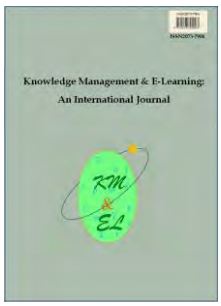
*Knowledge Management & E-Learning, Vol.15, No.3. Sep 2023*

# Concept maps for formative assessment: Creation and implementation of an automatic and intelligent evaluation method

**Tom Bleckmann**
**Gunnar Friege**
Leibniz University Hannover, Germany




**Knowledge Management & E-Learning: An International Journal (KM&EL)**
ISSN 2073-7904

**Recommended citation:**
Bleckmann, T., & Friege, G. (2023). Concept maps for formative assessment: Creation and implementation of an automatic and intelligent evaluation method. *Knowledge Management & E-Learning, 15*(3), 433–447. https://doi.org/10.34105/j.kmel.2023.15.025

*Knowledge Management & E-Learning, 15(3), 433–447*

# Concept maps for formative assessment: Creation and implementation of an automatic and intelligent evaluation method

Tom Bleckmann*

Institute for Didactics of Mathematics and Physics
Leibniz University Hannover, Germany
E-mail: bleckmann@idmp.uni-hannover.de

Gunnar Friege

Institute for Didactics of Mathematics and Physics
Leibniz University Hannover, Germany
E-mail: friege@idmp.uni-hannover.de

*Corresponding author

**Abstract:** Formative assessment is about providing and using feedback and diagnostic information. On this basis, further learning or further teaching should be adaptive and, in the best case, optimized. However, this aspect is difficult to implement in reality, as teachers work with a large number of students and the whole process of formative assessment, especially the evaluation of student performance takes a lot of time. To address this problem, this paper presents an approach in which student performance is collected through a concept map and quickly evaluated using Machine Learning techniques. For this purpose, a concept map on the topic of mechanics was developed and used in 14 physics classes in Germany. After the student maps were analysed by two human raters on the basis of a four-level feedback scheme, a supervised Machine Learning algorithm was trained on the data. The results show a very good agreement between the human and Machine Learning evaluation. Based on these results, an embedding in everyday school life is conceivable, especially as support for teachers. In this way, the teacher can use and interpret the automatic evaluation and use it in the classroom.

**Keywords:** Formative assessment; Concept maps; Machine learning; Feedback

**Biographical notes**: Tom Bleckmann is a doctoral candidate at the Institute for Didactics of Mathematics and Physics of the Leibniz University Hannover, Germany. He is part of the PhD program LernMINT, which focuses on the development and evaluation of data-based intelligent methods and their meaningful integration into STEM subjects His research interests include formative assessment, machine learning, and technology-enhanced feedback.

Prof. Dr. Gunnar Friege is professor of physics education in the faculty of mathematics and physics at the Leibniz University Hannover, Germany. He has been involved in multiple disciplinary and interdisciplinary research in the areas of formative assessment, problem-solving and learning with simulations and inquiry learning. He is one of the speakers of the PhD program LernMINT.

## 1. Introduction

Through formative assessment, students benefit from regular and individual feedback on their strengths and weaknesses. In addition, teachers can not only support their students but also reflect on and optimize their own work. However, in order to be able to implement a formative assessment, it is necessary to have diagnostic information, which must first be collected and evaluated in order to be able to use it for further work in the teaching-learning process. A major problem is that measuring student performance, which is the basis for formative assessment, can take an enormous amount of time. Hunt and Pellegrino (2002), for example, argue for the use of more computer technology as follows: "*The purpose of the technology is to gather information about the student that can be summarized and presented to a teacher so that the teacher can adjust instruction to the dominant ideas present in the class.*"

This paper is also about a new approach that uses the advantages of technology to improve formative assessment: With the help of Machine Learning techniques, a concept map on the topic of mechanics is to be evaluated automatically so that teachers have timely and easy access to diagnostic information that they can use for the further teaching process. For this purpose, concept maps were collected in 14 different classes in Germany, which were then analyzed by two human raters using a four-point rating system. Based on these ratings, a supervised Machine Learning algorithm will be trained and tested. With the help of this automatic and fast evaluation, teachers should in future not only get a quick overview of the performance level of their class but also have the possibility to create individual feedback for their students. In order for this to succeed, this paper describes the results of the Machine Learning evaluation and discusses the didactic implementation of it.

The paper, therefore, starts with a theoretical background on formative assessment, concept maps and Machine Learning. Then the two research questions and the design of the study are presented. After the most important results have been presented and discussed, the paper ends with a short summary and a brief outlook on further research.

