Research in Engineering Design (2025) 36:3
https://doi.org/10.1007/s00163-025-00444-2

ORIGINAL PAPER

Towards an automated evaluation of design processes—an algorithm to predict critical situations during concept synthesis

Christoph Zimmerer[1] · **Christoph Wittig**[1] · **Sven Matthiesen**[1]

Received: 27 June 2024 / Accepted: 20 January 2025

Abstract
In design research, there is a need for research methods that allow for larger numbers of participants in empirical studies, as small sample sizes lead to less statistically reliable results. The number of participants is limited by current research methods, such as protocol analysis, interviews, or analysis of eye tracking videos, because they require a lot of manual work during the evaluation. This is particularly noticeable in studies that analyze the cognitive processes of designers, e.g., during concept synthesis. Therefore, the goal of this paper is to develop an algorithm that automates the analysis of eye tracking data to predict designer perceived difficulties—critical situations in which methodical support could be beneficial—during design processes. By linking eye tracking data with retrospective think aloud data, a dataset was created for training different machine learning algorithms. The dataset was further processed, and three algorithms were evaluated regarding their suitability for automated detection of difficulties. The best algorithm, a cascading one-against-all classifier, achieved an accuracy of 62% and a false-negative rate of 26.6%. Depending on the type of difficulty, different eye tracking features were relevant for the decisions of the algorithms, highlighting the importance of tailored feature selection for each type of difficulty. The findings suggest that the automated analysis of eye tracking data using machine learning potentially facilitates larger studies and statistically more reliable findings, representing a significant step toward more efficient and insightful analysis of design cognition.

Keywords Conceptual design · Design research · Eye tracking · Data-driven evaluation · Machine learning

1 Introduction

Design research is a complex and multifaceted field, in which understanding the processes involved in design and innovation is crucial. Researchers in this field face the challenge of obtaining meaningful and comprehensive data while facing constraints such as difficult access to participants and resource-intensive methods of analysis like observations and interviews (Escudero-Mancebo et al. 2023). This leads to a tendency for studies to be conducted more frequently with small numbers of participants and focus on detailed data analysis (Hay et al. 2020). However, the effort required for data collection, transcription and analysis is extremely time-consuming, with each hour of observation requiring up to 25 additional hours of work in data processing and analysis, as described by Ahmed et al. (2003). This resource-intensive process is a major obstacle to conduct larger studies that aim to achieve statistically meaningful results in design research.

Dinar et al. (2015) emphasizes the resource-intensive nature of data analysis, the small sample size in empirical studies and the lack of statistically significant and reliable results. To address these obstacles, Dinar et al. (2015) suggests that future studies could benefit from the development of computer-based data collection and automated analysis techniques. Recognizing the need for new research methods to improve resource-intensive data processing and analysis in empirical studies with designers, recent work has highlighted the potential of physiologic signals (Dinar et al. 2015; Gero and Milovanovic 2020). Physiologic signals such as electromyography (EMG), electrocardiography (ECG) and eye tracking (ET) have emerged as promising tools for investigating design-relevant research questions (Gero and Milovanovic 2020). These measurement methods are characterized by the fact that they generate quantitative

✉ Christoph Zimmerer
christoph.zimmerer@kit.edu

[1] Karlsruhe Institute of Technology (KIT), IPEK - Institute of Product Engineering, Kaiserstr. 10, 76131 Karlsruhe, Germany

Published online: 20 February 2025

data without influencing the participants as much as traditional survey methods such as concurrent think aloud. Due to its ability to capture the gaze behavior of designers, eye tracking in particular has considerable potential here, as the designer's eyes are the sensors through which information is recorded (Ullman 2010). In the field of design research, eye tracking studies have so far mainly been used to obtain information about what contextual information designers look at. For example, Matthiesen et al. (2013) have investigated how designers proceed when analyzing technical systems. However, the measurement of creativity, for example, based on individual metrics has also been successfully carried out (Sun et al. 2014).

The potential applications of these measurement methods are best illustrated by other areas of research. In the field of human-computer interaction (Zagermann et al. 2018), in driving simulators (Benedetto et al. 2011) or diagram understanding (Maier et al. 2014), eye tracking is used for measuring cognitive load. High cognitive load in turn is considered an indicator of critical situations of declining performance (Bruggen 2015) and therefore the need for support. Following this, the integration of physiologic signals into design research could open up new possibilities for recording and evaluating cognitive load during the design process. A decrease of performance in the processing of complex cognitive tasks occurs particularly when the cognitive load of the participants exceeds a critical limit (Paas and van Merriënboer 1994). Detecting cognitive load provides an important insight into the mental processes of designers and could help to identify critical situations in which difficulties arise (Zimmerer and Matthiesen 2021). The use of cognitive load as an automatically measurable metric could enable researchers not only to improve the efficiency of data collection and evaluation by finding the relevant and critical situations during design more quickly, but also to analyze and interpret the design processes more accurately.

Nevertheless, to achieve such automatization and the associated benefits, it is necessary to initially rely on traditional design research methods (Lohmeyer and Meboldt 2016). These include, for example, interviews or protocol analyses. These research methods are essential for the development of automated approaches, as only in this way the necessary interpretability of physiologic signals can be achieved. Combining these traditional research methods with new automated approaches presents a unique opportunity due to the inherent diversity among engineering designers, each with distinct needs that require tailored support (Badke-Schaub et al. 2011). Hence, the use of physiologic studies in design has the potential to advance our understanding of design (Nguyen et al. 2018).

Based on this, the problem at hand is that there is a need for suitable research methods that enable a reduced effort in the evaluation of empirical studies in design research. Large numbers of participants are complicated by some of the current research methods, such as protocol analysis, interviews, or analysis of eye tracking videos, because they require a lot of manual work during the evaluation. This is particularly noticeable in studies that analyze the cognitive processes of designers, e.g., during concept synthesis. Therefore, the goal is to develop an algorithm that automatically analyses eye tracking data and predict critical situations in which the participant had difficulties in the design process by identifying high cognitive load. This way, subjective load as actually perceived by the participant is measured, eliminating the need for interpretation by an external evaluator. By providing these situations that are of interest to the researcher, the overall aim is to speed up the evaluation of empirical studies and thus open up the possibility to increase the number of participants and the robustness of study results. To achieve the goal, the following research question will be answered:

> How can difficulties of participants in the processing of design tasks be automatically identified and classified using machine learning algorithms based on eye tracking data?

Three different classification approaches are evaluated to answer the question. The use of eye tracking will enable a more precise and less intrusive capture of the cognitive behavior of designers. At the same time, by using eye tracking data, there is also the potential to enable the evaluation to be carried out continuously over the entire process, taking into account the viewpoint of the participants and not of external evaluators. By investigating an automated approach to measuring cognitive load with high accuracy and low deviation using eye tracking data, researchers may be able to reduce the effort involved in analyzing such studies.

2 Background

This paper focuses on the use of eye tracking data to capture cognitive load. In the following, the background on eye tracking and existing algorithms for recording cognitive load are explained in detail.

2.1 Eye tracking and cognitive load

Eye tracking is a technology for monitoring and measuring a person's eye movements and gaze points when viewing a visual stimulus, such as a computer screen or a real-world environment. It is a powerful tool for understanding human visual perception and behavior (Shaikh and Zee 2018). The most important eye tracking measures include pupil diameter, blinks, fixations and saccades (Holmqvist 2011). The pupil diameter describes the size of the pupil surrounded by the iris and is typically measured in millimeters. Blinks, the

brief closure of the eyelids, can be recorded in frequency and duration. Fixations refer to the moments when the eyes are still and absorb visual information; their duration is measured in milliseconds. Saccades are rapid eye movements between two fixations and are usually measured in degrees of the visual angle between two fixations. They show how the eyes search for new visual information. By combining these measured variables, detailed insights into the visual and cognitive processes of subjects can be gained.

To measure cognitive load, there are various options. The most widespread is the use of questionnaires. The Nasa-TLX (Hart and Staveland 1988; Hart 2006) is the most common here, but alternative variants such as the overall workload scale or the Subjective Workload Assessment Technique are also used (Hill et al. 1992). However, the disadvantage of these questionnaire-based survey methods is that the time periods for which the cognitive load is surveyed are relatively long or even only possible for entire tasks.

