Informatics in Education, 2023, Vol. 00, No. 00, –

DOI: 10.15388/infedu.2023.11
ACCEPTED VERSION (WITHOUT THE JOURNAL LAYOUT).

Assessing the Learning of Machine Learning in K-12: A Ten-Year Systematic Mapping

Marcelo Fernando RAUBER[1,2], Christiane GRESSE VON WANGENHEIM[1]

[1]*Graduate Program in Computer Science, Department of Informatics and Statistics, Federal University of Santa Catarina, Florianópolis/SC, Brazil*
[2]*Instituto Federal Catarinense, Campus Camboriú/SC, Brazil*
email: marcelo.rauber@ifc.edu.br, c.wangenheim@ufsc.br

Received: February 2022

Abstract. Although Machine Learning (ML) has already become part of our daily lives, few are familiar with this technology. Thus, in order to help students to understand ML, its potential, and limitations and to empower them to become creators of intelligent solutions, diverse courses for teaching ML in K-12 have emerged. Yet, a question less considered is how to assess the learning of ML. Therefore, we performed a systematic mapping identifying 27 instructional units, which also present a quantitative assessment of the students' learning. The simplest assessments range from quizzes to performance-based assessments assessing the learning of basic ML concepts, approaches, and in some cases ethical issues and the impact of ML on lower cognitive levels. Feedback is mostly limited to the indication of the correctness of the answers and only a few assessments are automated. These results indicate a need for more rigorous and comprehensive research in this area.

Keywords: Assessment, Computing education, Machine Learning, K-12.

1. Introduction

Machine Learning (ML) is a technology that allows computers to learn directly from examples, data, and experiences (Royal Society, 2017). ML applications are today being applied in everything, such as recommendation systems, personal assistants with voice recognition, image recognition for disease diagnosis, and pattern detection for example for finding unusual financial activities (Royal Society, 2017). The Artificial Intelligence (AI) revolution requires citizens to be prepared for this new reality as it shifts the demand to technological, social, emotional, and high cognitive skills (World Economic Forum, 2021). In this context, it also becomes important to popularize ML concepts and techniques starting in K-12. Several initiatives have emerged aiming at the introduction of AI/ML education in K-12, including AI4K12 (2020) aiming at developing curricular guidelines for teaching AI, including also the teaching of ML. In practice, the teaching of ML is currently being introduced mostly by extracurricular courses typically addressing basic concepts of ML and neural networks, as well as learning algorithms, ethical issues among others (Marques *et al*., 2020).
Several reviews have already elicited the state of the art regarding the teaching of ML in K-12, mapping content, instructional strategies, and technology (Marques *et al*., 2020;

Cheng, 2021; Kandlhofer and Steinbauer, 2021; Queiroz *et al.*, 2021). Other studies synthesize AI learning experiences in K-12 (Zhou *et al*., 2020), while Sanusi and Oyelere (2020) and Tedre *et al*. (2020) review pedagogical and technological strategies for AI education, and Steinbauer *et al*. (2021) discuss teaching AI with respect to education modus, level, concepts and tools. Yet, covering the wider scope of AI education, fewer details are presented specifically on teaching ML.

Furthermore, an issue less regarded so far is the assessment of the students' learning of ML, considering assessments to be an essential part of an effective learning process (Hattie and Timperley, 2007). Assessments guide student learning and provide feedback to the students by letting them know their level of performance on a task, how their performance relates to good performance and what to do to close the gap between those (Sadler, 1989; Ihantola *et al*., 2010; Stegeman *et al*., 2016). The assessment also helps teachers to determine the extent to which the learning goals are being met (Ihantola *et al*., 2010).

Yet, despite the many efforts aimed at dealing with the assessment of computational thinking (Grover and Pea, 2013; Grover *et al*., 2015) and reviews on the assessment of the learning of computational thinking (Cutumisu *et al.*, 2019; Da Cruz Alves *et al.*, 2019; Tikva and Tambouris, 2021), so far there seems to be a lack of research on the assessment of the learning of ML. In this regard, the main contribution of this systematic mapping is the identification, characterization, and analysis of models for the assessment of the student's learning of ML in the context of K-12.

2. Background

2.1 Teaching ML in K-12

Although there have been some historical AI teaching initiatives in schools from the 1970s (Papert and Solomon, 1971; Kahn, 1977) and, involving neural networks, in the 1990s (Bemley, 1999), there has been a rapid expansion of computing education in K-12 worldwide over the last few years. More recently, initiatives for teaching AI/ML in K-12 are emerging (Marques *et al*., 2020), with some countries such as China mandating that all high school students learn about AI (Jing, 2018). Furthermore, existing computing curriculum guidelines such as the CSTA K-12 Computer Science Framework (CSTA, 2017) are being extended focusing specifically on AI through curricular guidelines being developed by the AI for K-12 Working Group (AI4K12). To frame these guidelines, "big ideas" in AI that every student should know are defined (Touretzky *et al*., 2019a; Touretzky *et al*., 2019b) (Fig. 1).

While AI is the science and engineering of making intelligent machines that have the ability to achieve goals the way humans do, ML is a subfield of AI that deals with the field of study of giving computers the ability to learn without being explicitly programmed (by building a mathematical/statistical model based on collected data) (Mitchell, 1997). In this context "Big Idea 3" of the curricular guidelines refers to ML, expecting the students to learn (Touretzky *et al*., 2019b):

- What is learning?

- Approaches to ML (e.g., regression, instance-based, bayesian algorithms, support vector machines, decision trees, clustering, artificial neural networks (ANN))
- Types of learning algorithms
- Fundamentals of neural networks
- Types of neural network architectures
- How training data influences learning
- Limitations of ML

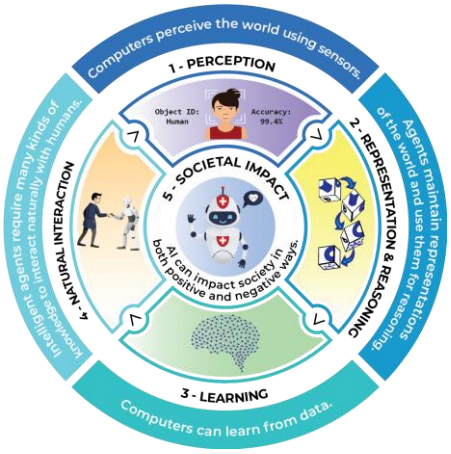


Fig. 1. Big Ideas for Teaching Artificial Intelligence in K-12 (Touretzky *et al*., 2019)

Following the trend of adopting active learning methodologies for teaching computational thinking, ML concepts are taught through a mix of expositive lessons, demonstrations, but mainly through hands-on exercises and/or projects. Most of the current ML courses target beginners, applying the Use-Modify-Create cycle (Lee *et al.*, 2011), providing a scaffolding that allows the student to first inspect and manipulate pre-defined examples, then to modify until on the Create stage, encouraging students to develop their own ML projects. Adopting active learning strategies, on any of these stages, students typically create ML models as a result of their learning.

2.2 Assessment of the student learning

Assessment is a social-historical construct inherent in the teaching-learning process and has evolved and transformed with the pace of the process and the scientific theories involved. The assessment of student learning can be defined as (Huba and Freed, 2000):

> *"Assessment is the process of collecting and analyzing information from a variety of sources in order to develop a deep understanding of what students know, understand, and can do with their knowledge as a result of their educational experiences."*

The major purpose of assessment is to determine at what level the learning objectives have been met. Thus, assessments are directly related to the learning objectives, which provide a framework to create adequate ways to assess student learning (Seel *et al*., 2017; Morrison *et al*., 2019).

There exist diverse types of assessments, which can be classified with regard to different perspectives. And, although there does not exist a consensus on the classification of assessment methods, in this study, we focus on quantitative assessments, in which variables are systematically measured that describe results numerically allowing the analysis of their reliability and validity (Rothwell, 2016; Creswell and Creswell, 2018).

As part of assessments, the data collection strategy refers to how data is collected in order to analyze the students' learning. There exist several methods for collecting data as part of the quantitative assessment (Table 1). These methods commonly use test items or other artifacts in order to analyze the students' learning. A definition of an item is given by Osterlind (1989):

> *"A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form for answering; and, it is intended to yield a response from an examinee from which performance in some psychological construct (such as an knowledge, ability, predisposition, or trait) may be inferred."*

Table 1. Examples of quantitative assessment methods and item formats

Assessment method	**Items format**	**Examples of item**
Test/Quiz	Selected-response	- Multiple-choice - Binary choice (True/False) - Matching - Interpretative
	Brief-constructed response	- Short answer - Completion - Label a Diagram
Performance-based assessment	Artifact (e.g., trained neural network, intelligent mobile application)	- Rubric - Gamification
Student self-assessment	Self-report inventories Self-questionnaire	- Multiple-choice - Likert-scale - Short answer
Observation	Observer annotations (formal or informal)	- Observer annotations - Rubric - Checklist
Interview	Oral questioning - Informal questioning - Structured interview	- Rubric

Source: adapted from McMillan (2018) and Rothwell (2016)

Tests and quizzes are typically used to assess the acquisition of competencies, especially knowledge. In order to establish a baseline, it is common to use pre-tests to determine knowledge before the learning intervention, and afterwards a post-test which allows comparing the performance. Selected-response formats present items with two or more possible answers, while constructed-response formats leave the student to create their own response. Performance-based assessments assess the learning based on a more extensive and elaborate answer or artifact created by the student, typically adopting rubrics (Morrison *et al*, 2019) to systematically assess the achievement of learning objectives based on the created results defining a scale covering different levels of competence (McMillan, 2018). Student self-assessment refers to the assessment of the learning by the students themselves based on their own perception. Typically questionnaires are used to collect this kind of information, their attitudes, and beliefs. Data can also be collected through the observation of learners to determine whether their behaviors, performance, interactions, or other variables have changed or improved. The observations are annotated either formally or informally, using, for example, rubrics or checklists. Interviews or questioning are conversations that aim to gather feedback and input directly from the students through a structured or unstructured approach.

Learning levels. Assessments can also be defined in accordance with the specific levels of learning to be achieved, for example, based on Bloom's Taxonomy (Bloom *et al.*, 1956; Anderson and Krathwohl, 2001). In Bloom's revised taxonomy, the cognitive process dimension presents six progressive dimensions of increasing cognitive complexity (Anderson and Krathwohl, 2001) (Fig. 2).