## 2. Theoretical background

### *2.1. Formative assessment*

Assessment in education can take place at the class level as well as at the student level and have different purposes (Schütze et al., 2018). A distinction is made between summative and formative assessment (Maier, 2010). In summative assessment, the focus is on a final assessment of teaching-learning processes, which is represented by grades or selection decisions (Maier, 2010). Accordingly, summative assessment is only carried out at the end of a teaching unit or in a final examination at the end of a school year.

In formative assessment, on the contrary, the basic idea is not to give grades or do a final assessment but to provide and use feedback and diagnostic information. On the basis of this, further learning or further lesson planning should be adaptive (Souvignier & Hasselhorn, 2018). Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers,

to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited (Black & Wiliam, 2009). For this reason, formative assessment can be carried out at both class and individual levels (Schütze et al., 2018) and often follows cyclical phases (Wolf, 2014). This means that the formative assessment phases are repeated regularly in class and not just done once. According to Bell and Cowie (2001), these independent and repetitive phases are gathering information, interpreting it, and then acting on it.

According to Wiliam and Thompson (2008), formative assessment can be broken down into five key features, based on Ramaprasad's (1983) three process dimensions of learning state ("*Where the learner is right now*"), learning goal ("*Where the learner is going*") and identifying steps to the learning goal ("*How to get there*"), as well as the people involved (e.g., teachers, peers, learners) (Schütze et al., 2018) (see Fig. 1).

<table>
<tr><th></th><th>Where the learner is going</th><th>Where the learner is right now</th><th>How to get there</th></tr>
<tr><td>Teacher</td><td>1. Clarifying learning intentions and criteria for success</td><td>2. Engineering effective classroom discussions, questions and learning tasks that elicit evidence of learning</td><td>3. Providing feedback that moves learners forward</td></tr>
<tr><td>Peer</td><td>Understanding learning intentions and criteria for success</td><td colspan="2">4. Activating students as instructional resources for one another</td></tr>
<tr><td>Learner</td><td>Understanding learning intentions and criteria for success</td><td colspan="2">5. Activating students as the owner of their own learning</td></tr>
</table>

**Fig. 1.** The five key strategies of formative assessment, Adapted from Wiliam and Thompson (2008)

Fig. 1 shows that the learning goals and success criteria must first be formulated clearly and comprehensibly by the teacher to all students, the individual learner and the peers. Then, different opportunities, such as formative performance tests or diagnostic instruments, should be provided by the teacher to collect diagnostic information (Wolf, 2014). These form the basis for further actions in the formative assessment process (Schütze et al., 2018). The resulting information can be used in different ways. On the one hand, the teacher can use the information to adapt further teaching, e.g., in the form of adapted instructions or repetition of certain topics. On the other hand, the teacher can also provide individual feedback for each student (Schütze et al., 2018). Further possibilities of feedback can also be peer assessments, where students assess each other, or completely independent self-assessments (Schütze et al., 2018). Thus, peers can also play a crucial role in the formative assessment process, providing further opportunities besides the classical teacher-student interaction. In the end, the feedback should show further learning steps for the students and promote self-regulation competence (Wolf, 2014).

Looking at the learning effectiveness of formative assessments, many empirical studies find that lower-performing students in particular can benefit from them (Maier, 2010). In his meta-study, Hattie (2009) was also able to prove a high effect, but the positive results are viewed in a much more differentiated way in more recent studies (Hattie, 2009;

Souvignier & Hasselhorn, 2018). The biggest effect on student performance is expected from so-called short-cycle feedback (Hattie, 2009; Wiliam & Leahy, 2007). This refers to an immediate sequence of the first three key features from Fig. 1 (i.e., goal setting, learning action and feedback), all of which should take place within one lesson (Hattie, 2009; Maier, 2010).

However, this aspect is difficult to implement in reality because, on the one hand, most modern school classes consist of a large number of students and thus one-on-one conversations between students and teachers are not possible. On the other hand, the whole formative assessment process takes a lot of time (Hunt & Pellegrino, 2002). There is no point in formative assessment by a teacher if the teacher cannot identify, analyze, and respond to the problems of individual students (Hunt & Pellegrino, 2002).

How finally the performance measurement is conducted can be very different (Wolf, 2014), because the aspect of formative is not a characteristic of the method that is used, but the way how it is used (Maier, 2010). For example, portfolio tasks, one-minute papers or student response systems are typical formative assessment methods (Schütze et al., 2018). Thus, concept maps can also be used as a summative but also as formative assessment (Ruiz-Primo & Shavelson, 1996). For this reason, the use of concept maps for formative assessment will be discussed in more detail in the next section.

### *2.2. Concept maps*

A concept map consists of different concepts, which are usually enclosed in circles or boxes, and of relationships between the concepts (Cañas et al., 2012). The different concepts are from a certain subject area and are arranged in a certain form (e.g., hierarchically) (Ley, 2015). The relationships between the concepts are represented by a labelled arrow, whereby an arrow can only link two concepts (Ley, 2015). The resulting composition of two concepts and a labelled arrow is also called a proposition (Cañas et al., 2012).