This is where the think aloud method (Ericsson and Simon 1993) offers potential, as the participants can indicate where the cognitive load was particularly high in relation to the respective situation. By using retrospective think aloud (RTA), this can also be done without influencing the cognitive load during the actual task processing. At the same time, RTA has already been proven to be a valuable research method for analyzing designers procedures in the design process, especially in combination with eye tracking. (Ruckpaul et al. 2014a).

Numerous studies have already been carried out to establish a connection between various eye tracking metrics and cognitive load (Marquart et al. 2015; Andrzejewska and Stolińska 2016; Chen et al. 2016, 2019; Zimmerer et al. 2023). Skaramagkas et al. (2023) present an overview of the frequency with which the different metrics are used to measure cognitive load. The most frequently used metrics are those based on pupil diameter, followed by fixation-based measures. In the latter case, the duration and the number are mainly used: Saccade-based features, like amplitude and velocity, and blink rate and blink duration are also frequently used.

2.2 Algorithms to detect cognitive load

The assessment of cognitive load through algorithms based on eye tracking metrics has been a subject of interest in several studies (Halverson et al. 2012; Nourbakhsh et al. 2013; Borys et al. 2017; Chen et al. 2019; Wu et al. 2020). The core findings on the most important aspects are described below.

Halverson et al. (2012) investigated the utility of eye movement and pupil-related metrics in assessing workload during complex tasks. They extracted eye metrics from time series using windows of varying sizes (1–30 s) and overlapped them to retain information. Using support vector machines and ten eye metrics, they aimed to predict workload levels. Their results demonstrated the effectiveness of pupil size and the percentage of eye closure (PERCLOS) in workload prediction, with a mean accuracy of 81%, though some outliers only achieved 16%. (Halverson et al. 2012) Despite the high accuracy, the study faced significant issues with high variability and lacked temporal resolution as short intervals were selected for the windows, but these were only associated with the overall workload of the entire task, limiting its applicability for continuous measurement. However, this continuous measurement is a necessary requirement to transfer the findings to design research.

Nourbakhsh et al. (2013) focused on using eye blink features combined with galvanic skin response (GSR) to classify different levels of cognitive load. Participants completed arithmetic tasks of varying difficulty levels, and the classification was done using SVM and Naïve Bayes algorithms. Their results showed that both GSR and blink features effectively classified stress levels, with combined features further improving accuracy. The study achieved 75% accuracy for binary classification (BC) and 53.6% for four-class classification. (Nourbakhsh et al. 2013) However, using arithmetic tasks means that the study lacked temporal resolution as the cognitive load was again only considered for the hole task. In addition, the tasks used do not represent complexities encountered in design research, limiting its direct applicability.

Borys et al. (2017) too conducted a study on classifying cognitive workload during arithmetic tasks using eye activity and in addition EEG features. They aimed to differentiate between cognitive workload states and a no-task control condition using various classification approaches. Their results showed a 90% accuracy for BC using SVM with six eye tracking features, while multi-class classification reached a maximum of 73% accuracy. (Borys et al. 2017) Despite the high accuracy, the results are insufficient for the problem at hand due to the lack of temporal resolution and the contrast between task processing and pause. Even if the results here are promising, the comparison in relation to the task-free phases is not directly transferable, since in relation to the activities of design it is not a matter of measuring against no workload but against a usual workload.

Chen et al. (2019) presented a cognitive load assessment method considering individual differences in eye movement data during flight tasks with varying difficulties. They performed k-means clustering before classification to divide the participants into two groups, for each of which an SVM classifier was then trained. This approach intended to improve classification results by accounting for individual physiologic differences. Despite this innovative approach, the study achieved a maximum accuracy of only 43.7%. (Chen et al. 2019) Furthermore, it also lacked temporal resolution.

Wu et al. (2020) examined the correlation between eye tracking metrics and perceived workload in robotic surgical tasks. Surgical trainees participated in simulated robotic tasks, where pupil diameter and gaze entropy were found to differentiate workload levels, both increasing with task difficulty. Using these eye tracking features, a classification model achieved an accuracy of 84.7% in predicting the workload of a given task. (Wu et al. 2020) However, this study focuses on discrete tasks. This limits its applicability to continuous workload assessment that is necessary for application in design research.

2.3 Summary of the state of research

To summarize, current preliminary work in the field of automated and continuous-time measurement of subjective difficulties is only suitable for the requirements of design research to a limited extent. Although the studies considered show great potential in the combination of eye tracking metrics and machine learning to classify cognitive load, their applicability to the specific aim of design research is limited. The main reasons for this are either the lack of accuracy or the lack of temporal resolution and the high scatter in the results. The research methods analyzed are often not able to measure subjective difficulties continuously over time, which is an essential requirement for design research. Instead of classifying the entire task according to cognitive load, it is necessary to identify specific time periods within a task in which the participant had difficulties. This opens up the possibility of analyzing specific segments of interest and gaining deeper insights into the design process.

Furthermore, most existing studies rely on discrete, well-defined tasks, such as arithmetic or simulated surgical procedures, which differ fundamentally from the iterative and dynamic nature of design tasks. These approaches often lack the ability to link temporal workload fluctuations to the phases and activities of the design process, a critical aspect for understanding when and why difficulties arise. Addressing these limitations is essential for advancing the applicability of automated measurement methods in design research.

3 Methodology

In the following chapter, the data set, consisting of eye tracking data and difficulties of the participants, and its creation are presented first (chapter 3.1). This is followed by a description of the procedure used to develop machine learning algorithms to answer the research question (chapter 3.2).

3.1 Dataset

The data set used to train the algorithm was obtained in advance as part of a participant study. To make the results easier to understand, the study and thus the development of the data set are briefly presented below.

3.1.1 Participants

In the dataset used, a total of 28 student participants were involved, with an average age of 23.5 years (SD = 2.3), ranging from 20 to 29 years old. Regarding their experience in the field of design, the mean value was 9.3 semesters (SD = 3.4), ranging from 5 to 14 semesters of experience. Among the participants, 21 had a background in mechanical design, while 7 were in the mechatronics course. 13 participants already had their bachelor's degrees. In terms of vision, all participants either had normal vision or vision corrected to normal. Prior to their involvement in the study, all participants were required to affirm their participation through a declaration provided by the KIT data protection office.

3.1.2 Tasks

The tasks in the study related to mechanical design, more specifically sheet metal design. This manufacturing process was chosen for several reasons. First, a large number of individual process steps, such as laser cutting, bending and welding, must be taken into account in sheet metal design, which means that numerous production-related constraints have to be taken into account. Second, preliminary investigations have shown that the application of the manufacturing process quickly led to difficulties for the participants, which is why it was well suited for generating a sufficiently large data set. Each participant was given three different tasks to complete:

The first task was to generate different concepts for a given anchoring base using an alternative manufacturing process. The existing design was declared as a prototype and was to be optimized in terms of manufacturing costs for mass production. Several alternative concepts were to be generated from the nine-part welded design as a sheet metal bending design to reduce the number of parts and thus the manufacturing costs. At least three alternative concepts had to be generated.

In the second task, a first prototype of a commissioning unit was presented. This was designed as a quick, simple prototype made of high-density fiberboard (HDF). The task was to solve the problem of jamming when delivering different colored cubes and at the same time to convert the given design into a final design. For this purpose, sheet metal bending design should be used instead of HDF panels.

In the third task, a latch design was provided that is used to lock drawers. This assembly, already designed as a sheet metal bending construction, was to be optimized due to signs of wear in the application. The participants were supposed to reduce the number of moving parts relative to each other by redesigning the latch.

3.1.3 Data collection and data set

The three tasks were completed in a randomized order to minimize sequence effects. First, the participants read through the respective task. Once they had understood them, they began working on the task, drawing alternative concepts on paper while their eye movements were recorded using eye tracking glasses, the Tobii glasses 2. Each task took 10 min to complete. After completing the task, the participants watched their own video recording and commented retrospectively on the situations in which they had difficulties and the types of difficulties they encountered.

The raw data, recorded by the Tobii eye tracking glasses at a sampling rate of 100 Hz, were first checked for completeness. Data sets with more than 25% missing data points were removed entirely. This resulted in a data set of 81 usable recordings.

