ($\mu = -0.02$, $\sigma = 0.19$); $W = 166663020$, $p < 0.001$. In contrast, the opposite was observed for (veterinary) medicine students, who exhibited improved performance on the slope tasks ($\mu = 0.05$, $\sigma = 0.2$) compared to the tasks associated with the area under the curve ($\mu = -0.01$, $\sigma = 0.24$); $W = 81328200$, $p < 0.001$. As there were no overall differences across the sample, and because the improvements in the performance of the students on different types of graph comprehension tasks were in opposition, we used the mean graph comprehension gain between the slope and area tasks for the remaining analyses. The mean graph comprehension gain for physics and (veterinary) medicine students across the three subjects tested is presented in Fig. 3.

The graph comprehension gain of physics and medical students across the three different subjects assessed by the test material (mathematics, physics, and medicine) was compared via a 2 (discipline) × 3 (subject) ANOVA. Graph comprehension gain did not differ between physics ($\mu = 0.05$, $\sigma = 0.14$) and medical students ($\mu = 0.05$, $\sigma = 0.19$; $F(1,48) = 0.01$, $p = 0.9$, $\omega^2 = -.02$). However, graph comprehension gain did differ significantly between subjects; $F(2,48) = 4.11$, $p = 0.02^*$, $\omega^2 = 0.11$; this can be considered a medium effect [62]. Bonferroni *post hoc* tests showed that the graph comprehension gain in medical items ($\mu = -0.02$, $\sigma = 0.19$) was significantly smaller than the graph comprehension gain in physics items ($\mu = 0.13$, $\sigma = 0.16$).

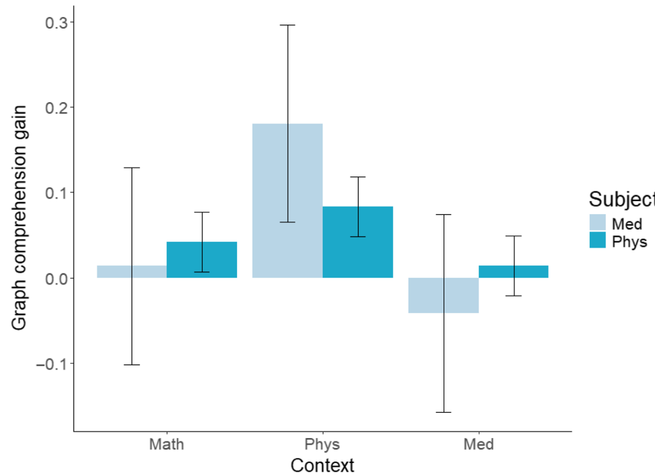

FIG. 3. Mean graph comprehension gain for students from each discipline across the different subjects. Error bars indicate the standard deviation of the results.

B. Total viewing time

This section presents the results of the viewing times analyses conducted on the eye-tracking data. The effects of the independent variables on the viewing time were first assessed using linear regression. This revealed several independent variables that should be considered when designing the machine-learning model, specifically viewing times on AOIs and the test date. Based on these results, an SVM was trained and the results were compared with optimized and non-optimized machine-learning algorithms as described in Section III E [46]. We also investigated the feature importance of the final SVM model.

1. Statistical analysis

We used multiple linear regression to determine the effects of the following independent variables on the total viewing time (dependent variable): the students' discipline, parameters associated with specific AOIs (Fig. 2), the test date, the subject of each task (Fig. 1), and the concept tested. In this way, the effects of all possible influences on visual behavior were evaluated. The results of linear regression models are presented in Table II. An ANOVA revealed that each step significantly improved the fit; $p < 0.001$. The first model—which only included the participants' discipline to predict their viewing time—did not perform well, adjusted $R^2 = 0.0002$. The model's performance was significantly increased by adding the AOI as a predictor variable, adjusted $R^2 = 0.02$. Adding the test date as a variable did not increase performance, adjusted $R^2 = 0.02$. However, adding the subject assessed by each item improved the model, adjusted $R^2 = 0.023$. Adding the specific graph comprehension concept tested did not improve the model. These results suggest that 23% of the variance in viewing time can be explained by this linear regression model, which includes the viewing time associated with each AOI and the subject assessed by the task.