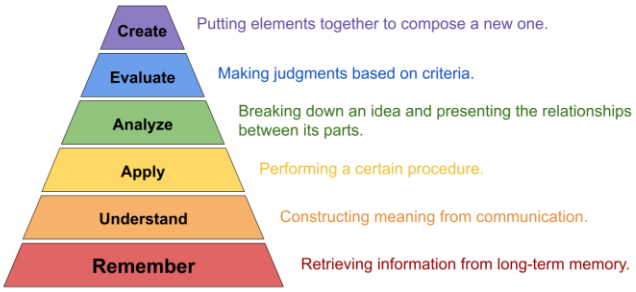


Fig. 2. Bloom's revised taxonomy

Feedback. Assessment primarily aims at providing feedback to the students in order to guide her/his learning process to achieve learning objectives (Frey and Fisher, 2011, McMillan, 2018). Feedback seeks to stimulate and provoke student reflection on their responses to an assessment by providing specific information that facilitates and directs the learning of a topic. It can be delivered by an instructor or in an automated way (McMillan, 2018; Morrison *et al*., 2019). Automation has the potential to reduce the workload of teachers (Ala-Mutka and Järvinen, 2004), helps to ensure consistency and accuracy of assessment results, and to eliminate bias (Ala-Mutka, 2005; Romli *et al*., 2010). Feedback should be clearly related to the learning objective identifying gaps in the students' learning, while at the same time providing informative and constructive support

on how to improve the learning. In order to be effective, it should also be relevant and specific to the students' answers/artifacts/behaviors and given in a timely manner accompanying the students' learning process (Seel *et al*., 2017; Morrison *et al*., 2019).

Reliability and validity of assessments. In order to be effective, it is important to evaluate the quality of assessments in terms of their reliability and validity (Morrison *et al*., 2019). Reliability "refers to the consistency of assessment scores" provided by an instrument (Moskal and Leydens, 2000). Although there is no simple and direct way to measure reliability, attributes of homogeneity, stability, and equivalence are considered (Heale and Twycross, 2015). Typically, homogeneity is analyzed, also called internal consistency, which indicates the consistency between items of a data collection instrument, while equivalence (inter-rater reliability) refers to the consistency between different raters, and stability (intra-rater reliability) to consistency over time of the same assessor (Kimberlin and Winterstein, 2008). Reliability is usually analyzed with respect to internal consistency using Cronbach's alpha coefficient analysis (Cronbach, 1951) or interrater reliability using Cohen's kappa (k) coefficient (Cohen, 1960).

The evaluation of validity on the other hand is "the process of accumulating evidence that supports the appropriateness of inferences made from student responses for specified assessment uses" (Moskal and Leydens, 2000). The evaluation of validity can be related to content, construct and/or criterion (Moskal and Leydens, 2000, DeVellis, 2017). Content validity refers to the level at which a student response reflects their knowledge in the area of interest and the appropriateness/comprehensiveness of the assessment instruments. Construct validity refers to the degree at which the construct of thought is internal to each individual and is only partially displayed as a result or explanation through an assessment. Criterion validity refers to the level at which the results of the evaluation correspond to a current or future event, or how they can be generalized to more relevant activities. Typically, validity analysis is done through Factor Analysis (Glorfeld, 1995), Correlation Matrix (DeVellis, 2017), or Item Response Theory (DeVellis, 2017).

3. Definition and Execution of the Systematic Mapping Study

In order to elicit the state of the art and practice on how the students' learning of ML is assessed in the context of K-12 education, we conducted a systematic mapping study following the procedures defined by Petersen and colleagues (Petersen *et al*., 2008, Petersen *et al*., 2015). Aiming at the analysis of objective measurements of the students' learning, we focus specifically on quantitative assessments models.

3.1 Definition of the review protocol

The main research question is "What quantitative models exist for the assessment of student learning of ML in K-12"?

This research question is refined into the following analysis questions:

AQ1. Which instructional units aimed at teaching ML in K-12 exist that also include a quantitative assessment of the students´ learning?

AQ2. What are the characteristics of these assessments in terms of learning level, learning objectives, content, and type?

AQ3. What instructional feedback is presented in what way to those involved, and what assessments are automated and how?
AQ4. How were the assessments designed and evaluated?

Data source. The search was conducted in major digital repositories in the field of Computing, including Scopus, IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, arXiv, SocArXiv, ERIC, Web of Science, and Wiley with access through the Capes Portal[1]. In addition, Google Scholar searches were conducted to complement the search, minimizing the risk of omission (Piasecki *et al.*, 2018). We also searched for relevant publications in the MIT Media Lab repository due to their research in this area.
Inclusion and exclusion criteria. We included only peer-reviewed articles in the English language published in the last 10 years, which present a form of quantitative assessment of student learning in the context of teaching ML in K-12. On the other hand, articles that do not present a quantitative assessment of student learning are excluded. We also excluded articles presenting ML assessments on other educational stages.
Quality criteria. Only articles that present substantial information in order to allow the extraction of relevant information regarding the analysis questions are considered. Other artifacts, not presenting substantial information, such as summaries or one-page abstracts, blogs, videos were excluded.
Definition of search terms: According to the research question, the search string was defined by identifying core concepts and synonyms, as shown in Table 2. The terms ML, Deep Learning, AI, and Data Science express the main concepts to be investigated and commonly appear in the literature. The terms K-12, kids, children teen, and school are commonly used in the educational context to indicate the educational level. The terms assessment, grading, learn, education, teach, course, and MOOC were used to narrow the search to focus on learning assessment.

Table 2
Search terms and synonyms/translations

Core concept	Keywords and Synonyms
Machine Learning	“Machine Learning”, “Deep Learning”, “Artificial Intelligence”, “Data science”
Assessment	assess*; grading; learn*; education; teach*; course; mooc;
K-12	K-12; kids; children; teen*; school*;

Using these keywords, a generic search string has been defined and calibrated:

```
("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR
"Data science") AND (assess* OR grading OR learn* OR education OR teach*
OR course OR mooc) AND ("K-12" OR kids OR children OR teen* OR school*)
```

[1] A web portal for access to scientific knowledge worldwide, managed by the Brazilian Ministry of Education for authorized institutions, including universities, government agencies and private companies (www.periodicos.capes.gov.br).

This generic search string has been adapted in conformance with the specific syntax of each of the repositories (Table 3). Queries were performed considering the fields title, abstract, and keywords, whenever this option was available, otherwise, only the abstract was searched. When possible, the field of knowledge was limited to Computer Science. The publication date has been limited to 2011-2021.

Table 3
Calibrated search strings for each of the repositories

Repository	Search string
ACM Digital Library	[[Abstract: "machine learning"] OR [Abstract: "deep learning"] OR [Abstract: "artificial intelligence"] OR [Abstract: "data science"]] AND [[Abstract: assess*] OR [Abstract: grading] OR [Abstract: learn*] OR [Abstract: education] OR [Abstract: teach*] OR [Abstract: course] OR [Abstract: mooc]] AND [[Abstract: "k-12"] OR [Abstract: kids] OR [Abstract: children] OR [Abstract: teen*] OR [Abstract: school*]] AND [Publication Date: (01/01/2011 TO 12/31/2021)]
ArXiv	Query: size: 50; date_range: from 2011-01-01 to 2021-12-31; classification: Computer Science (cs); include_cross_list: True; terms: AND abstract="Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data science"; AND abstract=assess* OR grading OR learn* OR education OR teach* OR course OR mooc; AND title="K-12" OR kids OR children OR teen* OR school*
ERIC	("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data science") AND (assess* OR grading OR learn* OR education OR teach* OR course OR mooc) AND ("K-12" OR kids OR children OR teen* OR school*)
IEEE Xplore	(("Abstract":"Machine Learning" OR "Abstract":"Deep Learning" OR "Abstract":"Artificial Intelligence" OR "Abstract":"Data science") AND ("Abstract":assess* OR "Abstract":grading OR "Abstract":learn* OR "Abstract":education OR "Abstract":teach* OR "Abstract":course OR "Abstract":mooc) AND ("Abstract":"K-12" OR "Abstract":kids OR "Abstract":children OR "Abstract":teen* OR "Abstract":school*))
MIT Media Lab	All listed publications were considered (http://appinventor.mit.edu/explore/research)
Science Direct	Year: 2011-2021 Title, abstract, keywords: ("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data science") ("K-12" OR kids OR children OR teen OR school) Observation: It only accepts 8 Logical operators and does not accept *. I deleted the *. I searched through the ML and K12 block (above), and on these results, I applied the filter: assess OR grading OR learn OR education OR teach OR course OR mooc
SCOPUS	(TITLE-ABS-KEY ("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data science")) AND (TITLE-ABS-KEY (assess* OR grading OR learn* OR education OR teach* OR course OR mooc)) AND (TITLE-ABS-KEY ("K-12" OR kids OR children OR teen* OR school*)) AND (PUBYEAR > 2010) AND (LIMIT-TO (SUBAREA, "COMP"))
SocArXiv	("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data science") AND (assess* OR grading OR learn* OR education OR teach* OR course OR mooc) AND ("K-12" OR kids OR children OR teen* OR school*)

SpringerLink	("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data science") AND (assess* OR grading OR learn* OR education OR teach* OR course OR mooc) AND ("K-12" OR kids OR children OR teen* OR school*)
Web of science	(AB=(("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data science") AND (assess* OR grading OR learn* OR education OR teach* OR course OR mooc) AND ("K-12" OR kids OR children OR teen* OR school*))) AND PY=(2011-2021)
WILEY	(Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data science") AND (assess* OR grading OR learn* OR education OR teach* OR course OR mooc) AND ("K-12" OR kids OR children OR teen* OR school*)" in Abstract
Google Scholar	In an anonymous tab in the browser. ("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence" OR "Data science") AND (assess* OR grading OR learn* OR education OR teach* OR course OR mooc) AND ("K-12" OR kids OR children OR teen* OR school*)

3.2 Search execution

The search was conducted from June to August 2021 by the author and reviewed by the co-author. The initial searches resulted in a total of 75.451 results (Table 4).

In the first stage, we selected potentially relevant articles based on the titles, abstracts, and keywords of the 500 most relevant articles (when available) returned as the result of the searches in each repository in accordance with the inclusion and exclusion criteria. Many artifacts have been excluded at this stage as they refer to the application of ML techniques in any field of education, e.g., the teaching of handwriting skills (Xue *et al*., 2021), autism spectrum disorders (Stevens *et al*., 2019) attention-deficit hyperactivity disorders (Tor *et al*., 2021), or dyslexia (Usman *et al*., 2021), rather than focusing on teaching ML. In addition, due to a different meaning of the term "Deep Learning" in the educational field, referring to meaningful learning by students (Fullan *et al*., 2014) rather than the one intended here as a sub-field of AI, several articles have also been excluded (e.g., Akhter *et al*., 2021; Chlap *et al*., 2021; Gao *et al*., 2021).

