Informatics in Education, 2022, Vol. 21, No. 3, 479–500


DOI: 10.15388/infedu.2022.18

# A Proposal for Performance-based Assessment of the Learning of Machine Learning Concepts and Practices in K-12

Christiane GRESSE VON WANGENHEIM[1],
Nathalia DA CRUZ ALVES[1], Marcelo F. RAUBER[1,2],
Jean C. R. HAUCK[1], Ibrahim H. YETER[3]

[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*
[3]*National Institute of Education, Nanyang Technological University, Singapore*
*e-mail: c.wangenheim@ufsc.br, nathalia.alves@posgrad.ufsc.br, marcelo.rauber@ifc.edu.br, jean.hauck@ufsc.br, ibrahim.yeter@nie.edu.sg*

Received: June 2021

**Abstract.** Although Machine Learning (ML) is used already in our daily lives, few are familiar with the technology. This poses new challenges for students to understand ML, its potential, and limitations as well as to empower them to become creators of intelligent solutions. To effectively guide the learning of ML, this article proposes a scoring rubric for the performance-based assessment of the learning of concepts and practices regarding image classification with artificial neural networks in K-12. The assessment is based on the examination of student-created artifacts as a part of open-ended applications on the *use* stage of the Use-Modify-Create cycle. An initial evaluation of the scoring rubric through an expert panel demonstrates its internal consistency as well as its correctness and relevance. Providing a first step for the assessment of concepts on image recognition, the results may support the progress of learning ML by providing feedback to students and teachers.

**Keywords:** assessment, education, rubric, machine learning, K-12.

## 1. Introduction

Machine Learning (ML) has become part of our everyday life deeply impacting our society. Different from Artificial Intelligence, focusing on theory and development of computer systems able to perform tasks that normally require human intelligence, Machine Learning focuses on the development of systems that learn and improve from experience on their own without having to be explicitly programmed. Currently, ML

is one of the most rapidly growing areas within artificial intelligence (Holzinger *et al.*, 2018). Recent progress in ML has been specifically achieved by deep learning approaches using neural networks, dramatically improving the state-of-the-art in image recognition, object detection, and speech recognition in many domains (Jordan and Mitchell, 2015; LeCun *et al.*, 2015).

Yet, most do not understand the technology behind it, which can make ML mysterious or even scary, overshadowing its potential positive impact on society (Evangelista *et al.*, 2018; Ho and Scadding, 2019). Thus, to demystify what ML is, how it works and what are its impact and limitations, there is a growing need for public understanding of ML (House of Lords, 2018; Tuomi, 2018). Therefore, it becomes important to introduce basic concepts and practices already in school, empowering students to become more than just consumers, but also creators of intelligent solutions (Kandlhofer *et al.*, 2016; Royal Society, 2017; Touretzky *et al.*, 2019). Knowledge about ML concepts, the ability to use and create ML models, together with the ability to critically analyze benefits, social and ethical aspects of AI, are becoming key skills of the 21st century to educate the next generation as responsible citizens (Steinbauer *et al.*, 2021; Touretsky *et al.*, 2019).

And, although being a complex knowledge area, studies have shown that children are able to learn ML concepts from a relatively young age (Hitron *et al.*, 2019). The introduction to this kind of complex knowledge also has the potential to improve children's everyday skills as well as to better prepare them to deal with challenges that arise as a result of the use of ML (Kahn *et al.*, 2020).

It may also encourage more students to choose computing careers and provide adequate preparation for higher education taking into consideration a major shift in the labor market with a fast-growing need for ML-literate workers (Tuomi, 2018; Touretsky *et al.*, 2019). Thus, teaching ML at K-12 not only helps young people to understand this emerging technology and how it works but can also inspire future ML users and creators to get acquainted with the world, to understand it, and to change it (Pedró *et al.*, 2019; Webb *et al.*, 2021).

As indicated by the curricular guidelines for teaching Artificial Intelligence (Touretzky *et al.*, 2019), teaching AI in K-12 should also include Machine learning represented by Big Idea #3 – Learning (Fig. 1). Following these guidelines, teaching ML on this educational stage should include an understanding of basic ML concepts, such as learning algorithms and fundamentals of neural networks, as well as limitations and ethical concerns related to ML.