A concept map can be created in many different ways. Ruiz-Primo and Shavelson (1996) have tried to classify the different approaches with the help of the criteria task demands and task constraints. For concept maps, this means they can vary greatly in structure and content. For example, some formats give the students complete freedom and no guidelines, to very strong constraints, such as fill-in-the-map tasks, where the students only have to fill in gaps within the map (Cañas et al., 2012). The choice of task format also has an impact on the level of difficulty of the concept map. Closed forms make the task easier for the students, whereas open forms make the task more difficult (Ley, 2015). The tools to create concept maps can also be varied. It is possible to create concept maps with paper and pencil or to use concept map software such as the IHMC CMap Tool (https://cmap.ihmc.us/cmaptools) (Cañas et al., 2004).

There are many studies in the literature on the use of concept maps as an assessment strategy (see Cañas et al., 2012) or as a research tool (e.g., Campbell, 2022). When creating or editing concept maps, students need not only to understand individual concepts but also to be able to understand their relationships with each other (Llinás et al., 2020). For this reason, concept maps can represent the students' knowledge structure (Ley, 2015). Among other things, misunderstandings about certain connections can be traced and students' prior knowledge can be evaluated (Stracke, 2004). This information can then be used to meaningfully adapt further learning and teaching (Hay et al., 2008). However, regardless of how concept maps are used for formative assessment, they need to be analyzed at some

point. One possibility is to use the strategy proposed by Novak and Gowin (1984) of counting structural attributes such as the number of hierarchical levels or the number of cross-links. This assessment system is well-suited for evaluating the overall quality of students' concept maps (Cañas et al., 2012). However, such a graph theoretical evaluation can only provide limited information about the quality of the content of the individual components (Ley, 2015). A solution could be a qualitative approach in which the individual propositions are analyzed and evaluated in more detail. Although this leads to a higher gain in knowledge and is more useful for formative assessment, this approach also costs a lot of time and is therefore difficult to implement during the normal school day (Hartmeyer et al., 2018). But in order for concept maps to be used quickly and easily by teachers as a formative assessment method in the classroom, they must be able to be evaluated in a time-efficient manner (Ley, 2015). However, since existing computer-based evaluation systems use rather superficial strategies (Anohina & Grundspenkis, 2009), Machine Learning techniques will be used in this work. For this reason, the next section will discuss this topic in more detail.

### *2.3. Machine learning*

Machine learning is a subfield of artificial intelligence and is generally divided into three types: supervised learning, unsupervised learning, and reinforcement learning (Bonaccorso, 2017). Supervised learning is based on statistical algorithms that analyse data with a known structure (Plaue, 2021). This means that the solution to a problem is already available, and the results of the machine learning model can then be compared with it. In other words, the learning process of the models is "*supervised*" (Müller & Guido, 2017). Typical tasks for a supervised machine learning approach are regression problems, such as the prediction of a stock price, and classification problems, such as spam detection of e-mails (Richter, 2019). In contrast, in unsupervised learning, no solutions or no characteristics of the data set are known a priori (Plaue, 2021). The goal is to identify patterns in the data set that are not due to random noise (Richter, 2019). A typical example of unsupervised learning is cluster analysis (Plaue, 2021). The third area of machine learning is reinforcement learning. In this type, the system learns based on effects on the environment, since it receives an evaluation for each output and can thus optimize itself (Pfannstiel, 2022). Classic examples of reinforcement learning are the finding of game strategies or the independent learning of robots (Richter, 2019). For this work, only supervised learning is relevant, which is why the following aspects only refer to it.

The workflow for creating machine learning models usually follows a similar pattern (see Fig. 2). First, the data is put into a form that the computer can handle. Often the images or texts that should be used are transformed into a reel-valued vector $X \in R^n$ with $n \in \mathbb{N}$, using various methods, whereby n can be very large depending on the problem (Richter, 2019). After this is done, the data is split into a training set and a testing set (see Fig. 2). Here, there are also various options, such as a random division into e.g. 90% training and 10% testing data or the use of a cross-validation strategy (Hastie et al., 2009).