2. Machine-learning analysis

The insights gained from the multiple linear regression were used to guide the design of the machine-learning algorithm by looking at the influence of the individual variables. Participants' performance was predicted using the viewing time on each AOI for every item because this seemed to be the most relevant variable based on the linear regression results. Reducing the number of features can also increase the efficiency of the machine-learning model [12,61]. Separate models were built for the pretest and post-test datasets, as this was a significant variable in the regression analysis. This study aimed to replicate the results of Klein *et al.* [7] via machine-learning methods. Specifically, we used a machine-learning method tailored to small datasets to train an SVM; the final model was evaluated using MCC to judge its performance on unbalanced datasets and to reduce bias [46]. The results revealed that correct and incorrect solvers could be distinguished based on the AOIs, although the MCC values of the final model were low (pretest results: MCC $= 0.23$, $p = 0.001$; post-test results: MCC $= 0.22$, $p = 0.001$). The performance of the SVM was compared to a KNN and an RF model; these models had slightly lower MCC scores (KNN pretest results: MCC $= 0.08$, $p = 0.02$; KNN post-test results: MCC $= 0.13$, $p = 0.01$; rf pretest results: MCC $= 0.14$, $p = 0.01$; rf post-test results: MCC $= 0.18$, $p = 0.01$).

It is important to note that the method used to generate these models (Sec. III E) was designed specifically for small datasets and is not a conventional practice. Analyses using classical train-test-split methods led to different results. Using a simple SVM without cross validation with a 60–40 train-test split, produced a model that scored well on a variety of performance metrics (pretest metrics: accuracy $= 0.78$, F1 $= 0.87$.; post-test metrics: accuracy $= 0.82$, F1 $= 0.90$). However, the MCC of the model was 0.0 in both cases, which suggests that the method developed for small datasets led to more reliable results.

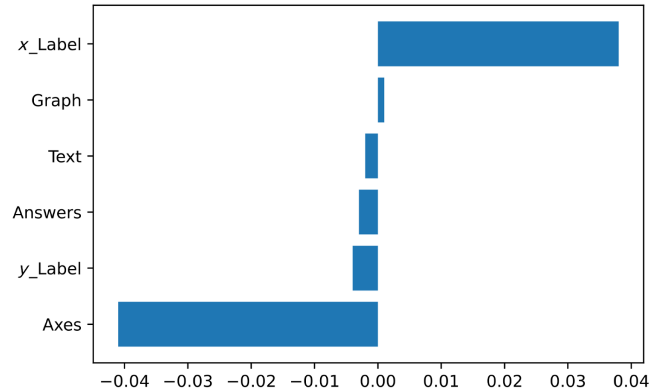

FIG. 4. Feature importance coefficients derived from the simple SVM.

The feature importance of the simple SVM instantiated with a linear kernel supported these results. Feature importance coefficients suggested that the x-label and the axes were the most important features when attempting to distinguish correct from incorrect solvers based on their visual behavior (Fig. 4). It is important to note that these coefficients were calculated on the training data.

Permutation feature importance uses the test data to calculate feature importance; this technique randomly shuffles the features, which generally causes the model to perform worse. A feature's relevance is determined by how much worse the model performs on the test set when the feature is randomly permutated. Permutation feature importance for the simple SVM revealed that the viewing time on the axes was the most important feature in distinguishing between correct and incorrect solvers, followed by the viewing time spent on the graph (Fig. 5). Notably, the feature importance coefficient regarding viewing time spent on the AOIs associated with the axes of the graph changed from negative to positive when permutation feature importance was considered; in addition, the high x-label coefficient observed in the standard feature importance was not observed when permutation feature importance analysis was conducted. These results suggest that the simple SVM was not optimally trained, although the prediction's accuracy appeared to be high.

V. DISCUSSION

This section discusses the first-semester physics majors and (veterinary) medicine students. We investigated graph comprehension gain and differences in visual behavior for students faced with math, physics, or medical graph comprehension tasks. We first compare the task performance of the physics and (veterinary) medicine students for each of the three subjects (math, physics, and medicine). We then evaluate the participants' visual behavior based on their total viewing time on each AOI (Fig. 2).