As ML is a complex knowledge area, it is important to carefully define the sequence of learning goals to be achieved with sufficient scaffolds for novices to start to create ML models with little instruction in the beginning to keep students engaged (low threshold) while also being able to support sophisticated programs with the learning progression (high ceiling). In this context, active learning that stresses action and direct experience is crucial to make ML transparent and enable students to build correct mental models (Wong *et al.*, 2020). As a part of the human-centric development of an ML model, students can explore several tasks from preparing a dataset, selecting an appropriate learning algorithm, training the ML model, and evaluating its performance (Lwakatare *et al.*,

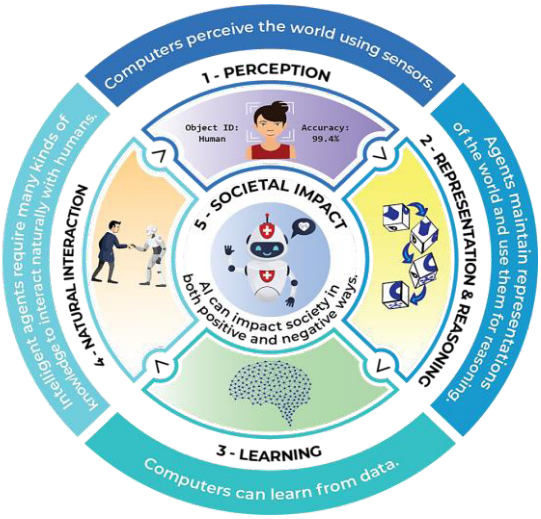



Fig. 1. 5 Big Ideas for teaching Artificial Intelligence in K-12 (Touretzky *et al.*, 2019).

2019; Ramos *et al.*, 2020). Representing a complex area, the best approach is to start with lower-level competencies and then progress upwards. In order to guide the learning progression focusing on the application of ML concepts and practices, often the *Use-Modify-Create* cycle (Lytle *et al.*, 2019) is also applied for ML education. Following this cycle, students are introduced to ML topics by *using* and analyzing a provided ML artifact as well as learning how to develop a predefined ML model, then *modifying* one, until *creating* their own ones.

Traditionally, Machine Learning has been taught mostly in higher education (McGovern *et al.*, 2011; Torrey, 2012). And, although there are many programs today that focus on coding and robotics, K-12 education still needs to embrace the teaching of Artificial Intelligence, including ML (Hubwieser *et al.*, 2015). However, various initiatives promoting ML education in K-12 have lately emerged, including several countries such as China introducing artificial intelligence and ML into curricula in primary and secondary schools (Marques *et al.*, 2020; Yang, 2019).

These instructional units teach competencies varying from presenting what is ML, to specific ML techniques, with an emphasis on artificial neural networks as well as the impacts of ML. Because of the complexity, several instructional units address only the most accessible processes, such as data management, while others cover the complete ML process in a simplified way black-boxing to different degrees some of the underlying ML processes. Typically, visual tools, such as Google Teachable Machine (Google, 2020) or customized solutions such as LearningML (Rodríguez García *et al.*, 2020) or PIC (Tang *et al.*, 2019) are adopted at this educational stage not requiring any programming. This allows students to execute an ML process in an interactive way using a train-feedback-correct cycle, enabling them to evaluate the current state of the model and take appropriate actions (Gresse von Wangenheim *et al.*, 2021). Most of these tools are

available for free online using resources in the cloud to train the ML models enabling their adoption in schools with common computer labs and internet connections (Gresse von Wangenheim *et al.*, 2021). These tools also allow the easy deployment of the created ML models into popular block-based environments, such as App Inventor, Scratch, or Snap!, which are used to teach computing in K-12.

As a part of the learning process, it is important to assess the students' learning by providing feedback to both the student and the teacher (Hattie & Timperley, 2007). For effective learning, students need to know their level of performance on a task, how their own performance relates to good performance and what to do to close the gap between those (Sadler, 1989). Despite the many efforts to address the assessment of computing education in K-12 settings, more emphasizes have been on computational thinking, algorithms & programming, and modeling & simulation (Lye & Koh, 2014; Tang *et al.*, 2019; Yasar *et al.,* 2016), while most instructional units on ML currently do not propose rigorous assessment solutions (Marques *et al.*, 2020). Few ML courses include rather simple quiz-based assessments, while performance-based assessments are basically non-existent. As one of the few existing studies, Sakulkueakulsuk *et al.* (2018) proposes an assessment based on the performance of the ML model created by the students, while AI Family Challenge (Technovation Families, 2019a) and Exploring Computer Science (2019) assess the outcome or students' presentation through rubrics. However, no further information on their design or evaluation has been encountered, thus their effectiveness and evidence for validity have remained questionable.