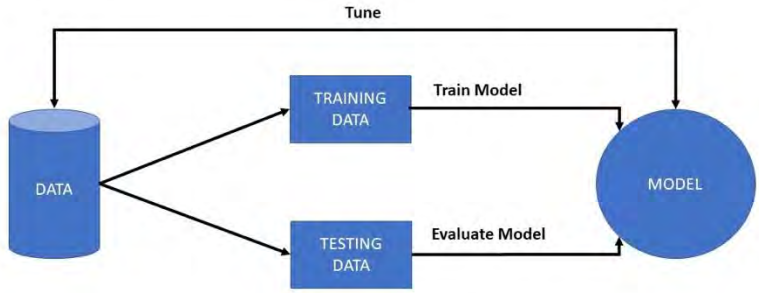



**Fig. 2.** Workflow of a supervised machine learning algorithm, Adapted form Mahesh (2019)

The training set is then used to train the model, which means that the algorithm should be able to recognize patterns in the data. So these patterns do not have to be specified by the programmer but must be learned by the algorithm from the data (Plaue, 2021). To see how successful this learning process is, test data that is not in the training set is given to the model for prediction. To optimize the model, the hyperparameters of the model can also be adjusted or tuned (see Fig. 2). These parameters serve as weights for the algorithm and must be selected by the human beforehand (Hirschle, 2022). The selection of the appropriate algorithm for a problem can be described as an iterative process, as different models are tested and then the model with the best performance is selected (Krüger & Krell, 2020). Typical supervised learning algorithms include Support Vector Machines, Decision Trees and Naive Bayes (Mahesh, 2019).

Since in a supervised machine learning approach the labels of the test data set are already known, various key indicators and metrics can be calculated to measure the performance of the model. The simplest metric for a classification problem is accuracy, which divides the correctly classified data by the total number of data (Bonaccorso, 2017). However, it is often useful to distinguish between the types of misclassification and to display them in the confusion matrix (see Fig. 3). From this the precision ($\frac{\text{True Positive}}{\text{False Positive + True Positive}}$), the recall ($\frac{\text{True Positive}}{\text{False Negative + True Positive}}$) or the harmonic mean of both the F1-score can be calculated (Ertel, 2021).

| | **Actually Positive** | **Actually Negative** |
|---|---|---|
| **Predicted Positive** | True Positive (TP) | False Positive (FP) |
| **Predicted Negative** | False Negative (FN) | True Negative (TN) |

**Fig. 3.** Confusion matrix for a binary classifier, Adapted from Ertel (2021)

Another method of measuring the agreement between humans and computers is Cohen's $\kappa$ which takes into account random agreement between human and computer-based results (Krüger & Krell, 2020). Here, values of $\kappa > .20$ are considered as slight

agreement, $.20 < \kappa < .40$ as fair agreement $.40 < \kappa < .60$ as moderate agreement $.60 < \kappa < .80$ as substantial agreement and $.80 < \kappa < 1.00$ as almost perfect agreement (Landis & Koch, 1977).

In science education research, the potential of Machine Learning to evaluate complex data and make accurate conclusions has also been recognized for different assessments (Zhai et al., 2021). For example, there has been research on the computer-assisted classification of written reflections by physics teachers in the preservice phase (Wulff et al., 2020), on the automatic coding of short text responses (Zehner et al., 2015) or on automated feedback on students' scientific argumentation (Zhu et al., 2017). Nevertheless, Zhai and colleagues (2020) were able to find research gaps in their review: The aim should not only be to enable immediate feedback but also to focus on the interpretability of machine learning results and how teachers and students deal with them. For example, studying how science teachers make instructional decisions on the fly based on the automated scores towards different types of science learning activities is lacking in the literature, but is essential to make better sense of and utilization of the ML-based science assessment (Zhai et al., 2020). The more autonomous Machine Learning models are used in science education contexts, such as in the case of the classification of student responses independently, the more important it is to inspect the results critically (Krüger & Krell, 2020).

## 3. Research questions

The theoretical background shows that formative assessment is one of the best ways to optimize learning and that concept maps are a good source of diagnostic information. However, detailed analysis is very time-consuming and thus timely feedback is not possible in everyday school life. Machine Learning could be a help for this problem, as it can quickly provide accurate evaluation and results. For these reasons, this paper tests a new evaluation approach for concept maps using Machine Learning techniques. The aim is to achieve a fast and accurate evaluation so that teachers and students can use the resulting information for a formative assessment. In order to achieve this goal, this paper focuses on the analysis of the first machine learning results and their didactic implementation in the classroom. This leads to the following research questions:

***RQ1:*** What is the level of agreement in the evaluation of a concept map between human and computer-generated ratings?

***RQ2:*** Can the developed algorithm be used for formative assessment?

## 4. Design of the study

A concept map on the topic of mechanics was developed for this study. The concept map consists of 11 central concepts of mechanics, such as uniform motion, speed or force, and different relationships between the concepts. In total, the concept map contains 19 different propositions (see Fig. 4).