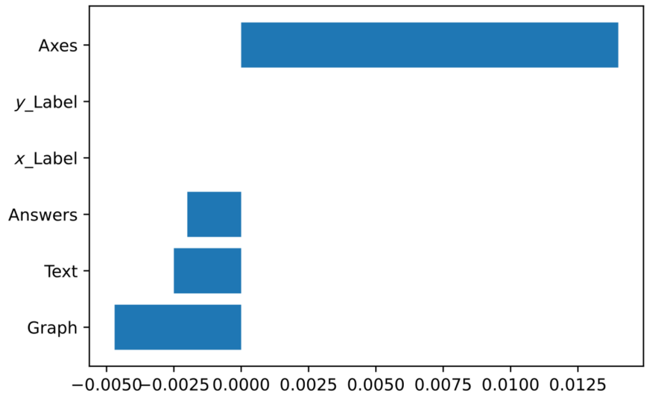

FIG. 5. Permutation feature importance of the simple SVM.

A. Comparison of graph comprehension gain between physics and medical students

Previous research found that physics students performed better at graph comprehension tasks than psychology [8] and economics students [6,7]. In this study, we compared physics with (veterinary) medicine students and found no significant differences in graph comprehension gain over the first semester. These results suggest that first-semester physics and medicine students have similar graph comprehension skills when faced with problems associated with mathematics, physics, and medicine.

The participants were tasked with solving isomorphic graph comprehension problems across three subjects and used the same visual behavior to solve each question. Participants' equal performance across different subjects suggests that all participants—i.e., both physics majors and (veterinary) medicine students—were able to apply the solution strategy regardless of subject. This also suggests that students were able to identify the relevant variables in all graphs regardless of subject and were able to relate the data presented in each graph to the subject of the question being asked. One explanation for this observation is that both physics majors and (veterinary) medicine students possessed a similar level of representational competence: Both physics and medicine students appeared to exhibit a comparable ability in terms of the use of problem-solving strategies in isomorphic representations across different subjects (i.e., familiarity with graphing conventions and connecting the representations to the concepts).

Another reason for the comparable performance between the physics majors and (veterinary) medicine students in this study compared to the social science students assessed in previous studies [6–8] could be the access to STEM courses: social sciences students typically only have one STEM course during the first semester (statistics). In contrast, all students participating in this study attended more than one STEM course. STEM courses provide students with a real-world context for the abstract mathematical concepts taught in their respective disciplines; this may occur concurrently with learning about their application through experimentation themselves, such as in physics labs. Thus, the description of phenomena in abstract mathematical terms could potentially enhance a student's ability to transfer these problem-solving skills to other subjects.

Each item was relatively simple (see examples in Fig. 1); some tasks were intended for 9th-grade students [39]—i.e., first-semester university students would be expected to solve them without much difficulty. This was reflected in the high degree of accuracy across all subjects and disciplines: 80% of the items were solved correctly. Another potential reason for this similar performance among the participants could be their relatively equal skill level as shown by comparable A-level grades of students from both disciplines. Furthermore, both physics majors and (veterinary) medicine students voluntarily chose disciplines that require a high degree of graph comprehension skills: the study of physics generally involves graphs containing experimental data, while graphs that depict patient data are common in medicine. Physics and (veterinary) medicine students would presumably also be familiar with the STEM courses that are taken as part of their degree and can be assumed to be proficient STEM learners. For example, physics majors are believed to have an aptitude for science [63]. Finally, students did not appear to improve between the pretest and post-test, which could indicate that they possessed an overall high level of graph comprehension skills even before starting their university studies.

In contrast to previous studies [6–8], we did not find any significant differences in graph comprehension between physics majors and medical students. This suggests that students who take STEM courses have a similar level of graph comprehension skills. Although we have assumed that STEM courses play a role in facilitating graph comprehension, we cannot conclude that STEM courses are the sole *reason* for the participants' comparable graph comprehension skills: Other factors, such as aptitude and self-selection of the university discipline, may also play a role. However, we did find that correct solvers used different visual processes to solve the graph comprehension tasks compared to incorrect solvers. This suggests that some students are more successful in identifying the relevant variables (correct solvers) compared to others (incorrect solvers), suggesting different levels of representational competence.