Therefore, this research aims to initiate the development of a scoring rubric for assessing the learning of ML concepts and practices focusing on image recognition with supervised learning. The rubric is defined as part of a performance-based assessment based on ML artifacts created by students as an outcome of the *use* stage in K-12. In this line, the following research questions were addressed:

(1) What is the evidence of internal consistency of the performance-based assessment scoring rubric?
(2) What is the evidence of content validity of the performance-based assessment scoring rubric?

## 2. Research Methodology

The development of the performance-based assessment is based on the method proposed by Moskal and Leyden (2000) and evidence-centered design (Mislevy *et al.*, 2003), including the following phases:

**Content domain analysis.** The content domain was analyzed through a systematic literature review on the definition of ML concepts and practices as well as learning objectives and evidence of these in outcomes created by middle- and high-school students.

**Definition of the scale for assessment**. As a part of an initial proposal of a scale, a scoring rubric has been defined to identify criteria with which the students' learning outcome is measured. It represents a descriptive scoring scheme (Brookhart, 1999;

Moskal, 2000) for performance-based assessments of ML artifacts created as learning outcomes (Kandlhofer *et al.*, 2016). Therefore, we identified the characteristics that are to be evidenced in a student's work to indicate proficient performance in relation to the respective learning objectives (Allen & Knight, 2009; Brookhart, 1999; Moskal, 2000). Then, for each criterion, performance levels were defined as descriptions of the different score levels.

**Evaluation of internal consistency and content validity.** To evaluate the initial proposal of the scale, we conducted an expert panel, in which the participants assess exemplary learning outcomes using the scoring rubric, and, afterward provide feedback through a questionnaire. The expert panel consisted of 16 professionals with relevant fields including machine learning and/or computing education and related areas including mathematics, computer graphics, and psychology. We evaluated internal consistency by analyzing inter-rater reliability, which relates to the issue that a student's score may differ among different raters. We used Fleiss' kappa coefficient based on the scores given by the participants concerning two ML models created as exemplary learning outcomes using the developed rubric (Fleiss *et al.*, 2003; Moskal & Leydens, 2000). Content validity was evaluated based on the questionnaire responses by analyzing correctness, completeness, clarity, and relevance, evaluating the extent to which criteria reflect the variables of the construct, and determining whether the measure is well-constructed (Moskal & Leydens, 2000; Rubio *et al.*, 2003). Content validity was analyzed through descriptive statistics and the content validity ratio proposed by Lawshe (1975). The results have been interpreted and discussed in the respective educational context.

## 3. State of the Art

To review the state of the art and practice on how ML concepts and practices are being assessed, we performed a systematic mapping study following the approach proposed by Petersen *et al.* (2008). Searching digital libraries in this field including ACM Digital Library, IEEEXplore, Scopus, arxiv, SocArXiv, and Google Scholar/Google to minimize the risk of omitting instructional units that may not have been published as scientific articles, we considered any instructional unit (e.g., course, activity, tutorial) that covers teaching ML in elementary to high school. As in several cases, we observed that courses do not necessarily focus exclusively on ML, but rather cover this topic as a part of a wider course on Artificial Intelligence (AI), we also searched for AI courses in order to minimize the risk of omission. Yet, instructional units on AI that do not cover ML topics were excluded as well as instructional units targeting other educational stages. As a result, a total of 14 instructional units were identified that also adopted some kind of assessment (Table 1) (Salvador *et al.*, 2021). Most focus on the assessment of basic ML concepts with some also covering neural networks and/or the impact of ML. The majority assesses learning on the remembering and understanding level following Bloom's taxonomy (Bloom *et al.*, 1956). Nine of the instructional units approach the assessment of learning the application of ML.