In the initial analysis of the time series recorded by the Tobii glasses, segments with more than 50 consecutive missing values—equivalent to half a second—were deleted. If the segments of consecutive missing data were shorter than half a second, the missing values were linearly interpolated. The data were then divided into two categories: pupil-based and event-based.

The pupil-based features, recorded directly by the glasses, included gaze point, gaze direction, pupil position, and pupil diameter. First and second derivatives were also calculated to resolve spatial dependencies of individual metrics.

The event-based features encompassed eye movements, categorized as fixation, saccade, and blink. These were identified using the Tobii I-VT (fixation) algorithm (Olsen and Matos 2012), which specifies the duration of the event and calculates the fixation point in the case of fixation. A detailed description of the algorithm can be found in Olsen and Matos (2012). Several motion-based features were derived from the event classification. For each motion type, the number of events, the occurrence rate, and the variance were calculated for the given windows. In addition, the percentage of time the eyes were closed [PERCLOS see (Wierwille et al. 1994; Dinges and Grace 1998)] was calculated for each time segment.

Building on the features described, their applicability to design studies lies in their ability to capture nuanced aspects of designers' cognitive and visual behavior during complex tasks. Pupil diameter and PERCLOS are generally known as indicators of cognitive load, helping to identify moments when designers are experiencing high workload. Fixation metrics, such as duration and variance, provide insights into how designers allocate their visual attention and saccade metrics can be linked to the evaluation of specific design elements or problem areas. By analyzing these variables in the context of specific design phases or activities, researchers can map physiologic responses to cognitive difficulties or decision-making moments, thus making the data directly relevant and interpretable for design studies.

The participants themselves documented the difficulties they encountered on a standard form. The participants received this form after completing each task and filled it out while watching their own video recording of the task again. They could pause and skip back and forth as they wished. These self-reported difficulties served as the primary data source.

To categorize the difficulties, an inductive approach was employed, meaning that the categories were not predefined by the researchers but emerged from the participants' documented responses during the study. After collecting all the responses, the researchers analyzed the descriptions and grouped them based on recurring themes and patterns in the data. Through this iterative process of qualitative coding, eight distinct categories of difficulties regarding the following design activities were identified: *'assembly'*, *'concept selection'*, *'embodiment design'*, *'function unclear'*, *'general overload'*, *'manufacturing process'*, *'no concept idea'* and *'visualization'*. This coding approach ensured that the categories were grounded in the participants' own experiences, enhancing the validity and relevance of the classification system. Figure 1 provides an overview of the number and distribution of the individual categories of difficulties that made up the final data set for algorithm development.

3.2 Algorithm development

This chapter describes the procedure for the algorithm development, consisting of data preparation followed by classification. In data preparation, the combined data from eye tracking and RTA is processed further. First, the data are segmented, the time series is divided into windows and the relevant features are calculated. The data are then standardized to ensure consistent scaling. Feature selection is used to identify the most important features to reduce the complexity of the model. To counter the imbalance of the distribution of classes, upsampling is used next. After data preparation, classification takes place, which includes both binary and multibinary classifiers. Figure 2 summarizes the procedure below.

3.2.1 Data preparation

During segmentation, the time series were prepared so that the features for the respective windows could be extracted. In this study, the recording time series are divided into shorter windows for which the respective features can then be calculated. Variable parameters here are the window length in seconds and the overlap of the windows in percent. This procedure allows the individual segments to be classified and the exact times at which the participant encounters a difficulty. A schematic representation of this procedure is shown in Fig. 3.

After segmentation, the data were standardized. This is a fundamental requirement for classifiers to ensure good classification quality. This process ensures that the features are comparable and do not have large influences due to their scaling, which is especially important for distance-based algorithms. For each feature type and its derivatives, the mean, median and maximum were calculated within the given time windows. To eliminate individual differences between the participants and the environmental conditions, the difference and the quotient of the individual data points with the mean value of the entire recording area were also used as additional features.

During feature selection, the most important features were determined from a pool of available features that had the greatest influence on the predictive ability of the model. The aim of feature selection is to reduce model complexity, avoid overfitting and reduce calculation time by removing less relevant or redundant features. First, a variance analysis was performed to identify and remove features with too little variance, as these are irrelevant for the classification. The feature spaces were then transformed into a normal distribution using a transformer, as some of the used classifiers require this for good functionality. The RobustScaler from scikit-learn was used here, as it is less susceptible to outliers and delivers more robust results. In a further step, the number of features used for training was reduced. For this purpose, an ANOVA analysis of variance was applied to the training data and a score was assigned to the features. The