B. Viewing times of correct and incorrect solvers

The second research question posed in this study concerns the total viewing time on specific AOIs for correct and incorrect solvers: Can differences in students' performance be accurately identified via machine-learning models trained on students' visual behavior? We addressed this question in two ways (Sec. III E): We first conducted a statistical analysis to determine the relative influence of each independent variable on the dependent variable viewing time. We then used the total viewing time on each AOI to predict performance. Finally, we compared several machine-learning models, which included an SVM trained using a method developed specifically for small datasets [46]. An SVM, a type of machine-learning-based regression, ensures that the models' feature importance is comparable to the variable importance obtained from a statistical regression.

Multiple linear regression revealed that the total viewing time did not depend on the participants' discipline, instead, it was most strongly influenced by the different AOIs (Table II). Specifically, longer viewing times were associated with increased time spent looking at the graph and the text and reduced time spent looking at the axes labels. The model's accuracy increased when the subject tested by the item was added as a variable (from adjusted $R^2 = 0.2$ to adjusted $R^2 = 0.23$; Table II). In particular, tasks that posed problems related to medicine and physics appeared to influence the viewing time; medical items required slightly longer viewing compared to physics items. These results were consistent with our overall observations on graph comprehension gain: medical tasks appeared to be more difficult than physics tasks (Fig. 3). Consequently, participants are likely to spend more time on tasks with greater difficulty. Participants' graph comprehension gain did not differ between physics, medicine, and mathematics (Sec. IV A).

This study also aimed to replicate previous findings regarding the differences in the visual behavior of correct and incorrect solvers [7] with machine-learning methods. The SVM model performed better than the KNN and rf models. We also found significant differences between correct and incorrect solvers in the pretest and the post-test. This suggests that machine-learning algorithms can distinguish between participants who either correctly or incorrectly approached a graphical comprehension task. These results are consistent with previous research, suggesting that the total viewing time spent on concept-specific areas can differ between correct and incorrect solvers [7,30]. Although the performance of the optimized machine-learning algorithm was not as good as the models presented in previous research, the results remained satisfactory. For example, Dzsotjan *et al.* [49] reported an F1 score of 0.66 when using a combination of the best features, which included features other than the viewing time on AOIs. It should be noted that an F1 score is also more likely to support the null hypothesis than an MCC score [46]. This is supported by the high accuracy (pretest: 0.78, posttest: 0.82) and F1 scores (pretest: 0.87, posttest: 0.90) obtained when using a simple SVM to predict participant performance. These results would suggest that a simple SVM might be a good model for the data; however, the MCC of the SVM model was 0.00 for both pretest and post-test datasets, indicating that this was not the case. In contrast, the MCC for the optimized model, which used cross-validation techniques and permutation tests, was much higher, suggesting that it is better suited to predict performance based on eye-tracking data.

It is important to note that the feature importance of the simple SVM model changed depending on how it was calculated. The coefficients used by the SVM for the training data suggested that the viewing time on the x-label and the axes were the most important features (Fig. 4). However, when feature importance was calculated using permutation tests on the test data, the second most relevant feature was the viewing time spent on the graph (though the axes were the most relevant features; Fig. 5). These observations were more consistent with the results of the multiple linear regression, which suggested that the viewing time spent on the graph was the most important variable. Therefore, we confirmed that machine learning seems to be a valid method to predict students' performance based on their visual behavior [12,49,50]; particularly, with an optimized machine-learning method developed for small datasets [46], which yields results comparable to classical statistical methods. The accuracy of the feature importance tests depends on the size of the test set; this study used a test set size of 40% to ensure there was enough data to calculate feature importance. As the feature importances derived from the simple SVM were similar but not identical to the results of multiple linear regression analysis, the size of the dataset appears to be too small for simple machine-learning algorithms. This highlights the importance of using an appropriate method to analyze small datasets.

C. Implications for practice

This study investigated the differences in graph comprehension skills between physics majors and (veterinary) medicine students and found that students from the two disciplines had relatively similar graph comprehension skills. This was in contrast to prior research that reported that STEM majors tended to outperform non-STEM social science students who traditionally do not take STEM courses as part of their curriculum [6–8]. The commonalities of the disciplines in our study and the differences in the results between our study and previous research suggest that students might benefit from taking courses that involve the graphical representation of data associated with specific real-world concepts [21] as is often the case in STEM courses. Nevertheless, there are likely to be other factors that influence students' graph comprehension skills, such as personal interest.