Table 1
Summary on assessments of ML concepts and practices

| Name | Reference | ML concepts | Learning level | Use-Modify-Create cycle | Assessment method | Type of feedback |
|---|---|---|---|---|---|---|
| **AI Family Challenge** | (Technovation Families, 2019a) | Basic ML concepts | Remembering, understanding, application | Use | Quiz: 1 multiple choice question | Automated correction of answers |
| | | Basic ML concepts | Remembering, understanding, application | Create | Quiz: 1 multiple choice question; Rubric: 9 criteria on 3-point performance levels | Automated correction of answers. Manual assessment by judges of the challenge |
| **Apps for good: ML course** | (Apps for Good, 2019) | Basic ML concepts Impact of ML | Remembering, understanding, application | Use, Create | Assessment questions | Manual evaluation by the instructor |
| **AI Literacy Workshop** | (Van Brummelen *et al.*, 2020) | Basic ML concepts Impact of ML | Remembering, understanding, application | Use, Modify, Create | Assessment questions | Manual evaluation by the instructor |
| **AI for Oceans** | (Code.org, 2019) | Basic ML concepts | Remembering, understanding | NA | Task completion | Automated indication of the degree of task completion, certificate. No analysis of correctness. Feedback for teachers for monitoring a class |
| **Curiosity Machine – build a neural network** | (Technovation Families, 2019b) | Neural networks | Remembering, understanding | NA | Quiz: 3 multiple-choice questions | Presentation of answers given, no automated correction |
| **Elements of AI** | (Elements of AI, 2019) | Basic ML concepts Neural networks Impact of ML | Remembering, understanding | NA | Exercises with 1–3 open-text or multiple-choice questions | Automated correction of multiple-choice answers. Presenting example answers and peer review for text answers. There is no grade, but the number of exercises completed is tracked. |
| **Alternate Curriculum Unit: AI** | (Exploring Computer Science, 2019) | Basic ML concepts Impact of ML | Remembering, understanding | NA | Rubrics with 8–14 criteria for the assessment of students' presentations | Manual evaluation by the instructor indicating the total points |
| **Machine Learning para Todos!** | (Gresse von Wangenheim *et al.*, 2020) | Basic ML concepts Neural networks Impact of ML | Remembering, understanding, application | Use | Single-question quizzes (multiple-choice, drag-and-drop, etc.), Rubric: 11 criteria on 3-point performance levels | Automated correction of quizzes, Manual evaluation by the instructor of performance-based assessment |
| **Developing Middle School Students' AI Literacy** | (Lee *et al.*, 2021) | Basic ML concepts Neural networks | Remembering, understanding | NA | Quizzes (true/false; open-text, etc.) | Manual evaluation by the instructor |

Continued on next page

Table 1 – continued from previous page

| Name | Reference | ML concepts | Learning level | Use-Modify-Create cycle | Assessment method | Type of feedback |
|---|---|---|---|---|---|---|
| **Introduction to ML: Image Classification** | (MIT App Inventor, 2019a) | Neural networks | Remembering, understanding, application | Use | Multiple choice test (3 questions) | Manual evaluation by the instructor |
| **Personal Image Classifier** | (MIT App Inventor, 2019b) | Neural networks | Remembering, understanding, application | Use | Multiple choice test (3 questions) | Manual evaluation by the instructor |
| **Ready AI AI+Me** | (ReadyAI, 2019) | Basic ML concepts Impact of ML | Remembering, understanding | NA | Single-question quizzes (multiple-choice, drag-and-drop, etc.) | Automated correction of answers Task completion tracking |
| **LearningML** | (Rodríguez García *et al.*, 2020) | Basic ML concepts Neural networks | Remembering, understanding, application | Use, Modify, Create | Quizzes, exercises, assessment question | Calculation of the performance of the created ML model |
| **Kids making AI** | (Sakulkueakulsuk *et al.*, 2018) | Basic ML concepts | Remembering, understanding | Use | Gamification allocating points in accordance with the performance of the created ML models | Manual evaluation by the instructor |

NA – not applicable, as these instructional units do not cover the application of ML concepts and practices

Most of the assessments are quite simple, in some cases consisting of single-question quizzes at the end of learning units or only monitoring task completion. An exception is Elements of AI (Elements of AI, 2019) assessing also the answers to exercises. Two courses teaching ML with MIT App Inventor (2019a, 2019b) propose tests composed of three multiple-choice questions for the assessment of the students at the end of the course. Considering that currently most of these ML courses are offered as extracurricular activities such lightweight assessment approaches may be adequate to prevent the demotivation of the students. Yet, the lack of a more rigorous assessment may impede better support for their learning and the improvement of these courses. Very few adopt performance-based assessment defining rubrics for the assessment of presentations (Exploring Computer Science, 2019), learning results (Technovation Families, 2019a), or based on performance measures of ML models created by the students (Sakulkueakulsuk *et al.*, 2018; Gresse von Wangenheim *et al.*, 2021). Due to the recentness of ML courses in K-12 education settings, most focus on the assessment of results from the *use* stage, with only Apps for Good (2019), Technovation Families (2019a), Rodríguez García *et al.* (2020), and Van Brummelen *et al.* (2020) adopting a *computational action* (Tissenbaum *et al.*, 2019) strategy that allows students to develop their own custom ML models that provide an impact on their lives and communities. To date, these assessments are to be performed manually by the instructors or judges (Technovation Families, 2019a). Only some of the quiz-based assessments are automated as a part of online courses. Instructional feedback to the student is typically limited to the indication of if the question(s) have been answered correctly,