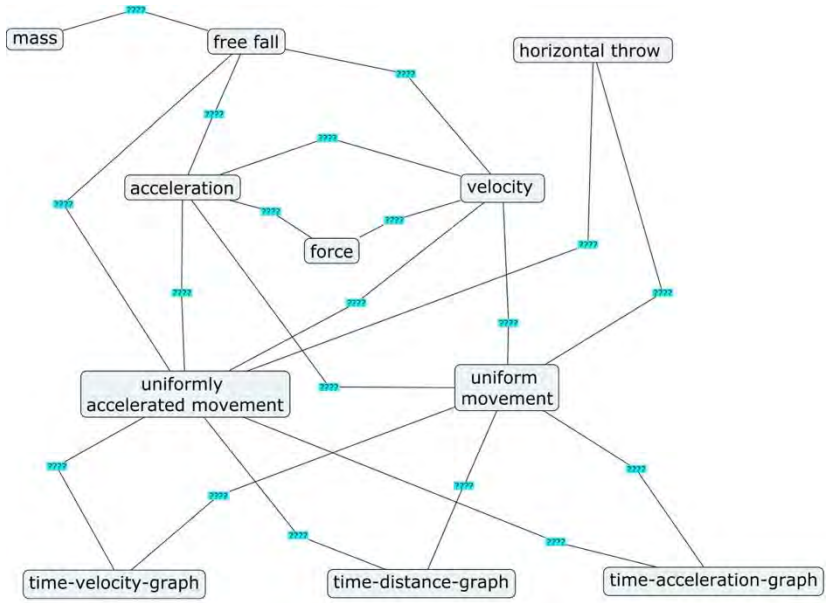



**Fig. 4.** Concept map on the topic of mechanics (translated from German)

The concept map software IHMC CMap Tool (Cañas et al., 2004) was used for the creation of the concept map and also for later editing by the students. The concept map was designed to be a fill-in-the-map task. This means that both the concepts and the connecting lines between them have already been generated. Therefore, the students only had to fill in the linking words in the blue boxes with no specific instructions on how to formulate the linking words (e.g., numbers, formulas, short sentences, etc.). A total of 45 minutes was planned for the study, with a short introduction beforehand and a questionnaire at the end. The actual working time, therefore, corresponded to about 30 minutes. A total of 230 (average age 16.9 years; female 49%, male 48%, diverse 3%) students from 14 different 11th grades took part in the study.

This rather highly directed format was chosen for two reasons: First, it was not possible to assess in advance whether the students had a certain amount of experience in working with concept maps since a completely open task format can be cognitively challenging without prior experience. Secondly, comparable concept maps should be created to enable a valid evaluation. For this reason, the highly directed format was chosen, forming a compromise between didactic interest and automatic evaluation. The students' concept maps were evaluated by two different raters. For this purpose, the concept maps were divided into individual propositions. The aim was to achieve a more detailed analysis than a pure dichotomous evaluation or to look only at the graph's theoretical aspects. For this reason, a general rating scheme was developed, which consists of four different categories: one category for (physically) wrong answers and three categories for correct answers (see Table 1). The scheme is based on the work of Fischler and Peuckert (2000) and Friege (2001) and has been adapted for this research. In order to achieve a high level of agreement between the two raters, coding examples were developed for each proposition. Thus, an agreement between the raters of 87% and a Cohen's $\kappa$ of .83 could be achieved.

**Table 1**
Rating scheme

| Category | Description | Keywords | Proportion |
|---|---|---|---|
| A | Wrong answers | | 38% |
| B | Superficial answers | "*is a*", "*has*", "*needs xy to calculate*", "*changes*" | 16% |
| C | Simple but more directed answers | "*increases/decreases*", "*x depends on y*" | 24% |
| D | Detailed answers | "*increases linearly/exponentially*", *"... proportional to..."* | 22% |

For a valid evaluation, all answers that were duplicated were deleted from the data set to avoid the later machine learning model recognising the answer again. Thus, the data set was reduced to a total of 2315 student responses, with the largest proportion being in the "wrong answers" category at 38% (see Table 1). The remaining answers were then used to develop different machine-learning classifiers.

## 5. Key findings

### *5.1. Machine learning model*

As described in section 2.3, a supervised Machine Learning approach was chosen. For this purpose, the individual propositions were evaluated by two raters (see section 4), which serve as the basis and dataset for the Machine Learning model. A 10-fold stratified cross-validation strategy (Pedregosa et al., 2011) was used to split the data into training and test data. The data set is divided into 10 equally sized blocks. One of the blocks is randomly selected for testing and the remaining nine blocks are used to train the model. On the one hand, this gives more answers for testing the algorithm and, on the other hand, it creates an even distribution of rating classes. The Machine Learning approach was selected so that an algorithm was trained on the entire data set and not, for example, on individual propositions. Firstly, the response frequencies of the propositions are relatively small (average answers per proposition) and secondly, sentence embeddings were chosen to encode the propositions and obtain a semantically meaningful representation of them (Reimers & Gurevych, 2019).