This study replicated previous studies [7] distinguishing correct and incorrect solvers based on their eye movements using machine-learning methods. This suggests that machine learning has the potential to be successfully applied as a tool in STEM education research, especially in the context of eye movement analyses. A comparison with simple algorithms suggests that the method proposed by Steinert *et al.* [46] is appropriate for analyzing small datasets, such as those used in this study. These results suggest that eye-tracking measurements can be used as inputs for adaptive learning systems [9]. Applications that use eye-tracking metrics will likely become increasingly relevant for teachers, especially in conjunction with personalized systems that can support individual learners. For example, machine-learning applications could help teachers identify students encountering problems with the course content.

D. Limitations and future research

There are several limitations to this research. The most relevant limitation is the low number of participants. Another limitation is that the medical items used in the pretest and post-test had not been previously tested, which may affect the comparability with the isomorphic items associated with mathematics and physics. Furthermore, other factors that may influence graph comprehension skills—such as self-selection of the study discipline or the number of STEM courses that each student took—were not addressed in our research questions. Future research should extend this research into other potential variables that could influence the development of graph comprehension skills.

Future research should also validate the medical test items used in this study to assess their comparability with test instruments used in previous research. Comparisons with other STEM disciplines, such as mathematics or computer science students, could also provide further insights as to why students of some disciplines outperform others in graph comprehension tasks.

While SVMs appear to be well suited to making predictions about correct and incorrect solvers, this study indicated a relatively low MCC score; future research may wish to improve the model presented in this study by testing other types of machine-learning algorithms, such as neural networks. The model performance may also improve by adding more features.

Finally, future research should consider the real-world application of eye-tracking data in adaptive learning systems and their ability to support students during the learning process. These systems should ideally be accessible and capable of operating in real time. Indeed, SVMs can be used in real time [55,56] and web-based eye tracking can now be easily implemented on certain websites [10]. These tools would provide students with individual support while also providing teachers with immediate information on whether students have issues with a specific type of problem or whether a particular student needs more assistance.

VI. CONCLUSION

Graph comprehension is an important skill for students of various disciplines. Physics majors appear to have better graph comprehension skills compared to students from social sciences disciplines. This study compared the graph comprehension skills of physics and medical students across three different subjects: math, physics, and medicine. Each participant attempted these tasks at the beginning and the end of the first semester. There were no significant differences in graph comprehension gain or visual behavior between physics majors and medical students, suggesting that students from these disciplines had similar graph comprehension skills. One reason for this finding could be that physics majors and medicine students often engage with graphs due to the multitude of STEM courses in their respective disciplines.

An investigation of the visual processing of correct and incorrect solvers across both disciplines using a novel machine-learning approach tailored to small datasets revealed that eye movements could be used as a predictor for performance. This demonstrates that machine-learning analyses can be a valuable tool for education research even for small datasets, as long as this is accounted for during the analysis.

V. R., Y. D., S. K., M. B., D. K., J. B., J. S., M. F., and J. K. conceptualized the work; V. R., Y. D., S. K., and J. K. determined the methodology; V. R. conducted the formal analysis; S. S. developed the machine-learning algorithm tailored to small datasets; V. R. and Y. D. conducted the investigation; V. R., Y. D., S. K., M. B., D. K., J. B., J. S., M. F., and J. K. validated the data; V. R. curated the data; V. R. prepared the original draft; V. R., Y. D., S. K., M. B., S. S., D. K., J. B., J. S., M. F., and J. K. reviewed and edited the manuscript; V. R. provided the visualizations; S. K. and J. K. supervised the project; J. K. dealt with the project administration.

[1] R. Mayer and L. Fiorella, *The Cambridge Handbook of Multimedia*, 3rd ed. (Cambridge University Press, Cambridge, England, 2012), 10.1017/9781108894333.

[2] Program for international student assessment (PISA). PISA 2022 mathematics framework, https://pisa2022-maths.oecd.org/ (2022).