without further guidance. Some also issue a certificate at successful course completion and to increase engagement, and Sakulkueakulsuk *et al.* (2018) adopt a gamification approach.

However, in general, the proposed assessments seem to be just emerging, lacking further information on how they have been designed or evaluated, especially when comparing them to research on assessment in computing education in K-12 in general. As a consequence, there seems to be no information on the reliability and validity of such assessments available.

## 4. Definition of the Performance-Based Assessment

Focusing on an active learning strategy taking students to create ML models with artificial neural networks, authentic assessment based on the created outcomes is an appropriate means allowing the openness of student responses, as opposed to, for example, multiple-choice assessments (Messick, 1996; Torrance, 1995). The assessment is based on the assumption that certain measurable attributes can be extracted from the artifacts created by the students during the learning process, evaluating whether the artifacts show that they have learned what they were expected to.

For performance-based assessment, typically scoring rubrics are adopted that define descriptive measures to separate levels of performance on a given task by delineating the criteria associated with learning activities (Moskal, 2000; Mc-Cauley, 2003; Whittaker *et al.*, 2001). By converting rubric scores, grades are determined in order to provide instructional feedback.

Here, we aim at the development of a scale that aims at assessing the proficiency of students on basic ML concepts. As a part of the scale, we define a scoring rubric establishing criteria used for scoring the created ML artifacts from the point of view of the instructor in the context of K-12 education, primarily middle and high school. The scoring rubric describes how observable variables summarize a student's performance in the task of developing an ML model for image recognition from the work products that are produced by the student during this task.

As currently, almost every student is a novice to ML, we focus on the *use* stage of the learning cycle on which students start to develop pre-defined ML models, for example, by following a step-by-step tutorial. The assessment is defined in conformity with the K-12 Guidelines for Artificial Intelligence (Touretzky *et al.*, 2019) referring to Big Idea 3 – Learning, AI literacy as defined by Long and Magerko (2020), covering general computing topics as proposed by the Computer Science Teachers Association (CSTA, 2017). Here, we focus exclusively on learning objectives related to the development of ML models using a supervised learning approach for image recognition enabling students to become creators of intelligent solutions (Kandlhofer *et al.*, 2016; Long & Magerko, 2020; Sulmont *et al.*, 2019; Touretzky *et al.*, 2019).

Building a human-centric manner ML application is an iterative process that requires students to complete a sequence of phases on the *use* stage (Amershi *et al.*, 2019;

Mathewson, 2019) with the help of visual ML tools such as Google Teachable Machine (Carney *et al.*, 2020; Gresse von Wangenheim *et al.*, 2021; Gresse von Wangenheim *et al.*, 2020):

**Data management:** During this step data is either collected or pre-assembled datasets are provided that may be low-dimensional to facilitate understanding or be messy on purpose to demonstrate issues of bias (D'Ignazio, 2017; Sulmont *et al.*, 2019). The data is cleaned by excluding messy images. and leaving it more balanced, including the same number of images for each category. For supervised learning, the datasets also need to be labeled. The data set is typically split into a training set to train the model and a test set to perform an unbiased performance evaluation of the model on unseen data.

**Model learning:** A ML model is typically built upon pre-trained models that have been proven effective in comparable situations by training the model with the data and using a specific learning algorithm. Training parameters, for example, the learning rate, epochs, and batch size are specified to improve performance. After the transfer learning step, the performance of the model can also be improved by hyper tuning the learning.

**Model evaluation:** The model can be tested with new images that have not been used for training. In addition, performance metrics (e.g., accuracy) can be analyzed and interpreted, identifying possible improvement opportunities. The performance results can also be visualized as a confusion matrix, a table that in each row presents the number of examples of predicted categories while each column represents the number of examples of actual classes, facilitating the identification of data that is not classified correctly.