In the second step, we analyzed the complete articles of potentially relevant ones in accordance with the established inclusion/exclusion and quality criteria. During this step, articles presenting ML courses without the presentation of assessment were excluded, (e.g., Fry and Makar, 2021; Kahn and Winters, 2021; Lasser *et al.*, 2021). In accordance with our research objective, we also excluded articles that do not present a quantitative assessment approach, such as (Burgsteiner *et al.*, 2016; Druga and Ko, 2021; Erickson and Chen, 2021; Frischemeier *et al*., 2021).

Table 4
Number of artifacts identified per stage of selection

Base	No. of search results	No. of analyzed artifacts	No. of potentially relevant artifacts	No. of relevant artifacts
ACM Digital Library	367	367	58	9
ArXiv	40	40	5	2
ERIC	633	500	9	4
IEEE Xplore	791	500	32	4
MIT Media Lab	85	85	10	2
ScienceDirect	145	145	2	1
SCOPUS	4.130	500	73	10
SocArXiv	10	10	3	1
SpringerLink	42.742	500	12	4
Web of science	2.808	500	45	4
WILEY	5.500	500	11	0
Google Scholar	18.200	500	24	1
TOTAL	**75.451**	**4.340**	**369**	**27 (without duplicates)**

4. Data Analysis

To answer the research question, we present our findings with respect to each of the analysis questions based on the relevant information we extracted from the articles. Extraction was performed by the first author and revised and discussed with the co-author until a consensus was reached. As not necessarily all information is presented explicitly in the articles, certain characteristics were inferred based on the information available in the articles. Variations in terminology were unified in accordance with the definitions presented in section 2. Appendix 1 and 2 detail the extracted information.

4.1 Which instructional units aimed at teaching ML in K-12 exist that also include a quantitative assessment of the students´ learning?

We found a total of 27 articles describing the quantitative assessment of student learning ranging from kindergarten to high school (Table 5).

Table 5
Relevant articles

Reference	Title
(Alexandre *et al*., 2021)	Why, What and How to Help Each Citizen to Understand Artificial Intelligence?
(Chattopadhyay *et al*., 2018)	A Middle-School Case Study: Piloting A Novel Visual Privacy Themed Module for Teaching Societal and Human Security Topics Using Social Media Apps
(Dryer *et al*., 2018)	A Middle-School Module for Introducing Data-Mining, Big-Data, Ethics and Privacy Using RapidMiner and a Hollywood Theme
(Estevez *et al*., 2019)	Gentle Introduction to Artificial Intelligence for High-School Students Using Scratch
(Gresse von Wangenheim *et al*., 2020)	Machine Learning for All - Introducing Machine Learning in K-12
(Henry *et al*., 2021)	Teaching Artificial Intelligence to K-12 Through a Role-Playing Game Questioning the Intelligence Concept
(Hitron *et al*., 2019)	Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes
(Hsu *et al*., 2021)	The Effects of Applying Experiential Learning into the Conversational AI Learning Platform on Secondary School Students
(Kahn *et al*., 2018)	AI Programming by Children using Snap! Block Programming in a Developing Country
(Kandlhofer *et al*., 2016)	Artificial intelligence and computer science in education: From kindergarten to university
(Kandlhofer *et al*., 2021)	EDLRIS: A European Driving License for Robots and Intelligent Systems
(Lee *et al*., 2021)	Developing Middle School Students' AI Literacy
(Melsión *et al*., 2021)	Using Explainability to Help Children Understand Gender Bias in AI
(Mike *et al*., 2020)	Equalizing Data Science Curriculum for Computer Science Pupils
(Mobasher *et al*., 2019)	Data Science Summer Academy for Chicago Public School Students
(Ng and Chu, 2021)	Motivating Students to Learn AI Through Social Networking Sites: A Case Study in Hong Kong
(Ossovski and Brinkmeier, 2019)	Machine Learning Unplugged - Development and Evaluation of a Workshop About Machine Learning
(Priya *et al*., 2021)	ML-Quest: A Game for Introducing Machine Learning Concepts to K-12 Students
(Rodríguez-García *et al*., 2020)	LearningML: A Tool to Foster Computational Thinking Skills Through Practical Artificial Intelligence Projects
(Rodríguez-García *et al*., 2021)	Evaluation of an Online Intervention to Teach Artificial Intelligence with LearningML to 10-16-Year-Old Students
(Sakulkueakulsuk *et al*., 2018)	Kids making AI: Integrating Machine Learning, Gamification, and Social Context in STEM Education
(Shamir and Levin, 2021)	Neural Network Construction Practices in Elementary School
(Tedre *et al*., 2020) and (Vartiainen *et al*., 2020)	Machine Learning Introduces New Perspectives to Data Agency in K—12 Computing Education; Machine learning for middle-schoolers: Children as designers of machine-learning apps

(Van Brummelen *et al*., 2020)	Teaching Tech to Talk: K-12 Conversational Artificial Intelligence Literacy Curriculum and Development Tools
(Vartiainen *et al*., 2021)	Machine learning for middle schoolers: Learning through data-driven design
(Williams *et al*., 2019)	A is for Artificial Intelligence: The Impact of Artificial Intelligence Activities on Young Children's Perceptions of Robots

There is a clear trend with a considerable annual increase of publications related to teaching and assessing ML in K-12, tied to the growing importance of AI/ML, as well as the growing trend of computer science education in K-12 worldwide (Fig. 3).

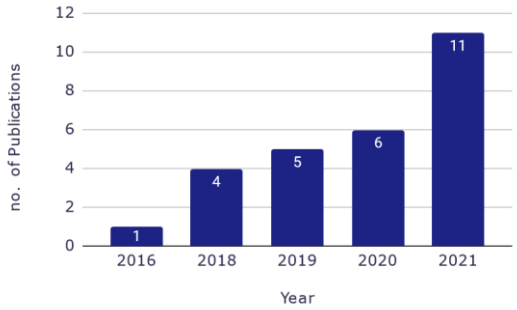


Fig. 3. Publications by year.

The majority of the identified instructional units that describe the assessment of the learning of ML target young people ranging from 12 to 14 years, while some courses focus on younger children from 10 years up as well as older ones up to 18 years. Exceptions are the courses presented by Williams *et al*. (2019) and Kandlhofer *et al*. (2016), which are the only ones aiming at kindergarten.

Most instructional units are offered in an extracurricular way, often as face-to-face workshops or summer camps. Due to the COVID-19 pandemic, several courses, which may have been designed for face-to-face classes originally, were modified or directly designed for online classes. The instructional units vary largely in terms of duration with mostly short courses from 16 minutes to 15 hours. Only five courses present longer courses with two (Kandlhofer *et al*., 2021; Mike *et al*., 2020) with more than 60 hours. Most of the instructional units present an introduction to the field of AI and ML, often through classification problems while also approaching ethical issues and the impact of ML.

Regarding ML concepts covered in the instructional units the vast majority of them initially present a theoretical introduction defining ML within the field of AI. Next, they present one or more ML approaches. It is also common to address the influence of training data on the results presented by the ML model (Lee *et al*., 2021; Alexandre *et al*., 2021; Van Brummelen *et al*., 2020; Gresse von Wangenheim *et al*., 2020; Rodríguez-García *et al*., 2020; Williams *et al*., 2019; Mobasher *et al*., 2019). Some courses address neural network fundamentals (Lee *et al*., 2021; Alexandre *et al*., 2021; Van Brummelen *et al*., 2020; Gresse von Wangenheim *et al*., 2020). Few instructional units explicitly address

the limitations of ML (Lee *et al*., 2021; Alexandre *et al*., 2021; Van Brummelen *et al*., 2020; Gresse von Wangenheim *et al*., 2020).
While some of the instructional units focus only on the levels of understanding (Chattopadhyay *et al*., 2018; Tedre *et al*., 2020; Vartiainen *et al*., 2020; Vartiainen *et al*., 2021) using diverse instructional methods (such as expositive lectures, tutorials, unplugged activities, or through a co-design process), most adopt active methodologies aiming at higher levels (apply, analyze, and create) of Bloom's taxonomy. Mapping the learning strategies on the use-modify-create cycle, we can also observe that the instructional units mostly either focus exclusively on the use level or cover the complete use-modify-create cycle. Only two courses (Kahn *et al*., 2018; Sakulkueakulsuk *et al*., 2018) cover the use-modify stages.

4.2 What are the characteristics of these assessments in terms of learning level, learning objectives, content, and type?

Most articles describe the assessment method but do not present or show a sample of the used instruments. The majority applies multiple assessment methods (Fig. 4), combining both qualitative and quantitative ones, partly due to the fact that the researchers also aim at the evaluation of the courses. In this regard, most studies report the application of an assessment before and after instruction, comparing the results. Yet in this context, assessment results are mostly presented in an accumulated way mostly for formative assessment purposes, not providing constructive feedback for the student's learning process.

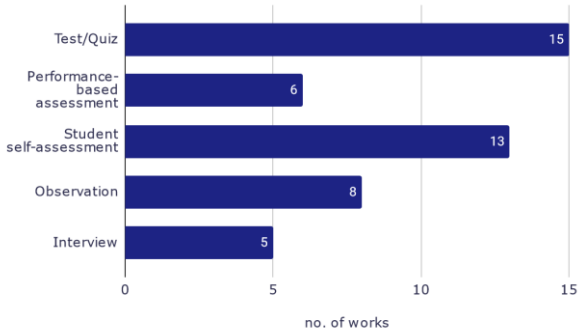


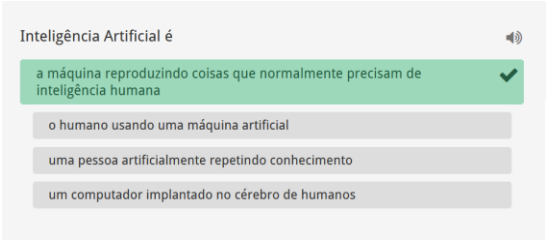
Fig. 4. Frequencies of adopted assessment methods

The content mostly assessed are basic ML concepts, ML approaches and in some cases ethical issues and the impact of ML on society. Most assessments are quite simple, aiming at lower cognitive levels (remember, understand and apply) of Bloom's Taxonomy in accordance with the introductory character of the courses currently. We also observed that although some of the instructional units cover several stages of the "use-modify-create" cycle, the assessments are not necessarily aimed at all of the covered stages, focusing mostly only on assessing learning on the use stage. This can be explained by the recent nature of the ML instructional units for K-12 and the early stage of development of the assessments. Only two articles indicate assessment on the modify stage, through a

non-automated gamification strategy giving points (based on the correct prediction of the ML models and trade operations between different groups) (Sakulkueakulsuk *et al.*, 2018) and through observation (Kahn *et al.*, 2018) conducted by 4 researchers during the learning process aiming to identify attention, interest, activity, eagerness to learn, learning atmosphere and circumstances of orderly learning. Some assessments also address ethical issues and the impact of ML on society.