As described in section 2.3 the process of choosing the algorithm was iterative, which means that different algorithms were tested. Among others, typical classification algorithms were tested, such as Multilayer Perceptron, Random Forest Classifier, or k-nearest Neighbors Classifier (Pedregosa et al., 2011). In the end, an Support Vector Machine Classifier (SVC) (Noble, 2006) provided the best results. For comparison, a so-called dummy classifier was built, which makes predictions without learning anything, but only based on the distribution of the rating categories in the training set (Pedregosa et al., 2011).

As can be seen in Table 2, the SVC provides a high accuracy of 80%. This means that 1,852 out of 2315 student responses were classified in the correct rating category. Since accuracy is not the best metric to use with an unbalanced data set, it is always useful to look at other metrics as well. However, both the weighted F1 score of 0.80 and Cohen's $\kappa$ of 0.73 also provide good agreement scores. The comparison with the dummy classifier also clearly shows that the trained Machine Learning model performs significantly better than a model that decides purely on the basis of statistical dependencies. It can be shown

that the model was able to classify the student answers into the same category as the human raters, although it did not quite reach the level of agreement that could be achieved between the two human raters (see Table 2).

**Table 2**
Overall results of the SVC and the dummy classifier

| Model | Weighted F1 | Accuracy | Cohen's $\kappa$ |
|---|---|---|---|
| SVC | 0.80 | 80% | .73 |
| Dummy | 0.26 | 26% | 0 |
| Human | | | .83 |

Table 3 shows that the SVC was able to achieve high values for precision and recall for each rating category A – D. This is of course also reflected in the F1 values. Only the results for rating category B dropped off slightly. One possible reason could be the small number of answers that were rated by the two human raters in this category in the entire data set (cf. Table 1). The more answers are available, the easier it is for the model to recognize patterns.

**Table 3**
Precision, recall and F1 of the SVC for the four rating categories

| Category | Precision | Recall | F1 |
|---|---|---|---|
| A | .80 | .83 | .82 |
| B | .76 | .74 | .75 |
| C | .83 | .80 | .81 |
| D | .81 | .81 | .81 |

### *5.2. Didactic implementation*

This work aims to develop a Machine Learning model that is able to evaluate the propositions of a concept map in mechanics and to integrate the resulting information into a formative assessment. The results from 5.1 have shown that the SVC is able to evaluate the student concept maps with a high computer-human agreement. Seeing these results in the context of the five key strategies of formative assessment from Wiliam and Thompson (2008), using the model can certainly be helpful in steps 2 and 3 (see section 2.1) as it provides diagnostic information. Since this is a formative assessment and not a summative one, a certain number of errors can also be tolerated, since in this case, no grades are to be given. The model and the resulting automatic evaluation can thus definitely be able to relieve the teachers in their workload. Once the students' concept maps are available, the algorithm takes less than a few seconds to fully evaluate the students' maps. This can save time in everyday school life. In addition, the teacher is able to provide highly individualized feedback for all students. Once the model has been trained, it does not matter whether 5, 25 or 50 concept maps are to be evaluated. The teacher can therefore decide whether to conduct a formative assessment at the class or individual level, depending on the specific application.

One disadvantage is that the model is only trained for this particular concept map. This means that a concept map from another subject area, such as electricity, cannot be analyzed so easily with this algorithm. Also, the closed concept map format has to be considered in the results. This format was deliberately chosen (see section 2.3), although a

more open format might have more potential for formative assessment. The results show that such a Machine Learning approach can work for such a closed format. It is not possible to say whether equally good performance can be achieved for more open-ended tasks. However, a look at the student responses shows that many of the answers are independent of the proposition and the context: "*free fall has an acceleration*", "*speed increases during free fall*", and "*horizontal throw is a uniformly accelerated motion*". Therefore, it can be assumed that a different algorithm may well be created for a more open concept map. However, this is a starting point for further research.

For implementation in everyday school life, the errors or the type of error must also be critically analyzed. Since the algorithm developed is to be the basis for a feedback tool, different aspects have to be taken into consideration depending on the use case. The results give hope for a tool that can be used by teachers. They get a quick overview and can provide feedback to their students. They are in control at all times and can correct or even reject wrongly classified answers if necessary. A sensitivity as high as possible, i.e., a high recall, would be desirable here. The situation is different if the feedback tool interacts directly with students. In this case, high precision would be more appropriate, as this would mean that the proportion of student answers that were classified as correct by the model would actually be very high. This could prevent conflicts between the validity of the model and the teacher. But regardless of how the intelligent methods are used in the classroom, all involved people must learn how to deal with the interpretation and the outcome of the data. Similar to fake news or the verification of sources, all results must be critically analyzed so that meaningful embedding in the classroom can take place. Of course, this requires further competencies to be learned by teachers and learners.