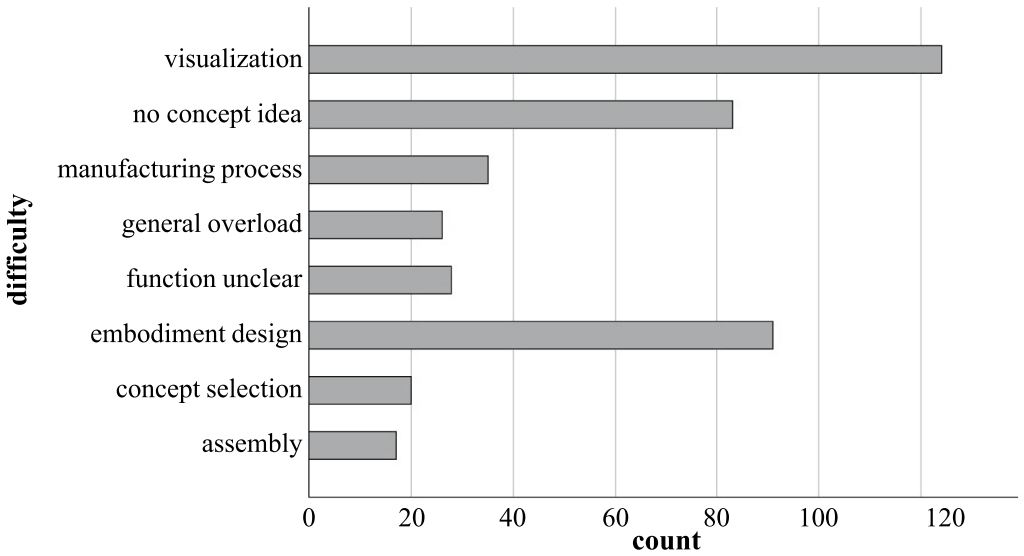

Fig. 1 Distribution of difficulties reported by the participants in the data set

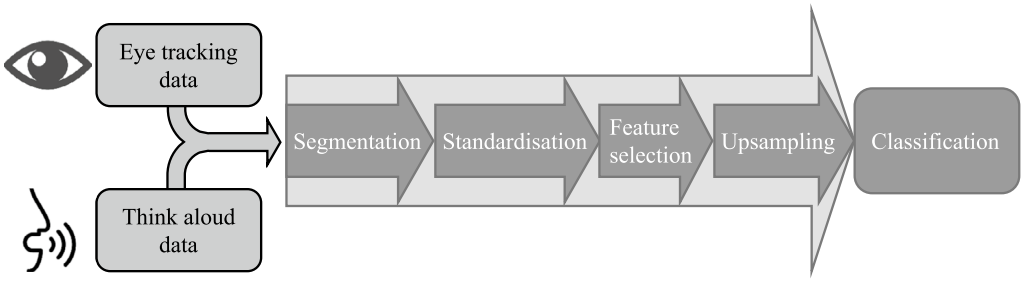

Fig. 2 Process of data preparation and classification

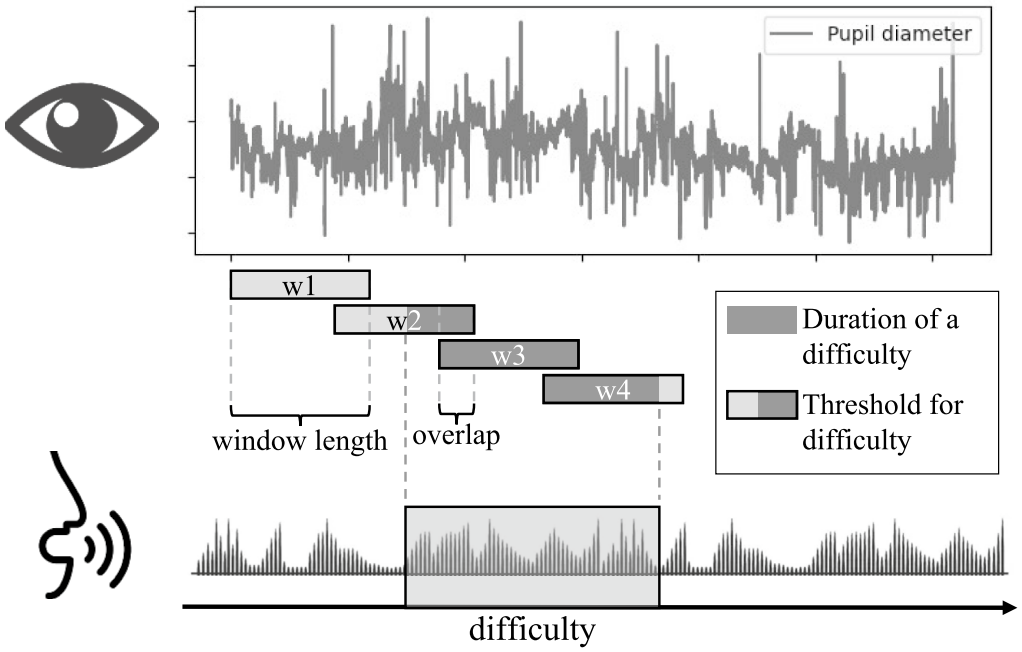

Fig. 3 Procedure for segmentation. This involves dividing the time series into individual overlapping windows and then assigning whether there is a difficulty in the respective windows

best features were selected based on this score. After varying the number of features to be selected, it was found that the best classification results could be achieved with the two highest-scoring features.

As a further step, upsampling was carried out. In the present data set, the classes are distributed unequally, which can lead to the model favoring the over-represented class and poorly classifying the under-represented class. In the BC 'difficulty'—'no difficulty', the difference between the two classes is still negligible. However, if individual difficulties are to be classified against the rest, as is the case with multibinary classifiers, problems can arise when training the classifiers. To solve this problem, the RandomUpsampler from imbalanced-learn was used. Here, individual data points of the less strongly represented class were duplicated to equalize the distribution and thus improve the classification quality. Upsampling strengthens the model capacity and increases the generalization capability, which leads to a more balanced and reliable classification.

3.2.2 Classification

3.2.2.1 Parameter study To determine the optimal combination of parameters for classification, a parameter study was carried out. Four variables were analyzed: the window length, the overlap of the windows, the threshold for classifying a window as a difficulty and the duration of a difficulty.

Three factor levels were considered for each of these variables:

- window length: 5 s, 10 s, 30 s;
- overlap: 0%, 25%, 50%;
- threshold for difficulty: 50%, 75%, 100%;
- duration of a difficulty: full duration, first 30 s, first occurrence only.

This resulted in a total of 81 combinations, each of which was tested to identify the ideal combination of parameters. The aim was to maximize the classification quality and ensure the most balanced and reliable classification possible.

3.2.2.2 Binary classification The established way to generate a two-class problem from a multi-class problem is

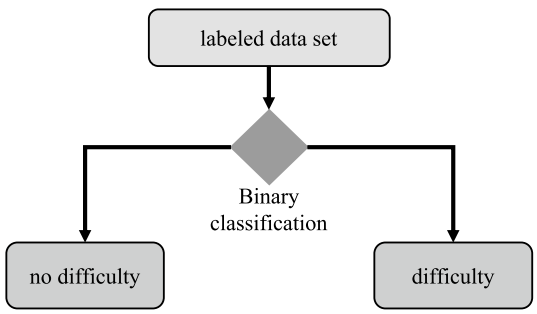

Fig. 4 Functionality of a binary classifier

to combine all difficulties into one class and adopt a BC approach (see Fig. 4). Eleven different classifiers were trained to determine the most appropriate one for the problem. The selection consisted of the following classifiers: logistic regression, support vector machine, k-nearest-neighbors, Gaussian Naive Bayes, passive aggressive classifier, Decision Tree Classifier, Random Forest Classifier, ridge classifier, gradient boost classifier, perceptron classifier, and multi-layer perceptron classifier.

3.2.2.3 Multibinary classification A different approach for solving multi-class problems in machine learning classification is multibinary classification. Here, each class is classified individually against all remaining classes, and the resulting binary classifiers are combined to form a multibinary classifier. This approach is called one-against-all (OAA). For each resulting BC problem, the best classifier and the best features are selected based on the training data. Each classifier then classifies the test data set, and each data point that is identified as a difficulty by at least one of the classifiers is labeled accordingly. Thus, the score calculation is again applied to the two-class problem on which the binary classifier is based. As an additional result, the multibinary classifiers provide additional information about the specific difficulty for each segment labeled as such.

There are various approaches for combining the training data for the individual binary classifiers. The OAA algorithm presented by Jeatrakul and Wong (2012) is used here. Based on this OAA algorithm, a variation of this approach was developed, the cascading one-against-all (cOAA).

In the OAA algorithm, the same training data set is always used and only the class assignment of each data point is changed. With the cOAA algorithm, on the other hand, a sequence is defined for the classification of the difficulties, starting with the difficulty that can be classified best. After a difficulty has been classified against the others, all data points belonging to this difficulty are removed from the training data set. As a result, the remaining data to be classified becomes smaller with each classification level, which should theoretically improve the classification quality of the subsequent binary classifiers.

A comparison of the two approaches is provided in Fig. 5.

3.3 Scoring metric

In this study, two metrics were used to evaluate the classifiers: the f1-score, which is a well-established metric, and the fn-score (false-negative-score), which is of particular interest for the use case. To determine the scores and standard deviations, a fivefold cross-validation was performed.

The f1-score, a harmonic mean of precision (how many of the predicted positives are correct) and recall (how many of the actual positives are correctly predicted), is widely used for unbalanced datasets and ranges from zero to one, with higher values indicating better performance of the classifier. A high f1-score close to one means that the classifier is performing very well at balancing precision and recall. Figure 6 explains the calculation of the score. A more detailed explanation of the score can be found in Sasaki (2007).

The fn-score is a measure of the false-negative rate. The false-negative rate indicates the percentage of data points that are true positives (segments with difficulties) but are incorrectly labeled as negatives (not difficulties). Since lower false-negative rates are better, the fn-score is calculated as

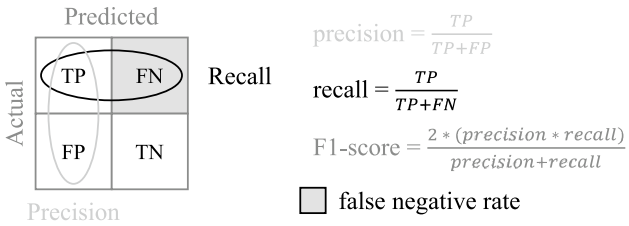

Fig. 6 Calculation of the f1-score

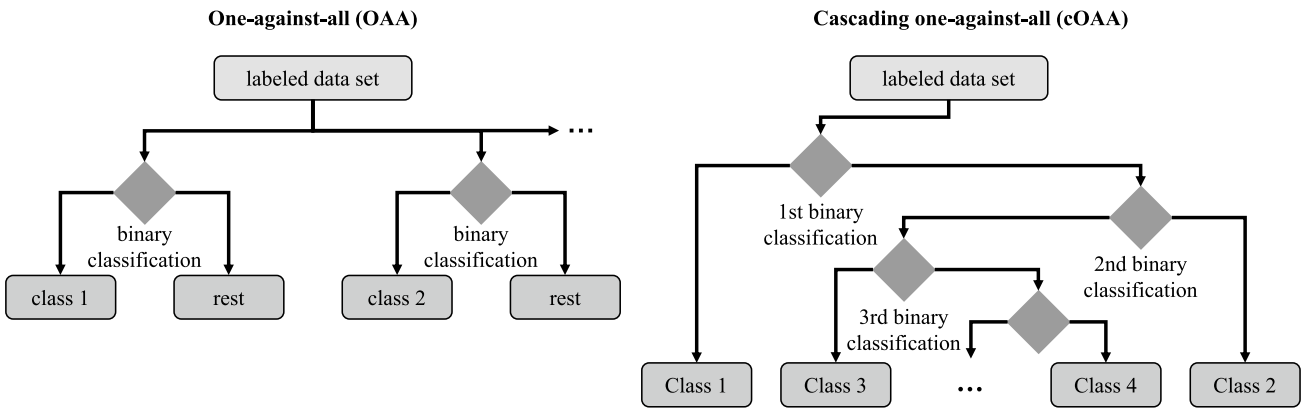

Fig. 5 Graphical comparison of the multibinary approaches one-against-all (OAA) and cascading one-against-all (cOAA)

the inverted fn-rate (1 − fn-rate) to make it more comparable with the f1-score. This means the fn-score also ranges from 0 to 1, with higher values indicating better performance. For example, an fn-score of 0.95 suggests the classifier correctly identifies 95% of the segments with difficulties.