Regarding the types of items used, most of the assessment methods use multiple-choice items, followed by Likert-style and short essays. Much fewer courses include performance-based assessment based on the analysis of artifacts created as a result of the learning process.

Test/Quizzes. Most studies used pre-and post-tests in order to assess knowledge acquisition and at the same time evaluate the effectiveness of the instructional unit. Some instructional units use quizzes during the learning activities. Most of these assessments have a formative character, not providing much feedback to the student, besides indicating if the question has been answered correctly or not. The most used item type is multiple-choice, followed by open-ended short essay items (Fig. 5).

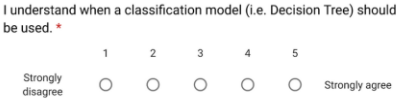


(Gresse von Wangenheim *et al.*, 2020)

(Williams *et al.*, 2019)

Fig. 5. Examples of tests/quizzes.

Student Self-Assessment. Another rather common assessment method is self-assessment, most often used to identify students' perceptions about the relevance and their learning of ML and AI topics covered by the instructional units. This kind of self-assessment is typically done using questionnaires with Likert-style scales, sometimes combined with short essay items (Fig. 6).



(Mobasher *et al.*, 2019)

Scale 1-5: How would you rate your interest in machine learning after this workshop?

Scale 1-5: Before attending this workshop, how would you rate your overall interest in the computing discipline?

(Chattopadhyay *et al.*, 2018)

Fig. 6. Examples of student's self-assessment.

Performance-based. Only six instructional units report the usage of a performance-based assessment, yet, most do not present details. Only two instructional units adopt rubrics (Fig. 7) for the performance-based assessment to analyze artifacts created by the students.

Criterion	Performance levels		
	Poor - 0 pt.	**Acceptable - 1 pt.**	**Good- 2 pt.**
Data management (LO5)			
Quantity of images	Less than 5 images per category	6 to 10 images per category	More than 10 images per category
Relevance of images	Several images are not related to the ML task (irrelevant) and/or at least one image contains unethical content (violence, nudity, etc.)	One image is irrelevant and no image containing unethical content	All images are related to the ML task and no image containing unethical content
Distribution of the dataset	Quantities of images by category vary greatly	Quantities of images by category vary little	All categories have the same quantity of images
Labeling of the images	Less than 20% of the images have been labeled correctly	Between 20% and 99% of the images have been labeled correctly	All images were labeled correctly
Data cleaning	There are several messy images (out of focus, several objects in the same image, etc.)	There is one messy image	No messy images were included in the dataset
Model training (LO6)			
Training	The model was not trained	The model was trained using standard parameters	The model was trained with adjusted parameters (e.g., epoch, batch size, learning rate)
Interpretation of performance (LO7)			
Tests with new objects	No object tested	1-2 object tested	More than 2 objects tested
Interpretation of tests	Wrong interpretation	(Not applicable)	Correct interpretation
Accuracy interpretation	Categories with low accuracy are not identified correctly and incorrect interpretation with respect to the model	Correctly identified categories with low accuracy, but incorrect interpretation with respect to the model	Correctly identified categories with low accuracy and the consequent interpretation with respect to the model
Interpretation of the confusion matrix	Misclassifications are not identified correctly and incorrect interpretation with respect to the model	Correctly identified misclassifications, but incorrect interpretation with respect to the model	Correctly identified misclassification and the consequent interpretation with respect to the model
Adjustments /Improvements made	No new development iterations have been reported	A new iteration with changes to the *dataset* and/or training parameters has been reported	Several iterations with changes to the *dataset* and/or training parameters were reported

(Gresse von Wangenheim *et al.*, 2020)

Covered topics		**Acceptable** (One point each)
Correctly programmed the neuron		
Programmed a tutorial agent which explains the following topics:	AI system	
	neuron	
	system parts	
	AND gate	
	OR gate	
	truth table	
	output status	
	training process	
	Sum of points:	

Adapted from (Shamir and Levin, 2021)

Fig. 7. Examples of performance-based rubrics

Gresse von Wangenheim *et al*. (2020) propose a rubric composed of eleven criteria regarding data management, model training, and interpretation of performance with respect to a Convolutional Neural Network (CNN) for the classification of recycling thrash created by the students. The rubric defines three performance levels (poor, acceptable, and good) with a detailed description of each one. Shamir and Levin (2021) define a rubric for the task to modify a simple neuron for the gates (and, or) in an Artificial Neural Network (ANN) to include explanations of the functionality, aiming to understand students' abstraction and reasoning on the proposed ANN. Only one performance level (acceptable) is defined.

Sakulkueakulsuk *et al*. (2018) assess learning results during various phases of the workshop in which students develop a ML model applied to classify mango fruits and evaluate its performance.

Observations and interviews. Most of the articles that report observations or interviews do not provide details, often only presenting a qualitative analysis. Some articles (Kandlhofer *et al.*, 2016; Tedre *et al.*, 2020; Vartiainen *et al.*, 2020, 2021; Ng and Chu, 2021; Shamir and Levin, 2021) used structured interviews (Fig. 8). One exception is presented by Hitron *et al.* (2019), who defined a rubric (a range from 0 to 3 for each criterion) for the analysis of the transcriptions of the short essay interview, aiming to identify if the children understand ML concepts/criterions of sample size, versatility, and negative examples. Additionally, in the final interview, when the children give examples of application in their lives of ML, the answers are categorized into 3 groups: accurate ML examples, non-ML examples, and fictional examples.

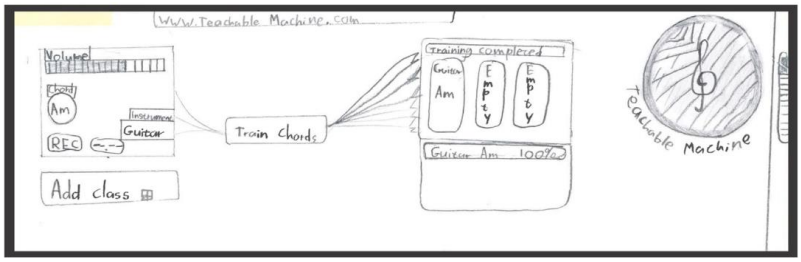
1. What is AI?
2. What does your artifact do?
3. How can it be extended?
4. What was difficult in creating it?
5. How was the course for you?

(Shamir and Levin, 2021)

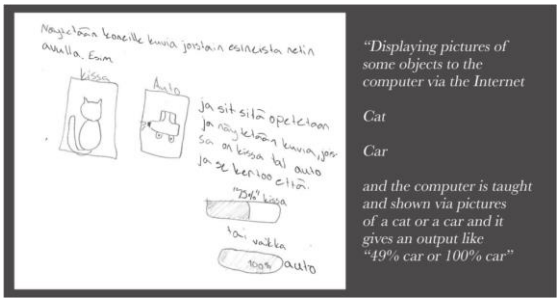
Fig. 8. Example of structured interview

Some articles (Kandlhofer *et al.*, 2016, Tedre *et al*., 2020; Vartiainen *et al*., 2020; Vartiainen *et al*., 2021) also report the use of drawings in order to assess the perception of (mostly younger) students on basic ML or AI concepts (Fig. 9), but without the use of a more explicitly defined protocol or rubrics.

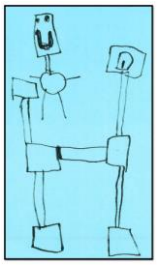
(Vartiainen *et al*., 2020)



(Vartiainen *et al*., 2021) and (Vartiainen *et al*., 2020)



(Tedre *et al*., 2020)



(Kandlhofer *et al.*, 2016)

Fig. 9. Examples of student drawings used for assessment

4.3 What instructional feedback is presented in what way to those involved, and what assessments are automated and how?

Most articles do not address details on if or which kind of feedback is provided. Given the emerging nature of these instructional units, often used in order to evaluate the course quality, currently there seems to be a lack of the provision of adequate feedback for the student's learning process.

Manual feedback. Most of the assessments are conducted and collected manually by the instructors. In some cases feedback has been given as part of a co-design process provided through an experienced person integrated within the group of learners that assists them in completing steps that require more technical knowledge in performing the task (Vartiainen *et al*., 2021; Tedre *et al*., 2020; Vartiainen *et al*., 2020; Kandlhofer *et al*. (2016). Sakulkueakulsuk *et al*. (2018) use a gamification mechanism to assess the learning during different phases of the workshop in which students develop a ML model

to classify the sweetness of mangoes based on their physical characteristics. As part of the gamification, students collect points for the accuracy of the models they trained and each correct prediction of a new sample.

The reported performance-based assessments also seem to limit feedback to the indication of the achieved performance levels (Gresse von Wangenheim *et al*., 2020), providing only implicitly an indication on how the students can improve their learning based on the next performance level definitions in the rubric. Shamir e Levin (2021) seems to provide only the indication of which criteria have been considered acceptable based on the artifact created by the students. Hitron *et al*. (2019) pre-defined a structured support providing feedback on if the child understood or misinterpreted the system's feedback regarding the accuracy of the model.

Automated feedback. Only a few papers report some automation of the assessment and feedback process. The online course "Machine Learning para Todos!" presented by Gresse von Wangenheim *et al*. (2020) includes quizzes as part of the interactive didactic material throughout the course with instantaneous correction indicating the correctness of the answers and brief explanations. Williams *et al*. (2019) reports that the robot being trained by the students gives feedback during activities aiming at predicting the children's next move or leading to rationalizing, providing auditive feedback on how the prediction works, which may help to develop the understanding of the student.

More indirectly, the platform created by Rodríguez-García *et al*. (2021), visualizes the precision and accuracy of the ML model, providing the students feedback on the quality of the model created by them. In a similar way, Hitron *et al*. (2019) report that students receive feedback on the accuracy of their trained model in real time as part of an activity recognizing a student gesture.