[3] L. Boels, A. Bakker, W. Van Dooren, and P. Drijvers, Conceptual difficulties when interpreting histograms: A review, Educ. Res. Rev. **28**, 100291 (2019).

[4] N. Glazer, Challenges with graph interpretation: a review of the literature, Stud. Sci. Educ. **47**, 183 (2011).

[5] G. Leinhardt, O. Zaslavsky, and M. K. Stein, Functions, graphs, and graphing: Tasks, learning, and teaching, Rev. Educ. Res. **60**, 1 (1990).

[6] S. Brückner, O. Zlatkin-Troitschanskaia, S. Küchemann, P. Klein, and J. Kuhn, Changes in students' understanding of and visual attention on digitally represented graphs across two domains in higher education: A postreplication study, Front. Psychol. **11**, 1 (2020).

[7] P. Klein, S. Küchemann, S. Brückner, O. Zlatkin-Troitschanskaia, and J. Kuhn, Student understanding of graph slope and area under a curve: A replication study comparing first-year physics and economics students, Phys. Rev. Phys. Educ. Res. **15**, 020116 (2019).

[8] A. Susac, A. Bubic, E. Kazotti, M. Planinic, and M. Palmovic, Student understanding of graph slope and area under a graph: A comparison of physics and nonphysics students, Phys. Rev. Phys. Educ. Res. **14**, 020109 (2018).

[9] K. Kennel, Using eye tracking data for enhancing adaptive learning systems, in *Proceedings of the 2022 Symposium on Eye Tracking Research and Applications, ETRA 2022* (2022), Vol. 34, pp. 1–3, 10.1145/3517031.3532195.

[10] X. Yang and I. Krajbich, Webcam-based online eye-tracking for behavioral research, Judgm. Decis. Mak. **16**, 1485 (2021).

[11] M. Sailer, E. Bauer, R. Hofmann, J. Kiesewetter, J. Glas, I. Gurevych, and F. Fischer, Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning, Learn Instr. **83**, 101620 (2023).

[12] S. Küchemann, P. Klein, S. Becker, N. Kumari, and J. Kuhn, Classification of students' conceptual understanding in STEM education using their visual attention distributions: A comparison of three machine-learning approaches, in *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU)* edited by H. C. Lane, S. Zvacek, and J. Uhomoibhi (SciTePress, Portugal, 2020), Vol. 1, pp. 36–46, 10.5220/0009359400360046.

[13] L. Boels, E. G. Moreno-Esteva, A. Bakker, and P. Drijvers, Automated gaze-based identification of students' strategies in histogram tasks through an interpretable mathematical model and a machine learning algorithm, Int. J. Artif. Intell. Educ. **34**, 931 (2023).

[14] A. Susac, A. Bubic, P. Martinjak, and M. Palmovic, Graphical representations of data improve student understanding of measurement and uncertainty: An eye-tracking study, Phys. Rev. Phys. Educ. Res. **13**, 020125 (2017).

[15] P. Shah and J. Hoeffner, Review of graph comprehension research: Implications for instruction, Educ. Psychol. Rev. **14**, 47 (2002).

[16] S. Friel, F. R. Curcio, and G. W. Bright, Making sense of graphs: Critical factors influencing comprehension and instructional implications, J. Res. Math. Educ. **32**, 124 (2001).

[17] E. G. Freedman and P. Shah, Toward a model of knowledge-based graph comprehension, in *Diagrammatic Representation and Inference. Diagrams 2002*, edited by M. Hegarty, B. Meyer, and N. H. Narayanan, *Lecture Notes in Computer Science Vol. 2317* (Springer, Berlin, Heidelberg, 2002), pp. 18–30, 10.1007/3-540-46037-3_3.

[18] E. Harrison, B. McBride, and M. C. Linn, Revision analysis of students' position-time graphs, CSCL **2** (2019), https://par.nsf.gov/biblio/10181648-revision-analysis-students-position-time-graphs.

[19] C. F. J. Pols, P. J. J. M. Dekkers, and M. J. de Vries, What do they know? Investigating students' ability to analyse experimental data in secondary physics education, Int. J. Sci. Educ. **43**, 274 (2021).