In this use case, this score is important as the algorithm is used for the automated evaluation of eye tracking studies. It is therefore crucial that as few segments with difficulties as possible are not detected and included in the manual evaluation. However, incorrectly classified difficulties can be easily identified and corrected by the study director.

4 Results

4.1 Parameter study

The parameter study was initially carried out with the binary classifier. This showed that all parameter combinations that included a shorter duration of the difficulty achieved significantly poorer results. It also turned out that the threshold for a difficulty only plays a minor role in the classification quality. Based on these findings, only the window length and the overlap of the windows were used as variable parameters for the more computationally intensive OAA and cOAA approaches to save computing capacity. A value of 75% was set for the difficulty threshold, and there was no limit to the duration of a difficulty in all further calculations.

The results of the parameter study are presented in Fig. 7. The three classification approaches are compared for each parameter combination of window length and overlap. The f1-score for the respective classifier is shown in light, the false-negative-score (fn-score) in dark grey. The lines at the top of the bars represent the standard deviation.

The binary classifier showed that the f1-score improved the shorter the window was and the less overlap the windows had. The fn-score remained largely constant across all parameter combinations, although it was also affected by a high standard deviation.

For the OAA classifier, the influence of window overlap was significantly greater than that of window length. The best results were obtained with large overlap of 50%, and the classification quality decreased with decreasing overlap. The fn-score was higher than for the binary classifier, and again the standard deviation was relatively high. There was no clear pattern in the dependence of the fn-score on the parameters.

The cOAA achieved the best results in the f1-score. A clear trend was that the f1-scores increased with greater overlap, with significantly better results being achieved with a window length of 5 s than with longer windows. The fn-score was also found to be significantly lower with shorter windows. Overall, the fn-scores were higher and the

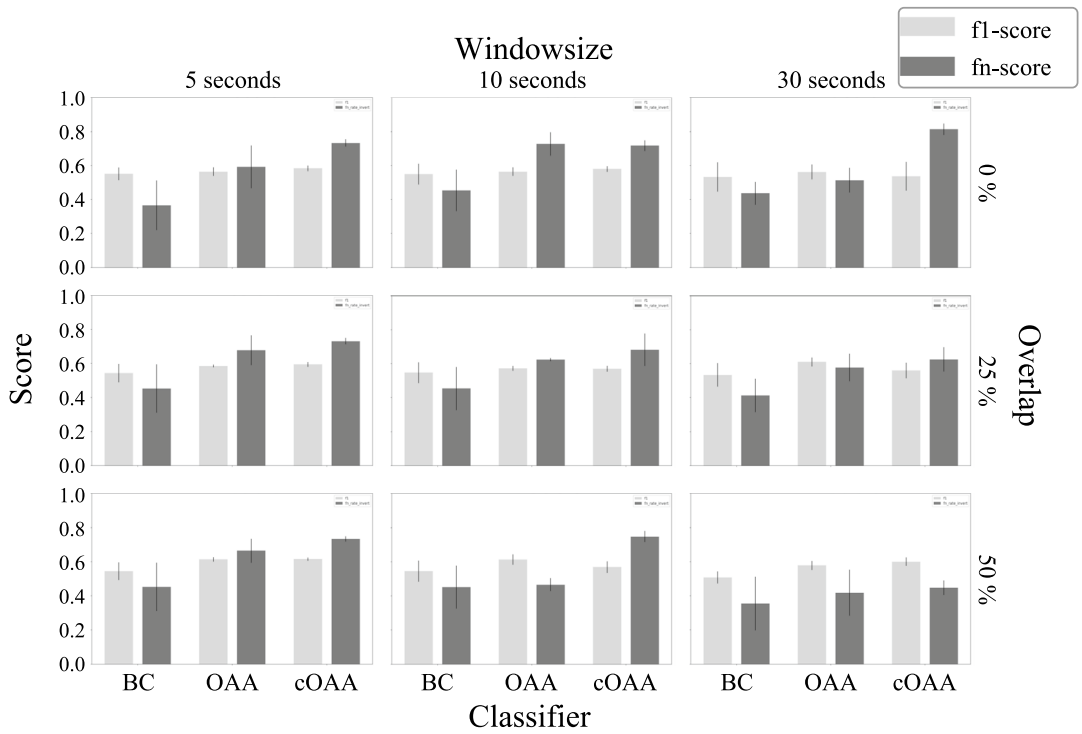

Fig. 7 Results of the parameter study. For the binary classification the best results are achieved using a window length of 5 s and an overlap of 0% (top left). The multibinary approaches in turn achieve the best results with a window length of 5 s and an overlap of 50% (bottom left)

standard deviations of both scores were lower than for the other classifiers.

These results emphasize the importance of the choice of window length and overlap for the performance of the classifiers. Especially the cascading OAA approach showed promising results in terms of accuracy and reliability of the classification.

After selecting the most suitable parameters for each case, the individual classification approaches are presented in detail below.

4.2 Binary classification

For the BC, all types of difficulties were summarized in the category 'difficulty'. This results in only the two classes 'difficulty' and 'no difficulty'. As can be seen from the parameter study, the best classification results are obtained with a window size of 5 s and an overlap of 0%. The feature selection using ANOVA showed that the features mean pupil diameter and median pupil diameter have the highest scores and are therefore used for classification.

Table 1 shows the results of the cross-validation. The f1-score lies between 0.419 and 0.552, with a mean standard deviation of 0.05. The Support Vector Machine classifier achieved the best results with an f1-score of 0.552 (SD = 0.037), followed by the Ridge classifier, which achieved comparable results with a score of 0.544 (SD = 0.062).

For the fn-score, there were significant differences between the classifiers. The fn-score lies between 0.176 and 0.704, with an average standard deviation of 0.15. The perceptron classifier achieved the best value here with 0.7038, but also showed a very high standard deviation of 0.26.

4.3 Multi binary classification

For multibinary classification, a parameter combination of a window size of 5 s and an overlap of 50% proved to be optimal. In contrast to BC, an individual selection of features and classifiers was made for each class of difficulty.

First, the best classifiers were determined individually for each difficulty. Figure 8 illustrates the f1-scores of the best classifiers for each specific difficulty in the multibinary classification. The different difficulties are listed on the x-axis: '*assembly*', '*Concept selection*', '*embodiment design*', '*function unclear*', '*general overload*', '*manufacturing process*', '*no concept idea*' and '*visualization*'. The y-axis represents the f1-score, which serves as a measure of the classification quality.

The bars in the figure represent the average f1-scores of the best classifiers for each difficulty, while the error bars at the top of the bars indicate the standard deviation of the f1-scores. This allows an assessment of the variability and reliability of the classification results.