4.5 How were the assessments designed and evaluated?

Estevez *et al*. (2019), Tedre *et al*. (2020), Vartiainen *et al*. (2020), Vartiainen *et al*. (2021), and Henry *et al*. (2021) indicate the adoption of a design-based research methodology, with a parallel development of design processes, evaluation, and theory-building. Some report the development of the assessment (as part of the design of the entire instructional unit) with a motivational focus. Shamir and Levin (2021) use the ARCS Model for instructional design, and the analysis of student motivation, similar to Ng and Chu (2021) based on the "Motivated Strategies for Learning Questionnaire".

Williams *et al*. (2019) based one of their assessments on "Wellman and Liu's Theory of Mind assessments", defining questions given through a story, and creating a colorful scene for each question. The other assessment is based on the "Perception of Robots Questionnaire". This questionnaire is similar to the previous one, presenting two different representations of robots and students have to select one or inform them they are equal.

Kahn *et al*. (2018) use a mixed-methods approach to sequential exploratory design, while Priya *et al*. (2021) use a set of questions in combination with the Technology Acceptance Model. Yet, several articles do not provide details on how the assessments have been developed.

The majority of studies do not report any evaluation of the reliability or validity of the presented assessments, presenting only the evaluation of the quality of the instructional

unit itself. Table 6 summarizes the studies that report some kind of evaluation of the assessment.

Table 6
Evaluation of assessments

Reference	Research Design		Reliability			Validity		
	Sample size	Age of students	Internal consistency	Inter-rater	Intra-rater	Content	Construct	Criterion
(Shamir and Levin, 2021)	7	12				X		
(Hsu *et al*., 2021)	46	12	X					
(Hitron *et al*., 2019)	30	13-13	X	X				

The evaluation of reliability through internal consistency has been conducted by Hitron *et al*. (2019) and Hsu *et al.*, (2021). As result Hsu *et al*. (2021) reported a Cronbach α value of 0.883 for the reliability of a self-assessment questionnaire with five items using a Likert scale. Hitron *et al*. (2019) also report a high interrater reliability (Kappa = 92%) of the coding performed by the researchers when labeling the essay items. Aiming to evaluate content validity Shamir and Levin (2021) did not inform specific results, but mentioned that several elementary school students and teachers reviewed the questions with a researcher analyzing the ability to read and understand the item.

5 Discussion

In general, we observed a common lack of focus on assessments among the currently emerging courses for teaching ML in K-12. A large number of instructional units do not mention any kind of assessment, which may also be related to the fact that this knowledge area is not yet part of curricular content, but is mostly introduced through extracurricular activities. Many articles approach the evaluation of the quality of the instructional units rather than specifically assessing the students' learning with the purpose to guide the students' learning process.

In accordance with the current focus of more introductory courses to teach ML to novices, most assessments are related only to the lower cognitive levels of Bloom's Taxonomy including the levels of remembering, understanding and application. And, although the majority of the courses apply an active learning strategy guiding the students to develop ML artifacts, these applications again are more related to the use stage of the "use-modify-create" cycle. Even courses covering the modify and create stage, typically limit assessments to the use stage. Thus, the proposed assessments, in general, seem to be rather simple, adopting more tests/quizzes than, for example, performance-based assessments. An interesting approach is the use of drawings as a means to assess students' perception and ideas related to ML, especially with younger learners.

Most assessments are not rigorously defined and often even the explicit relation of assessments with learning objectives is not given. Exceptions are rubrics defined by Gresse von Wangenheim *et al.* (2020) and Shamir e Levin (2021).

And, although several ML courses adopt also game-based learning strategies (Williams *et al*., 2019; Priya *et al*., 2021; Henry *et al*., 2021), only Sakulkueakulsuk *et al*. (2018) uses a gamification mechanism for the assessment of the students' learning. Another alternative strategy is used by Ng and Chu (2021), who embed the use of social media on Edmodo (similar to Facebook) for delivering asynchronous videos and tasks to students, but again do not use this for assessments.

We also observed that the assessments do not necessarily cover all of the learning objectives of the instructional units. The content mostly assessed are basic ML concepts, ML approaches and in some cases ethical issues and the impact of ML on society. Such a focus seems to be in accordance with the recent nature of ML courses in K-12 and the early stage of the design of assessments.

Another issue we observed is that the majority of the assessments is intended to be analyzed manually, while (also due to the COVID pandemic) the number of online ML courses is increasing. And, also as part of face-to-face classes an (at least partial) automation of these assessments could provide feedback more rapidly, with less bias, while at the same time reducing the instructor's effort and freeing the instructor for other activities with respect to the student's learning process.

We also observed that basically all assessments provide rather simple feedback in terms of a grade, points, or indication of performance level. Or, in the case of quizzes, only the indication on if the given answer is correct or not, sometimes completed by a brief explanation. This clearly indicates a need for improvement in order to provide more constructive feedback to the student which may help to effectively guide the learning process.

Despite the existence of systematic approaches to develop and evaluate learning assessments, such as Evidence Centered Design (Mislevy and Riconscente, 2005; Mislevy *et al.*, 2017) or Item Response Theory (Reise and Revicki, 2015; DeVellis, 2017), we observed a lack of methodology concerning the development of the proposed assessments. Exceptions are assessments presented by Estevez *et al*. (2019), Tedre *et al*. (2020), Vartiainen *et al*. (2020), Vartiainen *et al*. (2021), and Henry *et al*. (2021) adopting a design-based research methodology for the development of the design, evaluation, and theory-building processes. And, as most assessments have been presented in articles only as part of instructional units, there is a lack of more detailed information on the assessments themselves, e.g., explicitly presenting the adherence between assessment instruments and the learning objectives, assessment instruments, etc.

We also encountered only three studies that report an evaluation of the reliability or validity of the proposed assessments, including the evaluation of internal consistency of the assessment instrument, interrater reliability for labeling essay-type items, as well as in a rather informal way the evaluation of content validity.

The results of this systematic mapping clearly show the need for the development of ML assessments in K-12 in a more comprehensive way, covering in a more varied way also other ML tasks, such as object detection, and natural language processing, applied in diverse domains. This also includes the automation of the assessments as well as the

improvement of the feedback provided to the learner. This also becomes clear when compared, for example, with the assessment of the learning of computational thinking for which a larger number and more detailed and rigorous approaches have already been proposed (Da Cruz Alves *et al.*, 2019; Tang *et al*., 2020). Yet, as the ML concepts and processes differ, there is a need to design specific assessments. On the other hand, when integrating the learning of ML into a wider context including also the deployment of the ML model into software systems, assessments developed for the learning of algorithms and programming, user interface design, or skills such as creativity and collaboration can be adopted in a complementary way.
In the future, in order to ensure the reliability and validity of the assessments, it is also important to adopt in a wider way systematic approaches for the development and evaluation of the assessment in a more robust way and conduct large-scale studies analyzing the assessments.

Threats to validity. In order to mitigate the impact of factors that may affect the validity of our review, we adopted several strategies. A common bias is that positive results tend to be published more than negative ones, but this should be a minor factor since the impacts of the learning process did not serve as a selection criterion. Another risk is the omission of relevant papers. In this regard, the search string including synonyms was carefully constructed to include all potentially relevant articles. Threats to the selection of relevant instructional units and data extraction were mitigated by defining and documenting a strict protocol, with the careful establishment of inclusion and exclusion criteria and discussion between the authors until consensus was obtained. Data extraction was performed by the first author, inferred when not explicitly stated in the article, and carefully reviewed by the co-author.

6 Conclusion

In this article, we present a systematic mapping of quantitative assessments of the student's learning of ML in K-12 published during the last ten years. As a result, we identified 27 instructional units teaching ML that present quantitative assessments of the students' learning. The majority of the assessments seem to be conducted in a simple manner, mainly at a lower cognitive level related to theoretical content or the use stage of the "use-modify-create" cycle. The content predominantly assessed are basic ML concepts, ML approaches, and in some cases ethical issues and the impact of ML on society. This may be explained by the still-emerging nature of ML education in K-12, as well as by the fact that several reported assessments seem to have been applied in order to evaluate the quality of the course rather than the student's learning with the purpose to guide the students' learning process. This may also explain a lack of more constructive feedback in order to guide the students' learning process. Therefore, the results of this mapping study clearly show opportunities for research in this area to develop and extend the evaluation of the assessment in a robust way, including the automation of assessment and feedback in order to contribute effectively to the learning of ML in K-12.

Acknowledgments

This work was supported by the CNPq (National Council for Scientific and Technological Development), a Brazilian government entity focused on scientific and technological development [Grant no. 303674/2019-9].

References

AI4K12, (2020), Artificial Intelligence (AI) for K-12. Retrieved from https://ai4k12.org/.

Akhter M. P., Jiangbin Z., Naqvi I. R., AbdelMajeed M., and Zia T., (2021), Abusive language detection from social media comments using conventional machine learning and deep learning approaches. *Multimedia Systems*.

Ala-Mutka K. and Jarvinen H.-M., (2004), Assessment process for programming assignments, Proc. of the *International Conference on Advanced Learning Technologies*, IEEE, 181–185.

Ala-Mutka K. M., (2005), A Survey of Automated Assessment Approaches for Programming Assignments. *Computer Science Education*, 15(2), 83–102.

Alexandre F., Becker J., Comte M.-H., Lagarrigue A., Liblau R., Romero M., and Viéville T., (2021), Why, What and How to Help Each Citizen to Understand Artificial Intelligence? *Künstliche Intelligenz*, 35(2), 191–199.

Anderson L. W. and Krathwohl D. R., (2001), *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*, Longman.

Bemley J. L., (1999), Neural networks for precollege students, Proc. of the *International Joint Conference on Neural Networks.* IEEE, 4422–4427 vol.6.

Bloom B. S., Engelhart M. D., Furst E. J., Hill W. H., and Krathwohl D. R., (1956), Taxonomy of educational objectives: the classification of educational goals: handbook I: cognitive domain, New York, US: D. Mckay.

Burgsteiner H., Kandlhofer M., and Steinbauer G., (2016), iRobot: Teaching an Evaluated, Competencies-Based Introductory Artificial Intelligence Class in Highschools. Proc. of the *Advances in Artificial Intelligence*, Springer, 218-223.

Chattopadhyay A., Christian D., Ulman A., and Sawyer C., (2018), A Middle-School Case Study: Piloting A Novel Visual Privacy Themed Module for Teaching Societal and Human Security Topics Using Social Media Apps, Proc. of the *Frontiers in Education Conference*, IEEE, 1–8.

Cheng Z., (2021), Applicational Status Analysis of Artificial Intelligence Technology in Middle School Education and Teaching, Proc. of the *International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy*, Springer International Publishing, 171–178.