[20] Y. Okan, M. Galesic, and R. Garcia-Retarmero, How people with low and high graph literacy process health graphs: Evidence from eye-tracking, J. Behav. Decis. Making **29**, 271 (2016).

[21] J. Woolnough, How do students learn to apply their mathematical knowledge to interpret graphs in physics? Res. Sci. Educ. **30**, 259 (2000).

[22] M. Rau, Conditions for the effectiveness of multiple visual representations in enhancing STEM learning, Educ. Psychol. Rev. **29**, 717 (2017).

[23] A. A. diSessa and B. L. Sherin, Meta-representation: an introduction, J. Math. Behav. **19**, 385 (2000).

[24] A. Angra and S. M. Gardner, Development of a framework for graph choice and construction, Adv. Physiol. Educ. **40**, 123 (2016).

[25] D. Carraher and A. Schliemann, The transfer dilemma, J. Learn. Sci. **11**, 1 (2002).

[26] E. Alemdag and K. Cagiltay, A systematic review of eye tracking research on multimedia learning, Comput. Educ. **125**, 413 (2018).

[27] M. L. Lai, M. J. Tsai, F. Y. Yang, C. Y. Hsu, T. C. Liu, S. W. Y. Lee, M. H. Lee, G. L. Chiou, J. C. Liang, and C. C. Tsai, A review of using eye-tracking technology in exploring learning from 2000 to 2012, Educ. Res. Rev **10**, 90 (2013).

[28] M. A. Just and P. A. Carpenter, A theory of reading: From eye fixations to comprehension, Psychol. Rev. **87**, 329 (1980).

[29] L. Hahn and P. Klein, Eye tracking in physics education research: A systematic literature review, Phys. Rev. Phys. Educ. Res. **18**, 013102 (2022).

[30] V. Ruf, A. Horrer, M. Berndt, S. I. Hofer, F. Fischer, M. R. Fischer, J. M. Zottmann, J. Kuhn, and S. Küchemann, A literature review comparing experts' and non-experts' visual processing of graphs during problem-solving and learning, Educ. Sci. **13**, 1 (2023).

[31] K. A. Ericsson and W. Kintsch, Long-term working memory, Psychol. Rev. **102**, 211 (1995).

[32] H. L. Kundel, C. F. Nodine, E. F. Conant, and S. P. Weinstein, Holistic component of image perception in mammogram interpretation: Gaze-tracking study, Radiology **242**, 396 (2007).

[33] H. Haider and P. A. Frensch, Eye movement during skill acquisition: More evidence for the information-reduction hypothesis, J. Exp. Psychol. **25**, 172 (1999).

[34] A. Maries and C. Singh, Helping students become proficient problem solvers part I: A brief review, Educ. Sci. **13**, 156 (2023).

[35] A. Gegenfurtner, E. Lehtinen, and R. Säljö, Expertise differences in the comprehension of visualizations: a Meta-analysis of eye-tracking research in professional domains, Educ. Psychol. Rev. **23**, 523 (2011).

[36] S. Brams, G. Ziv, O. Levin, J. Spitz, J. Wagemans, A. M. Williams, and W. F. Helsen, The relationship between gaze behavior, expertise, and performance: A systematic review, Psychol. Bull. **145**, 980 (2019).

[37] J. A. Harsh, M. Campillo, C. Murray, C. Myers, J. Nguyen, and A. v. Maltese, "Seeing" data like an expert: An eye-tracking study using graphical data representations, CBE Life Sci. Educ. **18**, ar32 (2019).

[38] S. Becker, L. Knippertz, S. Ruzika, and J. Kuhn, Persistence, context, and visual strategy of graph understanding: Gaze patterns reveal student difficulties in interpreting graphs, Phys. Rev. Phys. Educ. Res. **19**, 020142 (2023).

[39] S. Ceuppens, L. Bollen, J. Deprez, W. Dehaene, and M. De Cock, 9th grade students' understanding and strategies when solving $x(t)$ problems in 1D kinematics and $y(x)$ problems in mathematics, Phys. Rev. Phys. Educ. Res. **15**, 010101 (2019).

[40] S. Becker, S. Küchemann, P. Klein, A. Lichtenberger, and J. Kuhn, Gaze patterns enhance response prediction: More than correct or incorrect, Phys. Rev. Phys. Educ. Res. **18**, 020107 (2022).