The overall classifier achieves a combined f1-score of 0.6137 and an fn-score of 0.6646. The highest classification score for a single difficulty was achieved for the difficulty *'concept selection'*, with an f1-score of 0.81. The difficulties *'visualization'*, *'manufacturing process'* and *'assembly'* are in a medium range with f1-scores between 0.73 and 0.74,

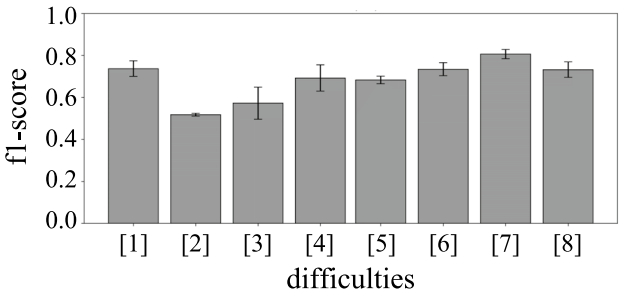

Fig. 8 Highest possible classification result for each class of difficulty using the most appropriate classifiers ([1]: visualization; [2]: no concept idea; [3]: embodiment design; [4]: general overload; [5]: function unclear; [6]: manufacturing process; [7]: concept selection; [8]: assembly)

Table 1 Comparison of the tested classifiers and their achieved scores

Classifier	f1-score	f1-score-std	fn-score	fn-score-std
Support vector machine	0.552	0.0371	0.3667	0.1467
Ridge classifier	0.544	0.0624	0.5554	0.1565
Gaussian Naive Bayes	0.5437	0.0514	0.4504	0.1482
Logistic regression	0.5434	0.0618	0.5562	0.158
Gradient boosting classifier	0.5424	0.0296	0.429	0.0748
MLP classifier	0.5398	0.0291	0.4177	0.141
k nearest neighbors	0.5179	0.014	0.4888	0.0743
Random forest classifier	0.5036	0.0242	0.4162	0.0844
Decision tree classifier	0.5026	0.0227	0.439	0.0867
Perceptron classifier	0.4292	0.1011	0.7038	0.26
Passive aggressive classifier	0.4185	0.0951	0.1755	0.3511

while '*general overload'* and '*function unclear'* achieve an f1-score of slightly below 0.7. The difficulties '*embodiment design'* and *'no concept idea'* have the lowest scores of 0.57 and 0.52 respectively and thus the lowest classification quality of all difficulties.

A low standard deviation, as for the difficulties *'no concept idea'*, *'embodiment design'*, *'function unclear'* and *'manufacturing process'*, indicates consistent classification performance. A medium standard deviation can be observed for the difficulty *'general overload'*, which indicates a somewhat greater variability in the results. The highest standard deviations can be seen for the difficulties *'visualization'*, *'concept selection'* and *'assembly'*, which indicates a higher variability of the classification results despite high f1-scores.

For each difficulty, the classifiers and features with the highest accuracy were used. Table 2 lists the features and classifiers selected in each case. The Decision Tree Classifier and the Random Forest Classifier were each used three times. The Gradient Boosting Classifier and the K-Nearest Neighbors Classifier were each used once. There are only a few overlaps among the features used: 'Pupil diameter (mean)', 'Fixation point (1st derivation, median)' and 'Gaze direction (2nd derivation, max)' were each used twice. Interestingly, there are no identical combinations of features for the different difficulties.

Looking at the features used, it can be seen that pupil-based features were used six times. Fixation-based features were used five times, while gazing direction was used as a feature four times. In addition, blinking was used once as a feature. It is noteworthy that saccade-based features were not used.

To further increase accuracy, a cascading OAA approach was also applied. The order of the individual classifications was determined in such a way that the best possible overall result is achieved.

Table 3 Change in the individual classification results due to the cascading one-against-all approach

Difficulty	f1-score OAA	f1-score cOAA
Concept selection	0.8059	0.8059
Visualization	0.7373	0.7436
General overload	0.6925	0.7588
Assembly	0.7325	0.7642
Manufacturing process	0.7345	0.7366
Function unclear	0.6837	0.6985
No concept idea	0.5176	0.5475
Embodiment design	0.5728	0.5586

Table 3 presents the changes in f1-scores between the OAA and the cOAA approaches for various difficulties. As shown in the table, the cascading approach led to an improvement in the f1-scores for most of the difficulties.

The *'concept selection'* difficulty maintained its f1-score of 0.8059, indicating no change. This lack of change can be attributed to the fact that this difficulty is classified first, with no data points yet removed, mirroring the situation in the OAA approach.

Improvements can be observed in all remaining categories except the *'embodiment design'* class. The greatest improvement can be observed in the *'general overload'* class. Here the score rises by 0.06 to 0.76. There was also a significant improvement of 0.03 in the *'assembly'* class. However, there was a slight deterioration observed in the *'embodiment design'* class, where the f1-score decreased from 0.5728 to 0.5586.

The final result of the classification is an overall f1-score of 0.615 (SD = 0.001). The false-negative score is even higher with 0.734 (SD = 0.016). For better

Table 2 Used feature OAA

Difficulty	Classifier	Used features
Assembly	Random Forest Classifier	Gaze direction (1st derivation, max) Pupil position (1st derivation, max)
Concept selection	Decision Tree Classifier	Gaze direction (2nd derivation, max), Fixation point (2nd derivation, median)
Embodiment design	Decision Tree Classifier	Pupil diameter (mean), Pupil diameter (2nd derivation, median)
Function unclear	K-Nearest Neighbors Classifier	Blink rate, Fixation point (2nd derivation)
General overload	Random Forest Classifier	Fixation point (1st derivation. max), Pupil diameter (mean)
Manufacturing process	Random Forest Classifier	Gaze point (1st derivation, max), Gaze direction (2nd derivation, max)
No concept idea	Gradient Boosting Classifier	Pupil diameter (mean), Pupil diameter (median)
Visualization	Decision Tree Classifier	Fixation point (2nd derivation, max), Fixation point (1st derivation, median)

Table 4 Final result of the three classification approaches measured by f1-score and fn-score as well as the associated standard deviations

Classifier	f1-score	f1-score-std	fn-score	fn-score-std
Binary Classifier	0.5520	0.0367	0.3667	0.1467
OAA	0.6137	0.0126	0.6646	0.0703
cOAA	0.6150	0.0089	0.7342	0.0162

comparability, Table 4 summarizes the results of all three approaches once again.

There were some changes in some of the classifiers due to the new approach:

In the *'visualization'*, *'embodiment design'* and *'concept selection'* classes, an improvement in prediction was achieved by changing the classifier used. All three classes used a Random Forest Classifier in the one-against-all approach.

While *'visualization'* and *'concept selection'* now achieve better results with a Decision Tree Classifier, the classifier of the *'embodiment design'* class switches to a support vector machine. In addition to the classifiers, there were also changes in the features used in two difficulty classes. In the *'embodiment design'* class, 'Pupil diameter (1st derivation, median)' was replaced by 'Pupil diameter (median)', while in the *'general overload'* class 'Pupil diameter (mean)' was replaced by 'Fixation point (2nd derivation, max)'.

4.4 Summary of results

The Binary Classifier achieved an f1-score of 0.5520 (SD = 0.0367), and a false-negative rate of 0.3667 (SD = 0.1467). The one-against-all (OAA) approach improved the performance, yielding an f1-score of 0.6137 (SD = 0.0126), and a clearly better false-negative score of 0.6646 (SD = 0.0703). The cOAA approach further enhanced the results, achieving the highest f1-score of 0.6150 (SD = 0.0089), and the best false-negative rate of 0.7342 (SD = 0.0162). A summary is shown in Table 4. Even if the cOAA does not differ considerably from the OAA regarding the absolute f1 score, the clear reduction of the standard deviation and the increased false-negative score should be noted.

5 Discussion

In this paper, the research question "How can difficulties in the processing of design tasks be automatically identified and classified using machine learning algorithms based on eye tracking data?" could be answered. The cOAA algorithm answered this question by demonstrating a high accuracy of 62% and a false-negative rate of 26.6%, utilizing eye tracking data to predict critical situations during design tasks. In the following, the implications are discussed, and the limitations and possible next steps are analyzed.

5.1 Measuring cognitive load via eye tracking

The integration of eye tracking data with traditional research methods represents a significant step forward in the field of design research. Compared to the accuracies reported in the state of research, our approach of combining eye tracking with retrospective think aloud shows notable improvements, especially for complex and extended tasks. Existing research often focuses on short, self-contained tasks like arithmetic problems (Nourbakhsh et al. 2013; Borys et al. 2017), surveyed by questionnaires achieving high accuracy but limited to these narrow scopes. In contrast, the RTA approach employed in this study is designed for longer time intervals and more comprehensive tasks, making them suitable for capturing data over entire design sessions.