Chlap P., Min H., Vandenberg N., Dowling J., Holloway L., and Haworth A., (2021), A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5), 545–563.

Cohen J., (1960), A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Creswell J. W. and Creswell J. D., (2018), *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, Los Angeles: SAGE.

Cronbach L. J., (1951), Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

CSTA, (2016), K–12 Computer Science Framework. Retrieved from http://www.k12cs.org.

Cutumisu M., Adams C., and Lu C., (2019), A Scoping Review of Empirical Research on Recent Computational Thinking Assessments. *Journal of Science Education and Technology*, 28(6), 651–676.

Da Cruz Alves N., Gresse Von Wangenheim C., and Hauck J. C. R., (2019), Approaches to Assess Computational Thinking Competences Based on Code Analysis in K-12 Education: A Systematic Mapping Study. *Informatics in Education*, 18(1), 17–39.

DeVellis R. F., (2017), *Scale development: theory and applications*, Fourth edition. SAGE.

Druga S. and Ko A. J., (2021), How do children's perceptions of machine intelligence change when training and coding smart programs?, Proc. of the *Interaction Design and Children*, ACM, 49–61.

Dryer A., Walia N., and Chattopadhyay A., (2018), A Middle-School Module for Introducing Data-Mining, Big-Data, Ethics and Privacy Using RapidMiner and a Hollywood Theme, Proc. of the *49th ACM Technical Symposium on Computer Science Education*, ACM, 753–758.

Erickson T. and Chen E., (2021), Introducing data science with data moves and CODAP. *Teaching Statistics*, 43(S1), S124–S132.
Estevez J., Garate G., and Graña M., (2019), Gentle Introduction to Artificial Intelligence for High-School Students Using Scratch. *IEEE Access*, 7, 179027–179036.
Frey N. and Fisher D., (2011), *The formative assessment action plan: practical steps to more successful teaching and learning*, ASCD.
Frischemeier D., Biehler R., Podworny S., and Budde L., (2021), A first introduction to data science education in secondary schools: Teaching and learning about data exploration with CODAP using survey data. *Teaching Statistics*, 43(S1), S182–S189.
Fry K. and Makar K., (2021), How could we teach data science in primary school? *Teaching Statistics*, 43(S1), S173–S181.
Fullan M., Langworthy M., and Barber M., (2014), *A rich seam: how new pedagogies find deep learning*. Retrieved from https://staging.oer4pacific.org/id/eprint/5/1/Rich%20seam.pdf
Gao Z., Dang W., Wang X., Hong X., Hou L., Ma K., and Perc M., (2021), Complex networks and deep learning for EEG signal analysis. *Cognitive Neurodynamics*, 15(3), 369–388.
Glorfeld L. W., (1995), An Improvement on Horn's Parallel Analysis Methodology for Selecting the Correct Number of Factors to Retain, *Educational and Psychological Measurement*, 55(3), 337-393.
Gresse von Wangenheim C., Marques L. S., and Hauck J. C. R., (2020), Machine Learning for All – Introducing Machine Learning in K-12, *SocArXiv*.
Grover S. and Pea R., (2013), Computational Thinking in K–12: A Review of the State of the Field. *Educational Researcher*, 42(1), 38–43.
Grover S., Pea R., and Cooper S., (2015), Designing for deeper learning in a blended computer science course for middle school students. *Computer Science Education*, 25(2), 199–237.
Hattie J. and Timperley H., (2007), The Power of Feedback. *Review of Educational Research*, 77(1), 81–112.
Heale R. and Twycross A., (2015), Validity and reliability in quantitative studies. *Evidence Based Nursing*, 18(3), 66–67.
Henry J., Hernalesteen A., and Collard A.-S., (2021), Teaching Artificial Intelligence to K-12 Through a Role-Playing Game Questioning the Intelligence Concept. *Künstliche Intelligenz*, 35(2), 171–179.
Hitron T., Orlev Y., Wald I., Shamir A., Erel H., and Zuckerman O., (2019), Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes, Proc. of the *Conference on Human Factors in Computing Systems*, ACM, 1–11.
Hsu T.-C., Abelson H., and Van Brummelen J., (2021), The Effects of Applying Experiential Learning into the Conversational AI Learning Platform on Secondary School Students. *Preview Version. Accepted by the IRRODL Special Issue "AI e-Learning and Online Curriculum".*
Huba M. E. and Freed J. E., (2000), *Learner-centered assessment on college campuses: shifting the focus from teaching to learning*, Allyn and Bacon.
Ihantola P., Ahoniemi T., Karavirta V., and Seppälä O., (2010), Review of recent systems for automatic assessment of programming assignments, Proc. of the *10th Koli Calling International Conference on Computing Education Research*, ACM Press, 86–93.
Jing M., (2018), China looks to school kids to win the global AI race. *South China Morning Post*. Retrieved from https://www.scmp.com/tech/china-tech/article/2144396/china-looks-school-kids-win-global-ai-race
Kahn K., (1977), Three interactions between AI and education. *Machine intelligence*, 8, 422–429.
Kahn K., Megasari R., Piantari E., and Junaeti E., (2018), AI Programming by Children using Snap! Block Programming in a Developing Country. Proc. of the *13th European Conference On Technology Enhanced Learning*, 2193(13), 14.
Kahn K. and Winters N., (2021), Learning by Enhancing Half-Baked AI Projects. *Künstliche Intelligenz*, 35(2), 201–205.
Kandlhofer M. and Steinbauer G., (2021), AI K–12 Education Service. *Künstliche Intelligenz*, 35(2), 125–126.
Kandlhofer M., Steinbauer G., Hirschmugl-Gaisch S., and Huber P., (2016), Artificial intelligence and computer science in education: From kindergarten to university, Proc. of the *IEEE Frontiers in Education Conference*, 1–9.
Kandlhofer M., Steinbauer G., Lassnig J., Menzinger M., Baumann W., Ehardt-Schmiederer M., et al., (2021), EDLRIS: A European Driving License for Robots and Intelligent Systems. *Künstliche Intelligenz*, 35(2), 221–232.

Kimberlin C. L. and Winterstein A. G., (2008), Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, 65(23), 2276–2284.

Lasser J., Manik D., Silbersdorff A., Säfken B., and Kneib T., (2021), Introductory data science across disciplines, using Python, case studies, and industry consulting projects. *Teaching Statistics*, 43(S1), S190–S200.

Lee I., Ali S., Zhang H., DiPaola D., and Breazeal C., (2021), Developing Middle School Students' AI Literacy, *Proc. of the 52nd ACM Technical Symposium on Computer Science Education*, ACM, 191–197.

Lee I., Martin F., Denner J., Coulter B., Allan W., Erickson J., et al., (2011), Computational thinking for youth in practice. *ACM Inroads*, 2(1), 32–37.

Marques L. S., Gresse Von Wangenheim C., and Hauck J. C. R., (2020), Teaching Machine Learning in School: A Systematic Mapping of the State of the Art. *Informatics in Education*, 19(2), 283–321.

McMillan J. H., (2018), *Classroom assessment: principles and practice that enhance student learning and motivation*, Seventh edition. Pearson Education.

Melsión G. I., Torre I., Vidal E., and Leite I., (2021), Using Explainability to Help Children Understand Gender Bias in AI, Proc. of the *Interaction Design and Children*, ACM, 87–99.

Mike K., Hazan T., and Hazzan O., (2020), Equalizing Data Science Curriculum for Computer Science Pupils, Proc. of the *20th Koli Calling International Conference on Computing Education Research*, ACM, 1–5.

Mislevy R. J., Haertel G., Riconscente M., Wise Rutstein D., and Ziker C., (2017), *Assessing Model-Based Reasoning using Evidence- Centered Design*, Springer International Publishing.

Mislevy R. J. and Riconscente M. M., (2005), *Evidence-Centered Assessment Design: Layers, Structures, and Terminology*, SRI International and University of Maryland.

Mitchell T. M., (1997), *Machine Learning*, McGraw-Hill.

Mobasher B., Dettori L., Raicu D., Settimi R., Sonboli N., and Stettler M., (2019), Data Science Summer Academy for Chicago Public School Students. *ACM SIGKDD Explorations Newsletter*, 21(1), 49–52.

Morrison G. R., Ross S. M., Morrison J. R., and Kalman H. K., (2019), *Designing effective instruction*, Eighth edition. Wiley.

Moskal B. M. and Leydens J. A., (2000), Scoring rubric development: Validity and reliability. *Practical assessment, research, and evaluation*, 7(1), 10.

Ng T. K. and Chu K. W., (2021), Motivating Students to Learn AI Through Social Networking Sites: A Case Study in Hong Kong. *Online Learning*, 25(1), 195-208.

Ossovski E. and Brinkmeier M., (2019), Machine Learning Unplugged - Development and Evaluation of a Workshop About Machine Learning, in *Informatics in Schools. New Ideas in School Informatics*,Springer International Publishing, 136–146.

Osterlind S. J., (1989), *Constructing Test Items*, Springer Netherlands.

Papert S., Solomon C., Soloway E., and Spohrer J. C., (1971), Twenty things to do with a computer, in *Studying the novice programmer*, Lawrence Erlbaum Associates, 3–28.

Petersen K., Feldt R., Mujtaba S., and Mattsson M., (2008), Systematic mapping studies in software engineering, Proc. of the *International Conference on Evaluation and Assessment in Software Engineering*, *12*, 1–10.

Petersen K., Vakkalanka S., and Kuzniarz L., (2015), Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64, 1–18.

Piasecki J., Waligora M., and Dranseika V., (2018), Google Search as an Additional Source in Systematic Reviews. *Science and Engineering Ethics*, 24(2), 809–810.

Priya S., Bhadra S., and Chimalakonda S., (2021), ML-Quest: A Game for Introducing Machine Learning Concepts to K-12 Students. *arXiv:2107.06206*.

Queiroz R.L., Ferrentini Sampaio F., Lima C., and Machado Vieira Lima P., (2021), AI from Concrete to Abstract: Demystifying artificial intelligence to the general public, *AI & SOCIETY*, 36(3), 877-893.

Reise S. P. and Revicki D. A., (2015), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*, Routledge, 484.

Rodríguez-García J. D., Moreno-León J., Román-González M., and Robles G., (2021), Evaluation of an Online Intervention to Teach Artificial Intelligence with LearningML to 10-16-Year-Old Students, Proc. of the *52nd ACM Technical Symposium on Computer Science Education*, ACM, 177–183.