## 6. Conclusion and future work

The paper presents an intelligent Machine Learning approach for the analysis of concept maps. The results show that the developed approach, which is based on a Support Vector Machine classifier, can quickly evaluate the student proposition of a mechanics concept map with a very good performance. The paper only dealt with German language concept maps, but it is possible to transfer the results to maps in other languages, such as English, since similar methods for generating features are also available in other languages. Furthermore, the paper shows that interdisciplinary cooperation between Computer Science and Science Education can not only provide new insights but also develop novel and beneficial methods for teaching and learning.

In a follow-up study, the tested and developed machine learning model will be rolled out in German schools in the next school year. In the first step, the algorithm and the concept map will be used at the beginning of the mechanics unit. With the help of automatic evaluation, teachers should get a quick and easy overview of the possible weaknesses and strengths of their students. In this way, they can design the rest of the unit from a more well-rounded perspective and adapt it to the prior knowledge of their students. In the second step, the same concept map is to be used again shortly before the exam. This time, the algorithm should on the one hand provide the students with direct hints and represent their changes to the first concept map. On the other hand, the teachers will receive renewed feedback on the students' level of performance. Due to the feedback categories developed, teachers and students get a much more accurate insight into the current state of knowledge instead of a simple consideration of right and wrong answers. How the teachers and

students will ultimately deal with the provided information and feedback will be exciting questions for this follow-up project.

## Author Statement

The authors declare that there is no conflict of interest.

## Acknowledgements

This work has been partly supported by the Ministry of Science and Education of Lower Saxony, Germany, through the Graduate training network "LernMINT: Data-assisted classroom teaching in the STEM subjects" (project no. 51410078).

## ORCID

*Tom Bleckmann* https://orcid.org/0009-0009-3186-6476

*Gunnar Friege* https://orcid.org/0000-0003-3878-9230

## References

Anohina, A., & Grundspenkis, J. (2009). Scoring concept maps. In *Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing* (pp. 1–6). ACM Press. https://doi.org/10.1145/1731740.1731824

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education, 85*(5), 536–553. https://doi.org/10.1002/sce.1022.abs

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Bonaccorso, G. (2017). *Machine learning algorithms*. Packt.

Campbell, L. O. (2022). Learner development through serial concept mapping. *Knowledge Management & E-Learning, 14*(2), 170–185. https://doi.org/10.34105/j.kmel.2022.14.010

Cañas, A. J., Hill, G., Carff, R., Suri, N., Lott, J., Gómez, G., … Carvajal, R. (2004, September). CmapTools: A knowledge modelling and sharing environment. In *Proceedings of the First International Conference on Concept Mapping* (CMC 2004) (pp. 13–20). Dirección de Publicaciones de la Universidad Pública de Navarra. Retrieved from https://thomaseskridge.com/assets/pdf/Canas-2004.pdf

Cañas, A. J., Novak, J. D., & Reiska, P. (2012). Freedom vs. restriction of content and structure during concept mapping – Possibilities and limitations for construction and assessment. In *Proceedings of the Fifth International Conference on Concept Mapping* (pp. 247–257). University of Malta. Retrieved from https://www.researchgate.net/profile/Alberto_Canas/publication/248392782_Freedom_vs_Restriction_of_Content_and_Structure_during_Concept_Mapping_-_Possibilities_and_Limitations_for_Construction_and_Assessment/links/00b7d51ded8390185c000000.pdf

Ertel, W. (2021). *Grundkurs künstliche intelligenz*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-32075-1

Fischler, H., & Peuckert, J. (2000). *Concept mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie*. Logos Berlin (Verlag).

Friege, G. (2001). *Wissen und Problemlösen: Eine empirische Untersuchung des wissenszentrierten Problemlösens im Gebiet der Elektrizitätslehre auf der Grundlage des Experten-Novizen-Vergleichs*. Logos Berlin (Verlag).

Hartmeyer, R., Stevenson, M. P., & Bentsen, P. (2018). A systematic review of concept mapping-based formative assessment processes in primary and secondary science education. *Assessment in Education: Principles, Policy & Practice, 25*(6), 598–619. https://doi.org/10.1080/0969594X.2017.1377685

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/b94608

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. Retrieved from https://apprendre.auf.org/wp-content/opera/13-BF-References-et-biblio-RPT-2014/Visible%20Learning_A%20synthesis%20or%20over%20800%20Meta-analyses%20Relating%20to%20Achievement_Hattie%20J%202009%20...pdf

Hay, D., Kinchin, I., & Lygo-Baker, S. (2008). Making learning visible: The role of concept mapping in higher education. *Studies in Higher Education, 33*(3), 295–311. https://doi.org/10.1080/03075070802049251

Hirschle, J. (2022). *Deep natural language processing: Einstieg in word embedding, sequence-to-sequence-modelle und transformer mit python*. Hanser.