Considering the application of this ML process in the *use* stage, other phases of the human-centric ML process are not considered, as requirements are typically pre-defined and the model deployment and monitoring phase may represent additional content in combination with other computing/programming courses. And, adopting deep learning, feature design is shifted to the underlying learning system along with classification learning. Furthermore, to support students in their first steps to start to understand ML, certain fine-grained details of the neural network structure may be concealed as black boxes to lessen cognitive load (Resnick *et al.*, 2000). Based on this domain analysis, the learning objectives are defined as presented in Table 2.

Table 2

Learning objectives related to the development of ML models on the use stage

| ID | Learning objective | Source |
|---|---|---|
| **LO1** | Collect, clean and label data for the training of an ML model; understand how ML algorithms are influenced by data | Based on the human-centric ML process (Amershi *et al.*, 2019; Touretzky *et al.*, 2019; Long and Magerko, 2020) |
| **LO2** | Train an ML model | (Touretzky *et al.*, 2019; Long and Magerko, 2020) |
| **LO3** | Evaluate the performance of the ML model | Based on the human-centric ML process (Amershi *et al.*, 2019; Long and Magerko, 2020) |

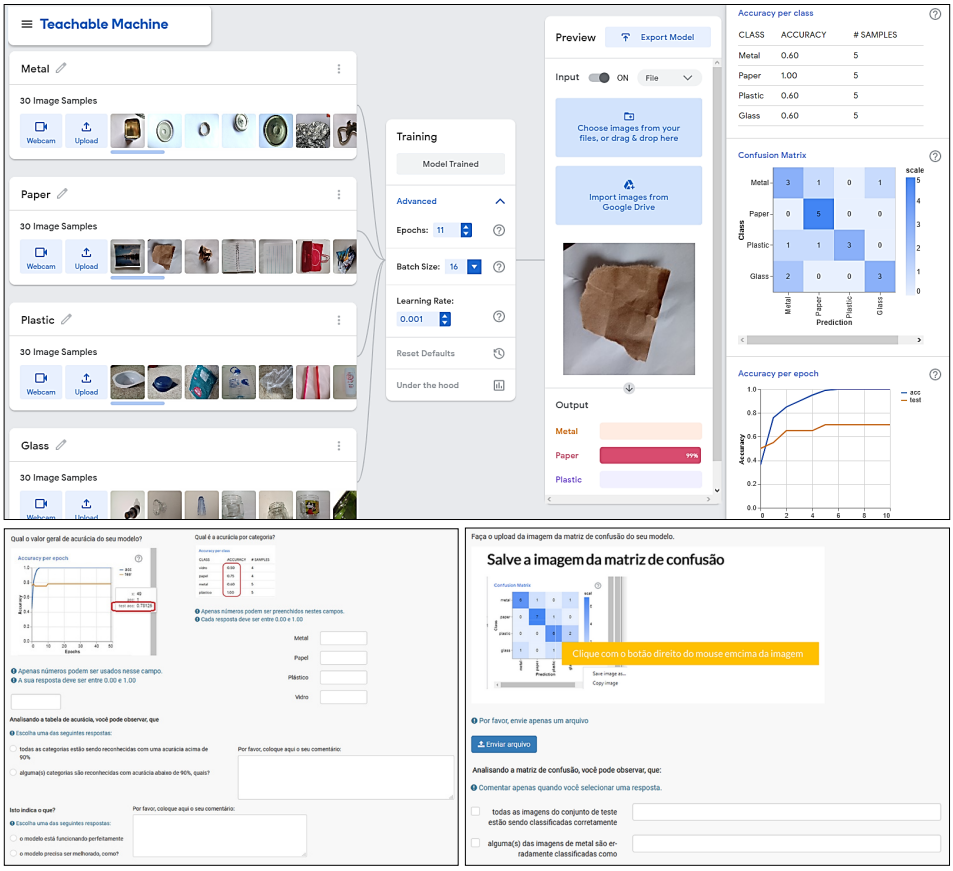



Fig. 2. Example outcome created with Google Teachable Machine and collected via an online questionnaire.

Concerning these learning objectives, we defined a performance-based assessment based on the artifacts developed by the students as outcomes of the learning process. Adopting a visual tool Google Teachable Machine (2020) for teaching the creation of ML models at this educational stage (Carney *et al.*, 2020), evidence for the achievement of these learning objectives can be obtained based on the ML artifacts developed by the students, including the prepared dataset, model training parameters, and evaluation results (Fig. 2).