Rodríguez-García J. D., Moreno-León J., Román-González M., and Robles G., (2020), LearningML: A Tool to Foster Computational Thinking Skills Through Practical Artificial Intelligence Projects, *Revista de Educación a Distancia*, 20(63), artic. 07.

Romli R., Sulaiman S., and Zamli K. Z., (2010), Automatic programming assessment and test data generation a review on its approaches, Proc. of the *International Symposium on Information Technology*, IEEE, 1186–1192.

Rothwell W. J., (2016), *Mastering the instructional design process: a systematic approach*, Fifth edition. Wiley.

Royal Society, (2017), *Machine learning: the power and promise of computers that learn by example.* Retrieved from royalsociety.org/machine-learning.

Sadler D. R., (1989), Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144.

Sakulkueakulsuk B., Witoon S., Ngarmkajornwiwat P., Pataranutaporn Pornpen, Surareungchai W., Pataranutaporn Pat, and Subsoontorn P., (2018), Kids making AI: Integrating Machine Learning, Gamification, and Social Context in STEM Education, Proc. of the *IEEE International Conference on Teaching, Assessment, and Learning for Engineering*, 1005–1010.

Sanusi I. T. and Oyelere S. S., (2020), Pedagogies of Machine Learning in K-12 Context, Proc. of the *IEEE Frontiers in Education Conference*, 1–8.

Seel N. M., Lehmann T., Blumschein P., and Podolskiy O. A., (2017), *Instructional Design for Learning: Theoretical Foundations*, Sense Publishers.

Shamir G. and Levin I., (2021), Neural Network Construction Practices in Elementary School. *Künstliche Intelligenz*, 35(2), 181–189.

Stegeman M., Barendsen E., and Smetsers S., (2016), Designing a rubric for feedback on code quality in programming courses, Proc. of the *16th International Conference on Computing Education Research*, Koli Calling, ACM, 160–164.

Steinbauer G., Kandlhofer M., Chklovski T., Heintz F., and Koenig S., (2021), A Differentiated Discussion About AI Education K-12. *Künstliche Intelligenz*, 35(2), 131–137.

Stevens E., Dixon D. R., Novack M. N., Granpeesheh D., Smith T., and Linstead E., (2019), Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. *International Journal of Medical Informatics*, 129, 29–36.

Tang X., Yin Y., Lin Q., Hadad R., and Zhai X., (2020), Assessing computational thinking: A systematic review of empirical studies. Computers & Education, 148, 103798.

Tedre M., Vartiainen H., Kahila J., Toivonen T., Jormanainen I., and Valtonen T., (2020), Machine Learning Introduces New Perspectives to Data Agency in K—12 Computing Education, Proc. of the *IEEE Frontiers in Education Conference*, 1–8.

Tikva C. and Tambouris E., (2021), Mapping computational thinking through programming in K-12 education: A conceptual model based on a systematic literature Review. *Computers & Education*, 162, 104083.

Tor H. T., Ooi C. P., Lim-Ashworth N. S., Wei J. K. E., Jahmunah V., Oh S. L., et al., (2021), Automated detection of conduct disorder and attention deficit hyperactivity disorder using decomposition and nonlinear techniques with EEG signals. *Computer Methods and Programs in Biomedicine*, 200, 105941.

Touretzky D., Gardner-McCune C., Martin F., and Seehorn D., (2019a), Envisioning AI for K-12: What Should Every Child Know about AI? Proc. of the *AAAI Conference on Artificial Intelligence*, 33(01), 9795–9799.

Touretzky D. S., Gardner-McCune C., Martin F., and Seehorn D., (2019b), K-12 Guidelines for Artificial Intelligence: What Students Should Know, Proc. of the *ISTE Conference*, 53.

Usman O. L., Muniyandi R. C., Omar K., and Mohamad M., (2021), Advance Machine Learning Methods for Dyslexia Biomarker Detection: A Review of Implementation Details and Challenges. *IEEE Access*, 9, 36879–36897.

Van Brummelen J., Heng T., and Tabunshchyk V., (2020), Teaching Tech to Talk: K-12 Conversational Artificial Intelligence Literacy Curriculum and Development Tools. *arXiv:2009.05653*.

Vartiainen H., Toivonen T., Jormanainen I., Kahila J., Tedre M., and Valtonen T., (2021), Machine learning for middle schoolers: Learning through data-driven design. *International Journal of Child-Computer Interaction*, 29, 100281.

Vartiainen H., Toivonen T., Jormanainen I., Kahila J., Tedre M., and Valtonen T., (2020), Machine learning for middle-schoolers: Children as designers of machine-learning apps, Proc. of the *IEEE Frontiers in Education Conference*, 1–9.

von Wangenheim C. G. von, Marques L. S., and Hauck J. C. R., (2020), Machine Learning for All – Introducing Machine Learning in K-12.
Williams R., Park H. W., and Breazeal C., (2019), A is for Artificial Intelligence: The Impact of Artificial Intelligence Activities on Young Children's Perceptions of Robots, Proc. of the *Conference on Human Factors in Computing Systems*, ACM, 1–11.
World Economic Forum, (2021), McKinsey: These are the skills you will need for the future of work. *World Economic Forum*. Retrieved from https://www.weforum.org/agenda/2021/06/defining-the-skills-citizens-will-need-in-the-future-world-of-work/
Xue G., Liu S., Gong D., and Ma Y., (2021), ATP-DenseNet: a hybrid deep learning-based gender identification of handwriting. *Neural Computing and Applications*, 33(10), 4611–4622.
Zhou X., Van Brummelen J., and Lin P., (2020), Designing AI Learning Experiences for K-12: Emerging Works, Future Opportunities and a Design Framework. *arXiv:2009.10228*.

Marcelo Fernando Rauber is a professor for Informatics at the *Instituto Federal Catarinense (IFC)*, Camboriú, Brazil. Currently he is a Ph.D. student of the Graduate Program in Computer Science (PPGCC) at the Federal University of Santa Catarina (UFSC), Florianópolis, Brazil, and, a research student at the initiative Computing at Schools/INCoD/INE/UFSC. He received a MSc. (2016) in Science and Technology Education from the Federal University of Santa Catarina, a specialization (2005) in Information Systems Administration from UFLA and a BSc. (2004) in Computer Science from UNIVALI. His main research interests are computing education and assessment.

C. Gresse von Wangenheim is a professor at the Department of Informatics and Statistics (INE) of the Federal University of Santa Catarina (UFSC), Florianópolis, Brazil, where she coordinates the Software Quality Group (GQS) focusing on scientific research, development and transfer of software engineering models, methods and tools and software engineering education. She also coordinates the initiative Computing at Schools, which aims at bringing computing education to schools in Brazil. She received the Dipl.-Inform. and Dr. rer. nat. degrees in Computer Science from the Technical University of Kaiserslautern (Germany), and the Dr. Eng. degree in Production Engineering from the Federal University of Santa Catarina.

Appendix 1 - Brief description of the instructional units included in the systematic review

Reference	Title	Target age (in years)	Brief description of the instructional unit	Country	Duration (in hours)
(Alexandre *et al.*, 2021)	Why, What and How to Help Each Citizen to Understand Artificial Intelligence?	15-Adult	Online initiative (over 9 months during the Covid-19 pandemic), acting as a core curriculum for AI. Cover AI general topics and central role of data in ML. Offering materials and links to other courses, aiming for 21st century skills.	France	10-20 (on-line)
(Chattopadhyay *et al*., 2018)	A Middle-School Case Study: Piloting A Novel Visual Privacy Themed Module for Teaching Societal and Human Security Topics Using Social Media Apps	11-13	Pilot workshop as part of the Google-IgniteCS program aiming to introduce privacy concepts and the potential of applying ML in this context, to increase the overall interests in privacy, ML as well as computing and cyber-security.	USA	1,5
(Dryer *et al*., 2018)	A Middle-School Module for Introducing Data-Mining, Big-Data, Ethics and Privacy Using RapidMiner and a Hollywood Theme	11-19	Research aiming to teach the K-12 public a basic understanding of data-driven intelligence, and an awareness of big-data related ethics and privacy, conducted through the Google-IgniteCS program and NSA GenCyber Camp.	USA	NI
(Estevez *et al.*, 2019)	Gentle Introduction to Artificial Intelligence for High-School Students Using Scratch	16-17	Face-to-face workshop to explain and experiment basic AI techniques (K-means and a simplified ANN) using a scaffolded design through incomplete algorithms with Scratch as a block-based programming environment.	Spain	1
(Gresse von Wangenheim *et al.*, 2020)	Machine Learning for All - Introducing Machine Learning in K-12	10-14	Face-to-Face/online course introducing basic ML concepts and the development of a first image recognition model.	Brazil	8
(Henry *et al*., 2021)	Teaching Artificial Intelligence to K-12 Through a Role-Playing Game Questioning the Intelligence Concept	10-14	Course through an unplugged role-playing game, inspired by the game "Guess Who?", aimed for children to discover the basic concepts of ML. In groups, successively assuming the role of a developer, a tester, or AI, his mission is simulating an ML system capable of identifying an animal, through a dataset of simple questions.	Belgium	3,3

(Hitron *et al*., 2019)	Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes	10-13	Gest tool used for studying children's learning of ML concept and the process of classification problems in supervised ML. Gest uses a hardware device with accelerometer and software modules that recognize the gestures (stopped, circle or square) of children and provides feedback about the accuracy detected. In three different arrangements, children collected and classified data by themselves and evaluated their sampling.	Israel	NI
(Hsu *et al*., 2021)	The Effects of Applying Experiential Learning into the Conversational AI Learning Platform on Secondary School Students	12	Comparison of teaching approaches (experiential learning cycle vs doing conversational projects) using a visual programming interface for conversational AI applications with Alexa and MIT App Inventor.	Taiwan	NI (over 6 weeks)
(Kahn *et al*., 2018)	AI Programming by Children using Snap! Block Programming in a Developing Country	16-17	Face-to-face workshop aimed to investigate the process of learning AI using Snap speech synthesis and ML.	Austria	NI
(Kandlhofer *et al*., 2016)	Artificial intelligence and computer science in education: From kindergarten to university	5-17	Four face-to-face proofs-of-concept from a developed AI literacy from kindergarten to university and their assessments.	Austria	NI
(Kandlhofer *et al*., 2021)	EDLRIS: A European Driving License for Robots and Intelligent Systems	NI (school students, apprentices, young people)	The EDLRIS initiative, a consortium who provides training, curriculum (including material), and certification for teachers (acting as multiplicators) and students (school students, apprentices, young people) over the AI or/and Robotics topics (each one subdivided in Basic and Advanced).	Austria and Hungary	60 (basic AI module include ML, with face-to-face and online tasks)
(Lee *et al*., 2021)	Developing Middle School Students' AI Literacy	10-14	During the Covid-19 pandemic, middle school students participated in a summer workshop, with two main objectives: to develop AI literacy as critical and ethical citizens and basic knowledge and skills at the subject.	USA	30 (on-line)
(Melsión *et al*., 2021)	Using Explainability to Help Children Understand Gender Bias in AI	10-14	During the Covid-19 pandemic, students used a visualization educational tool (Grad-CAM) attempting to address the binary gender bias in ML, covering the three main steps of a machine learning classifier: labeling, training, and evaluation.	Sweden	0,27 (on-line, 16 minutes)