[41] A. Daoud and D. Dubhashi, Statistical modeling: the three cultures, arXiv.2012.04570.

[42] A. Namoun and A. Alshanqiti, Predicting student performance using data mining and learning analytics techniques: A systematic literature review, Appl. Sci. **11**, 237 (2020).

[43] T. Kabudi, I. Pappas, and D. H. Olsen, AI-enabled adaptive learning systems: A systematic mapping of the literature, Comput. Educ. **2**, 100017 (2021).

[44] L. Breiman, Statistical modeling: The two cultures, Stat. Sci. **16**, 199 (2001).

[45] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. (O'Reilly Media Inc., Canada, 2019).

[46] S. Steinert, V. Ruf, D. Dzsotjan, N. Großmann, A. Schmidt, J. Kuhn, and S. Küchemann, A refined approach for evaluating small datasets via binary classification using machine learning, PLoS One **19**, e0301276 (2024).

[47] A. Field, J. Miles, and Z. Field, *Discovering Statistics Using R* (SAGE Publications, 2012).

[48] W. Xu and F. Ouyang, The application of AI technologies in STEM education: A systematic review from 2011 to 2021, Int. J. STEM Educ. **9**, 59 (2022).

[49] D. Dzsotjan, K. Ludwig-Petsch, S. Mukhametov, S. Ishimaru, S. Küchemann, and J. Kuhn, 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* (ACM, New York, 2021), pp. 467–471, 10.1145/3460418.3479358.

[50] S. Küchemann, S. Becker, P. Klein, and J. Kuhn, Gaze-based prediction of students' understanding of physics line-graphs: An eye-tracking-data based machine-learning approach, in *Computer Supported Education*, edited by H. C. Lane, S. Zvacek, and J. Uhomoibhi (Springer International Publishing, 2021), pp. 450–467, 10.1007/978-3-030-86439-2_23.

[51] R. J. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).

[52] S. Mozaffari *et al.*, Quantifying gaze-nased strategic patterns in physics vector field divergence, in *ICAART 2020. Lecture Notes in Computer Science*, edited by A. P. Rocha, L. Steels, and J. van den Herik (Springer, 2021), pp. 465–481, 10.1007/978-3-030-71158-0_22.

[53] N. S. Rebello *et al.*, Machine learning predicts responses to conceptual tasks using eye movements, presented at PER Conf. (2018), https://www3.cs.stonybrook.edu/~minhhoai/papers/PERC2018_Rebello.pdf.

[54] V. Bachurina *et al.*, A machine learning investigation of factors that contribute to predicting cognitive performance: Difficulty level, reaction time and eye-movements, Decis. Support Syst. **155**, 113713 (2022).

[55] D. Ifenthaler and C. Widanapathirana, Development and validation of a learning analytics framework: Two case studies using support vector machines, Tech. Know. Learn. **19**, 221 (2014).

[56] C. Verma, V. Stoffová, and Z. Illés, Prediction of students' awareness level towards ICT and mobile technology in Indian and Hungarian University for the real-time: preliminary results, Heliyon **5**, e01806 (2019).

[57] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.20.020138 for the original test items in German.

[58] Tobii. Tobii Pro Nano, https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-nano#specifications (2022).

[59] D. D. Salvucci and J. H. Goldberg, Identifying fixations and saccades in eye-tracking protocols, in *Proceedings of the Symposium on Eye Tracking Research & Applications —ETRA '00*, (ACM, New York, 2000), pp. 71–78, 10.1145/355017.355028.

[60] K. Holmqvist and R. Andersson, *Eye Tracking: A Comprehensive Guide to Methods, Paradigms and Measures* (Lund Eye–Tracking Research Institute, Lund, Sweden, 2017).

[61] W. Jia, M. Sun, J. Lian, and S. Hou, Feature dimensionality reduction: A review, Complex Intell. Syst. **8**, 2663 (2022).

[62] R. E. Kirk, Practical significance: A concept whose time has come, Educ Psychol Meas **56**, 746 (1996).

[63] B. E. Woolnough, Why students choose physics, or reject it, Phys. Educ. **29**, 368 (1994).