Hunt, E., & Pellegrino, J. (2002). Issues, examples, and challenges in formative assessment. *New Directions for Teaching and Learning, 2002*(89), 73–85. https://doi.org/10.1002/tl.48

Krüger, D., & Krell, M. (2020). Maschinelles lernen mit aussagen zur modellkompetenz. *Zeitschrift Für Didaktik Der Naturwissenschaften, 26*, 157–172. https://doi.org/10.1007/s40573-020-00118-7

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Ley, S. L. (2015). *Concept maps als diagnoseinstrument im physikunterricht und deren auswirkung auf die diagnosegenauigkeit von physiklehrkräften*. Retrieved from https://duepublico.uni-due.de/servlets/DocumentServlet?id=38141

Llinás, J. G., Macías, F. S., & Márquez, L. M. T. (2020). The use of concept maps as an assessment tool in physics classes: Can one use concept maps for quantitative evaluations? *Research in Science Education, 50*(5), 1789–1804. https://doi.org/10.1007/s11165-018-9753-4

Mahesh, B. (2019). Machine learning algorithms – A review. *International Journal of Science and Research,* 9(1), 381–386.

Maier, U. (2010). Formative assessment – Ein erfolgversprechendes konzept zur reform von unterricht und leistungsmessung?. *Zeitschrift Für Erziehungswissenschaft, 13*(2), 293–308. https://doi.org/10.1007/s11618-010-0124-9

Müller, A. C., & Guido, S. (2017). *Einführung in machine learning mit Python: Praxiswissen data science*. O'Reilly.

Noble, W. S. (2006). What is a support vector machine?. *Nature Biotechnology, 24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173469

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pfannstiel, M. A. (2022). *Künstliche intelligenz im gesundheitswesen*. Springer. https://doi.org/10.1007/978-3-658-33597-7

Plaue, M. (2021). *Data science*. Springer. https://doi.org/10.1007/978-3-662-63489-9

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*(1), 4–13. https://doi.org/10.1002/bs.3830280103

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-3

Richter, S. (2019). *Statistisches und maschinelles lernen*. Springer. https://doi.org/10.1007/978-3-662-59354-7

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33*(6), 569–600. https://doi.org/10.1002/(SICI)1098-2736(199608)33:6<569::AID-TEA1>3.0.CO;2-M

Schütze, B., Souvignier, E., & Hasselhorn, M. (2018). Stichwort – Formatives assessment. *Zeitschrift Für Erziehungswissenschaft, 21*, 697–715. https://doi.org/10.1007/s11618-018-0838-7

Souvignier, E., & Hasselhorn, M. (2018). Formatives assessment. *Zeitschrift Für Erziehungswissenschaft, 21*, 693–696. https://doi.org/10.1007/s11618-018-0839-6

Stracke, I. (2004). *Einsatz computerbasierter concept maps zur wissensdiagnose in der chemie: Empirische untersuchungen am beispiel des chemischen gleichgewichts*. Waxmann.

Wiliam, D., & Leahy, S. (2007). A theoretical foundation for formative assessment. In J. H. McMillan (Ed.), *Formative Classroom Assessment: Theory into Practice* (pp. 29–42). Teachers College Press.

Wiliam, D., & Thompson, M. (2008). Integrating assessment with learning: What will it take to make it work?. In C. A. Dwyer (Ed.), *The Future of Assessment: Shaping teaching and learning* (pp. 53–82). Routledge. https://doi.org/10.4324/9781315086545-3

Wolf, N. (2014). Formative leistungsmessung im naturwissenschaftlichen unterricht. Doctoral dissertation, Pädagogische Hochschule Schwäbisch Gmünd, Germany. Retrieved from https://phsg.bsz-bw.de/frontdoor/deliver/index/docId/14/file/Dissertation_FORMAL_Nicole_Wolf_Digital_Ohne_externe_Links.pdf

Wulff, P., Buschhüter, D., Westphal, A., Nowak, A., Becker, L., Robalino, H., … Borowski, A. (2020). Computer-based classification of preservice physics teachers' written reflections. *Journal of Science Education and Technology, 30*(1), 1–15. https://doi.org/10.1007/s10956-020-09865-1

Zehner, F., Sälzer, C., & Goldhammer, F. (2015). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement, 76*(2), 280–303. https://doi.org/10.1177/0013164415590022

Zhai, X., Krajcik, J., & Pellegrino, J. (2021). On the validity of machine learning-based next generation science assessments: A validity inferential network. *Journal of Science Education and Technology, 30*, 298–312. https://doi.org/10.1007/s10956-020-09879-9

Zhai, X., Yin, Y., Pellegrino, J., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

Zhu, M., Lee, H. S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education, 39*(12), 1648–1668. https://doi.org/10.1080/09500693.2017.1347303