The continuous collection of difficulty data through RTA allowed relatively high prediction accuracies even for extended and complex tasks such as concept development. A major advantage of RTA is that it does not impose any additional cognitive load on participants during the data collection process. This non-intrusive research method ensures that participants' natural workflow is not disrupted, resulting in more accurate and authentic data. In addition, RTA measures subjective load as actually perceived by the participant, eliminating the need for interpretation by an external evaluator.

In addition, combining eye tracking data with established qualitative survey methods such as interviews and protocol analysis provides a more nuanced understanding of the design process (Ruckpaul et al. 2014a). The results showed that different features and combinations of features were effective in detecting various difficulties. In particular, the integration of context-specific varying eye tracking metrics helped to achieve higher classification accuracy and a lower false-negative rate compared to the state of research, demonstrating its effectiveness in providing accurate and reliable data on design processes.

Especially the false-negative rate achieved by the cOAA approach supports the suitability of eye tracking data for identifying difficulties in design tasks. Lower false-negative rates have the great advantage that the algorithm does not leave any relevant situations undetected and therefore provides more reliable data for subsequent analysis. This is of crucial importance in design research, especially with regard to the evaluation effort, as the precise identification of design challenges is essential for this.

5.2 Benefits of automated evaluation

The classification results of the three approaches—Binary Classifier, one-against-all (OAA), and cOAA—demonstrate significant differences in their performance. The Binary Classifier achieves an f1-score of 0.5520 with a standard deviation of 0.0367. In contrast, the OAA approach shows an improvement with an f1-score of 0.6137 and a lower standard deviation of 0.0126. The cOAA approach achieves the best results with an f1-score of 0.6150 and the lowest standard deviation of 0.0089, indicating a more consistent performance.

The same picture emerges when looking at the false-negative rate. The binary classifier is far behind with an fn-score of 0.367, which corresponds to a false-negative rate of 63.3%. In a comparison of the multibinary classifiers, the OAA algorithm with an fn-score of 0.665 is also noticeably less accurate than the cOAA with an fn-score of 0.734. The standard deviation of 0.016 for the cOAA is also significantly lower than for the OAA (SD = 0.07). This means that the cOAA does not recognize around a quarter of the difficulties mentioned by the participants.

In conclusion, the cOAA approach, with its cascading strategy, offers a considerable improvement over both the Binary Classifier and the OAA approach, making it a more effective method for accurately and consistently classifying difficulties in design tasks. This advancement highlights the importance of using specialized classifiers and features tailored to specific types of difficulties, thereby enhancing the overall classification performance and reliability in design research studies. This facilitates larger-scale studies and statistically more reliable findings, making eye tracking a powerful research method in design research for capturing detailed and objective data on cognitive processes. With this approach, the efficiency of data collection and evaluation will be increased, and design processes thus better analyzed and interpreted.

The exact extent of the advantage of the approach developed has yet to be fully investigated. In particular, further work is needed to increase the accuracy of the classification models. While the accuracy of 73% of detected difficulties (based on the fn-score) achieved in this study is a very good step in the right direction and shows the potential and feasibility of this project, this value needs to be further increased to improve the practical applications and the reliability of the results.

5.3 Tailored feature selection

These findings have important implications for research into the measurement of cognitive load using eye tracking. In particular, the fact that different types of difficulties can be detected with different eye tracking features is a highly valuable finding.

Fixations are the relevant features for identification in several difficulty categories, especially in *'visualization'*, *'function unclear'*, and *'concept selection'*. This implies that extended fixations in these cases are associated with increased cognitive load and can therefore serve as an indicator of difficulties.

In contrast, the results show a close relationship between pupil diameter and difficulties such as *'embodiment design'* and *'no concept idea'*. As is already known from the state of research, a dilated pupil can indicate increased cognitive load (Marquart et al. 2015; Chen et al. 2016) and thus phases in the design process in which complexity is high or information processing is intensified. Cognitive processes such as thinking and imagination therefore appear to play a greater role in the difficulties presented here, in which pupil diameter plays a role.

Difficulties relating to *'general overload'* are between these two cases. One fixation and one pupil-based feature are relevant here. This may primarily be since many different types of difficulties are combined here and therefore no clear correlations to the relevant features could be identified during clustering.

For classes of difficulties relating to the *'manufacturing process'* or *'assembly'*, features on the gaze path are dominant. In these types of difficulties, sequences and interrelationships between different design features or elements of the structure often play an important role. This fact could explain the strong dependence on gaze behavior in these cases.

From this, it can be deduced that the present results emphasize the need for a differentiated consideration of eye tracking metrics depending on the context, in this case the type of difficulty. For the development of algorithms for recording cognitive load using eye tracking, it is therefore crucial when compiling training data not to focus purely on a one-dimensional measurement of cognitive load, but also to always consider the specific situation at hand.

Nevertheless, further research is needed to validate the accuracy and reliability of the eye parameters in different design contexts and to further improve the integration of eye tracking into the design research process. This study provides an important contribution to the study of cognitive processes in design and lays the foundation for future work on the use of eye tracking to capture difficulties in the design process.

5.4 Implications for the field of design research

The central problem, which was also highlighted in the introduction, is the difficulty of achieving statistically significant results, as design studies are often only carried out with a

small number of participants due to the enormous amount of evaluation required (Dinar et al. 2015; Cash et al. 2022). The present work shows that the developed machine learning algorithms and the cascading classification approach have the potential to significantly reduce the required evaluation effort. Based on the example that one hour of study material requires about 25 additional hours of work in data processing and analysis (Ahmed et al. 2003), even small improvements in the evaluation effort can provide benefits.

This may enable more extensive studies in future and thus improve the statistical robustness of the results. Thus, the use of these research methods could significantly increase the overall validity and reliability of design studies. A further benefit of this approach is that the automated evaluation of eye tracking data can reduce the influence of subjective influences by the evaluators of studies. The objectivity of the evaluation is thus considerably increased by the use of eye tracking data (Ruckpaul et al. 2014b).

6 Conclusion and outlook

The findings of this study represent an important contribution toward the goal of using machine learning algorithms for analyzing eye tracking data in design research. While not a definitive solution, the results demonstrate considerable progress in developing research methods that can automate and enhance the analysis of design processes. Traditional design research methods are often constrained by the intensive resources required for data collection and analysis. By automating these processes, especially the data analysis, this study addresses an essential need in the design community (Dinar et al. 2015; Gero and Milovanovic 2020) for more efficient and scalable research techniques.

The improved classification performance achieved through the cOAA approach is particularly noteworthy. Using this approach, increased accuracy of difficulty detection can be achieved through sequential classification, demonstrating the potential of this approach to improve the robustness of conclusions drawn from empirical studies in design research. In addition, the reduction in the false-negative rate is crucial to identifying real problems, making this approach well-suited for applications that rely on understanding and mitigating design challenges. This is consistent with the objectives outlined by Cash et al. (2022) and other researchers who have emphasized the importance of achieving statistically significant results in design studies. The success of the cOAA approach suggests that future research could benefit from similar approaches, thereby potentially transforming standard practices in the field.

Looking forward, there are several directions for future research that could further increase the impact of these findings. While the current study focused on eye tracking data, other physiologic signals such as EMG and ECG could further enrich the understanding of design processes. Exploring a wider range of machine learning techniques and data characteristics could provide even more comprehensive insights. In addition, future research should consider the long-term acceptance and integration of these research methods within the design community (Reich 2010) as well as the actual time saving potential and the increase in objectivity of the evaluation. These opportunities highlight the potential for continued innovation in design research methodologies toward more effective and efficient study designs.

Acknowledgements This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—"Recognizing cognitively demanding situations in design and measuring them for the semi-automated analysis of empirical studies in method development: AutoCodIng"—Project number 460444004.

Author contributions C. Zimmerer conducted the research of this contribution under the supervision of S. Matthiesen. C. Wittig assisted in the development of the algorithms. The authors wrote and edited the manuscript in close collaboration, reviewing each other's progress continuously.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare no competing interests.