(Mike *et al*., 2020)	Equalizing Data Science Curriculum for Computer Science Pupils	15-16	Pilot implementation and delivery of Data Science curriculum taught for 10th grade and for teachers in a public school, who incorporates both a broad view on data science and a data workflow, focusing on a deep understanding of ML.	Israel	90
(Mobasher *et al*., 2019)	Data Science Summer Academy for Chicago Public School Students	15-17	A week-long summer course in public High School was taught aiming to increase awareness about data science. including algo de ML?	USA	NI
(Ng and Chu, 2021)	Motivating Students to Learn AI Through Social Networking Sites: A Case Study in Hong Kong	12-15	During the Covid-19 pandemic, teachers have conducted a case study, redesigning three times previous lessons about ML and offered through Edmodo (a Social Network Site - SNS that resembles Facebook). Covering introduction to AI and ML fundamentals, ANN and Deep Learning. Lessons include interaction over SNS, videos, synchronous conferences, gamification activities, and assessment.	Hong Kong	24 (Estimated, online over 12 weeks)
(Ossovski and Brinkmeier, 2019)	Machine Learning Unplugged - Development and Evaluation of a Workshop About Machine Learning	15-17	Series of face-to-face workshops at High School (#28) and graduated trainees (#16). Over an action-oriented method, they simulated through an unplugged activity: a simple machine learning linear classification method (the results of a manual classification are marked in a pinboard with a wooden strip).	Germany	1,5
(Priya *et al*., 2021)	ML-Quest: A Game for Introducing Machine Learning Concepts to K-12 Students	15-18 (estimated, High School)	ML-Quest is an RPG 3D video game, with 3 phases, each one aims to introduce and work with a conceptual ML approach (Supervised Learning, Gradient Descent and KNN Classification).	India	Not informed. (on-line)
(Rodríguez-García *et al*., 2020)	LearningML: A Tool to Foster Computational Thinking Skills Through Practical Artificial Intelligence Projects	20 (aimed to include children)	A preliminary pilot face-to-face workshop using the workflow-based ML tool (LearningML), investigating if the use of LearningML and Scratch by students with no previous knowledge about ML and IA (but have some in development) can leave to learn the basics about ML and IA.	Spain	2
(Rodríguez-García *et al*., 2021)	Evaluation of an Online Intervention to Teach Artificial Intelligence with LearningML to 10-16-Year-Old Students	10-16	Based on previous works related to the LearningML web platform, the research investigates if the use of Learning ML by students with no previous knowledge about ML and IA can leave to learn the basics about ML and IA, through an Online course during the Covid-19 pandemic.	Spain	Not measured (on-line, intervention data collection occurs over 21 days)

(Sakulkueakulsuk *et al*., 2018)	Kids making AI: Integrating Machine Learning, Gamification, and Social Context in STEM Education	12-14	Gamification workshop taught for 7th to 9th grade, where the students are introduced to ML and develop an ML model applied to a local and familiar issue.	Thailand	9
(Shamir and Levin, 2021)	Neural Network Construction Practices in Elementary School	12	A curriculum applied as an extracurricular course defined for grades 3-5. In the 6th grade, AI topic has been introduced, focused on all aspects of ML, including artificial neural networks (ANN). A small group extra class course was conducted, with an approach on developing a neuron using a programmable learning environment (PLE) that is based on MIT scratch.	Israel	24 (not clear, 6-day course)
(Tedre *et al*., 2020) (Vartiainen *et al*., 2020)	Machine Learning Introduces New Perspectives to Data Agency in K—12 Computing Education; Machine learning for middle-schoolers: Children as designers of machine-learning apps	12-13	A pilot exploratory study case, where the students co-designed and design a ML application for solving meaningful everyday problems.	Finland	9
(Van Brummelen *et al*., 2020)	Teaching Tech to Talk: K-12 Conversational Artificial Intelligence Literacy Curriculum and Development Tools	13-17	Online workshop in which researchers evaluated the Conversational Agent interface for MIT App Inventor and a remote workshop guided by Long and Magerko´s AI design literacy.	EUA	12,5 (online over 5 days)
(Vartiainen *et al*., 2021)	Machine learning for middle schoolers: Learning through data-driven design	12-13	Course from the first study (Tedre *et al*., Vartiainen *et al*., 2020) emerged open questions, and this second study conducted, based on a design-oriented pedagogy, guides leave the students through an ideation process using “Google’s Teachable Machine 2” to develop an ML educational application, having as co-designers teachers and a group of Computer Science specialists.	Finland	8
(Williams *et al*., 2019)	A is for Artificial Intelligence: The Impact of Artificial Intelligence Activities on Young Children's Perceptions of Robots	4-6	Face-to-face workshops focusing in how training Data Influences ML focusing in Kindergarden students using the tool "PopBots" a two-part interactive system: A Lego Robot tand hat is independent, programable, and as an active role and a tablet for the activities and programming that consists of a blocks-based interface for programming and training a robot with a set of activities and assessments around ML and IA.	EUA	1

Appendix 2 - Summary of assessment characteristics

(NI - not informed)

Reference	Assessment method	Types of items	ML concepts assessed	Learning level	Use-Modify-Create stage(s)	Manual/Automated assessment
(Alexandre *et al*., 2021)	- Test/Quizzes	NI	NI	NI	NI	NI
(Chattopadhyay *et al*., 2018)	- Self-assessment	- Likert-style - Short essay	- Basic ML concepts (interest in the field)	Remember	None	Manual
(Dryer *et al*., 2018)	- Self-assessment	- Likert-style	- Basic ML concepts	Remember	None	Manual
(Estevez *et al*., 2019)	- Self-assessment	- Likert-style - Short essay	- Basic ML concepts (applications)	Understand	None	Manual
(Gresse von Wangenheim *et al*., 2020)	- Test/Quizzes - Performance-based	- Multiple choice - Drag-and-drop - Rubric	- What is learning? - Approaches to ML (Classification, ANNs, and ML process) - Fundamentals of neural networks - How Training Data Influences Learning - Limitations of ML (including impact of ML)	Apply	Use	Automated (Quizzes)
(Henry *et al*., 2021)	- Test/Quizzes - Observation	NI	- Basic ML concepts	NI	NI	NI
(Hitron *et al*., 2019)	- Interview	- Short essay - Rubric	- Basic ML concepts (data set)	Understand	Use	Manual, except by the the system acuracy in

						recognize a new gesture, in real time.
(Hsu *et al*., 2021)	- Test/Quizzes	- Multiple-choice	Approaches to ML (Conversational agents)	NI	NI	NI
(Kahn *et al*., 2018)	- Observation - Self-assessment	NI	- Approaches to ML (Conversational agents)	Apply	Use-Modify	Manual
(Kandlhofer *et al*., 2016)	- Test/Quizzes - Observation - Interviews - Self-assessment - Performance-based	- Likert-style - Short essay - Multiple choice - Semi-structured interviews - Program code	- Basic ML concepts (Interest in the field)	Apply.	Use-Modify.	Manual
(Kandlhofer *et al*., 2021)	- Test/Quizzes	- Multiple-choice	- Basic ML concepts - AI Ethics	NI	NI	NI
(Lee *et al*., 2021)	- Test/Quizzes - Interview - Observation	NI	- Basic ML concepts - AI careers	NI	NI	NI
(Melsión *et al*., 2021)	- Test/Quizzes - Self-Assessment	- Likert-style - Brief- constructed response - Multiple-choice	- Basic ML concepts - ML Process - Bias	Analise	Use	NI
(Mike *et al*., 2020)	- Observation - Performance-based - Test/Quizzes	NI	NI	NI	NI	Manual

(Mobasher *et al*., 2019)	- Self-assessment	- Likert-style - Multiple choice - Short essay - Checkbox	- Approaches to ML (Clustering, decision trees, K-nearest Neighbor) How Training Data Influences Learning	Remember	None	NI
(Ng and Chu, 2021)	- Interview - Observation - Self-Assessment - Test/Quizzes	- Semi-structured interviews - Likert-style - Brief- constructed response - True-or-False - Multiple-choice	NI	NI	NI	NI
(Ossovski and Brinkmeier, 2019)	- Self-assessment	- Likert-style	- Basic ML concepts	Understand	None	Manual
(Priya *et al*., 2021)	- Self-Assessment	- Likert-style	- Basic ML concepts	Remember	NI	NI
(Rodríguez-García *et al*., 2020)	- Self-Assessment - Test/Quizzes	- Likert-style - Multiple choice - Short essay	- Basic ML concepts	Understand	Use	Manual
(Rodríguez-García *et al*., 2021)	- Test/Quizzes	- Multiple choice - Likert-style	NI	NI	NI	NI System indicates accuracy of the ML model.
(Sakulkueakulsuk *et al*., 2018)	- Self-Assessment - Performance-based - assessment through gamification	- Likert-style - ML model accuracy	- Basic ML concepts	Apply	Modify	Manual

(Shamir and Levin, 2021)	- Test/Quizzes - Interviews - Performance-Based	- Selected- Response and Brief- constructed Response - Rubric - Semi- Structured Interview	- Basic ML concepts - ML process	Apply	Use	NI
(Van Brummelen *et al*., 2020)	- Test/Quiz - Performance-Based - Observation	- Multiple-choice - Selected- response - Short answer	- What is learning? - Approaches to ML - Types of learning algorithms by learning styles - Fundamentals of neural networks - Types of neural net architectures - How Training Data Influences Learning - Limitations of Machine Learning	NI	NI	Manual
(Williams *et al*., 2019)	- Self-Assessment - Test/Quizzes - Observation	- Likert-Style - Multiple choice - Informal observation	- What is learning? Approaches to ML (Knowledge-Based Systems, Classification) - Types of learning algorithms (Supervised learning) - How Training Data Influences Learning	Apply	None	Manual. Robot give feedback during activities.