References

Ahmed S, Wallace KM, Blessing LT (2003) Understanding the differences between how novice and experienced designers approach design tasks. Res Eng Des 14:1–11. https://doi.org/10.1007/s00163-002-0023-z

Andrzejewska M, Stolińska A (2016) Comparing the difficulty of tasks using eye tracking combined with subjective and behavioural criteria. J Eye Mov Res 9:1–16. https://doi.org/10.16910/JEMR.9.3.3

Badke-Schaub P, Daalhuizen J, Roozenburg N (2011) Towards a designer-centered methodology: descriptive considerations and prescriptive reflections. In: Birkhofer H (ed) The future of design methodology. Birkhofer, Herbert, pp 181–197

Benedetto S, Pedrotti M, Minin L, Baccino T, Re A, Montanari R (2011) Driver workload and eye blink duration. Transport Res F Traffic Psychol Behav 14:199–208. https://doi.org/10.1016/j.trf.2010.12.001

Borys M, Plechawska-Wójcik M, Wawrzyk M, Wesołowska K (2017) Classifying cognitive workload using eye activity and eeg features in arithmetic tasks. In: Damaševičius R, Mikašytė V (eds) Information and software technologies, vol 756. Springer International Publishing, Cham, pp 90–105

Bruggen A (2015) An empirical investigation of the relationship between workload and performance. Manag Decis 53:2377–2389. https://doi.org/10.1108/MD-02-2015-0063

Cash P, Isaksson O, Maier A, Summers J (2022) Sampling in design research: eight key considerations. Des Stud 78:101077. https://doi.org/10.1016/j.destud.2021.101077

Chen F, Zhou J, Wang Y, Yu K, Arshad SZ, Khawaji A, Conway D (2016) Robust multimodal cognitive load measurement. Springer International Publishing, Cham

Chen J, Zhang Q, Cheng L, Gao X, Ding L (2019) A cognitive load assessment method considering individual differences in eye movement data. In: 2019 IEEE 15th international conference on control and automation (ICCA). IEEE, pp 295–300. https://doi.org/10.1109/ICCA.2019.8899595

Dinar M, Shah JJ, Cagan J, Leifer L, Linsey J, Smith SM, Hernandez NV (2015) Empirical studies of designer thinking: past, present, and future. J Mech des 137:247. https://doi.org/10.1115/1.4029025

Dinges DF, Grace R (1998) PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance. US Department of Transportation, Federal Highway Administration, Publication Number FHWA-MCRT-98–006}

Ericsson KA, Simon HA (1993) Protocol analysis: verbal reports as data, 3rd edn. MIT Press, Cambridge

Escudero-Mancebo D, Fernández-Villalobos N, Martín-Llorente Ó, Martínez-Monés A (2023) Research methods in engineering design: a synthesis of recent studies using a systematic literature review. Res Eng Design 34:221–256. https://doi.org/10.1007/s00163-022-00406-y

Gero JS, Milovanovic J (2020) A framework for studying design thinking through measuring designers' minds, bodies and brains. Des Sci. https://doi.org/10.1017/dsj.2020.15

Halverson T, Estepp J, Christensen J, Monnin J (2012) Classifying workload with eye movements in a complex task. Proc Hum Factors Ergon Soc Annu Meet 56:168–172. https://doi.org/10.1177/1071181312561012

Hart SG (2006) Nasa-task load index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet 50:904–908. https://doi.org/10.1177/154193120605000909

Hart SG, Staveland LE (1988) Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv Psychol 52:139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hay L, Cash P, McKilligan S (2020) The future of design cognition analysis. Des Sci. https://doi.org/10.1017/dsj.2020.20

Hill SG, Iavecchia HP, Byers JC, Bittner AC, Zaklad AL, Christ RE (1992) Comparison of four subjective workload rating scales. Hum Factors 34:429–439

Holmqvist K (2011) Eye tracking: a comprehensive guide to methods and measures. Oxford Univ. Press, Oxford

Jeatrakul P, Wong KW (2012) Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm. In: The 2012 international joint conference on neural networks (IJCNN). IEEE, pp 1–8. https://doi.org/10.1109/IJCNN.2012.6252450

Lohmeyer Q, Meboldt M (2016) The integration of quantitative biometric measures and experimental design research. Exp Des Res 23:97–112. https://doi.org/10.1007/978-3-319-33781-4_6

Maier A, Baltsen N, Christoffersen H, Störrle H (2014) Towards diagram understanding: a pilot study measuring cognitive workload. In: Proceedings of international conference on human behaviour in design

Marquart G, Cabrall C, de Winter J (2015) Review of eye-related measures of drivers' mental workload. Proc Manuf 3:2854–2861. https://doi.org/10.1016/j.promfg.2015.07.783

Matthiesen S, Meboldt M, Ruckpaul A, Mussgnung M (2013) eye tracking, a method for engineering design research on engineers' behaviour while analyzing technical systems. In: International conference on engineering design

Nguyen P, Nguyen TA, Zeng Y (2018) Empirical approaches to quantifying effort, fatigue and concentration in the conceptual design process. Res Eng Des 29:393–409. https://doi.org/10.1007/s00163-017-0273-4

Nourbakhsh N, Wang Y, Chen F (2013) GSR and blink features for cognitive load classification. In: Human-computer interaction–INTERACT 2013: 14th IFIP TC 13 international conference, Cape Town, South Africa, September 2–6, 2013, proceedings, part I 14:159–166

Olsen A, Matos R (2012) Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. In: Eye tracking research and applications symposium (ETRA):317–320. https://doi.org/10.1145/2168556.2168625

Paas FG, van Merriënboer JJ (1994) Instructional control of cognitive load in the training of complex cognitive tasks. Educ Psychol Rev 6:351–371

Reich Y (2010) My method is better! Res Eng Des 21:137–142. https://doi.org/10.1007/s00163-010-0092-3

Ruckpaul A, Fürstenhöfer T, Matthiesen S (2014) Combination of eye tracking and think-aloud methods in engineering design research. Des Comput Cogn. https://doi.org/10.1007/978-3-319-14956-1

Ruckpaul A, Kriltz A, Matthiesen S (2014) Using eye tracking to understand the engineering designers' behaviour in synthesis driven analyzing processes—experiences in study design. In: International conference on human behavior in design

Sasaki Y (2007) The truth of the F-measure. Teach Tutor Mater 1:1–5

Shaikh AG, Zee DS (2018) Eye movement research in the twenty-first century—a window to the brain, mind, and more. Cerebellum 17:252–258. https://doi.org/10.1007/s12311-017-0910-5

Skaramagkas V, Giannakakis G, Ktistakis E, Manousos D, Karatzanis I, Tachos N, Tripoliti E, Marias K, Fotiadis DI, Tsiknakis M (2023) Review of eye tracking metrics involved in emotional and cognitive processes. IEEE Rev Biomed Eng 16:260–277. https://doi.org/10.1109/RBME.2021.3066072

Sun L, Xiang W, Chai C, Yang Z, Zhang K (2014) Designers' perception during sketching: an examination of Creative Segment theory using eye movements. Des Stud 35:593–613. https://doi.org/10.1016/j.destud.2014.04.004

Ullman DG (2010) The mechanical design process, 4th edition. In: McGraw-Hill series in mechanical engineering. McGraw-Hill Higher Education, Boston

Wierwille WW, Wreggit SS, Knipling RR (1994) Development of improved algorithms for on-line detection of driver drowsiness. In: international congress on transportation electronics (1994: Dearborn Mich.). Leading change

Wu C, Cha J, Sulek J, Zhou T, Sundaram CP, Wachs J, Yu D (2020) Eye-tracking metrics predict perceived workload in robotic surgical skills training. Hum Factors 62:1365–1386. https://doi.org/10.1177/0018720819874544

Zagermann J, Pfeil U, Reiterer H (2018) Studying eye movements as a basis for measuring cognitive load. In: Extended abstracts of the 2018 CHI conference on human factors in computing system. pp. 1–6. https://doi.org/10.1145/3170427.3188628

Zimmerer C, Nelius T, Matthiesen S (2023) Using eye tracking to measure cognitive load of designers in situ. In: Gero JS (ed) Design computing and cognition'22. Springer International Publishing, Cham, pp 481–495

Zimmerer C, Matthiesen S (2021) Study on the impact of cognitive load on performance in engineering design. In: Proceedings of the 23rd international conference on engineering design (ICED 21), vol 1, Gothenburg, Sweden, 2761–2770. https://doi.org/10.1017/pds.2021.537