Therefore, we define an initial scale defining the items to be measured to assess the ability to develop an ML model indirectly inferring the achievement of ML competencies. The criteria to be used in scoring the artifacts created by the students during the development of an ML model are defined as a rubric (Table 3). These scores can be used to provide instructional feedback guiding the students' learning as well as indicating improvement opportunities concerning the instructional unit. Performance levels are defined on a 3-point ordinal scale ranging from *poor* to *good* based on typical learning outcomes expected at this educational stage.

Table 3

Scoring rubric for application of ML concepts for image recognition – *use* stage

| Criterion | Performance levels | | |
|---|---|---|---|
| | Poor – 0 pt. | Acceptable – 1 pt. | Good– 2 pt. |
| **Data management (LO1)** | | | |
| C1. Quantity of images | Less than 5 images per category | 6 to 10 images per category | More than 10 images per category |
| C2. 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 |
| C3. 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 |
| C4. 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 |
| C5. 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 (LO2)** | | | |
| C6. 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 (LO3)** | | | |
| C7. Tests with new objects | No object tested | 1–2 object tested | More than 2 objects tested |
| C8. Interpretation of tests | Wrong interpretation | (Not applicable) | Correct interpretation |
| C9. 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 |
| C10. 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 |
| C11. 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 |

## 5. Evaluation of the Scoring Rubric

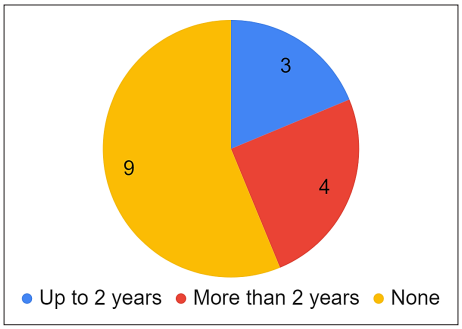
In order to analyze the quality of the scoring rubric, in terms of internal consistency and content validity, we conducted a series of scientific procedures including an expert panel review and statistical analysis to determine primarily psychometric properties of the rubric.

### 5.1. *Definition of the Evaluation*

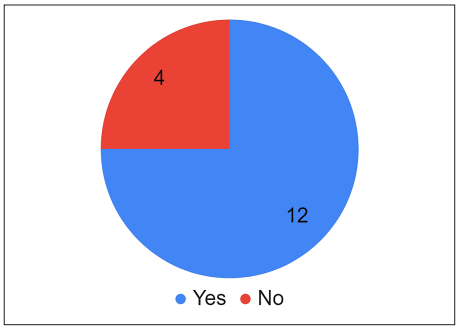
As a part of the evaluation, the experts performed assessments of two ML models created as exemplary learning outcomes using the developed rubric. The artifacts created as learning outcomes include the prepared dataset, the training parameters, as well as an evaluation report, documenting the tests run, and the analysis and interpretation of the evaluation results (Fig. 2). On purpose, we prepared one weak learning outcome (e.g., few images in the dataset, few test runs) and one strong one that satisfies almost all criteria at the highest performance level. Internal consistency was evaluated regarding the inter-rater reliability of the assessments by the experts. Once experienced with the application of the rubric, the experts provide further feedback concerning content validity with respect to correctness, completeness, clarity, and relevance (Lawshe, 1975; Moskal and Leydens, 2000; Rubio *et al.*, 2003). Each question is rated dichotomically by the experts, suggesting changes when necessary. The priority of each of the assessment criteria is rated on a 3-point ordinal scale ranging from not relevant to essential.

### 5.2. *Execution of the Evaluation*

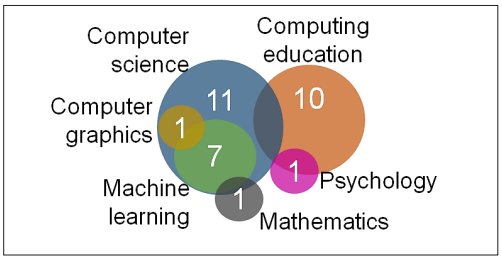
We systematically selected participants from the Computing in School initiative at the Federal University of Santa Catarina and external participants, who are recognized experts with academic and/or practical experience in the subject matter. The participants were invited via email explaining the objective of the evaluation and assuring confidentiality. Participation was voluntary. Instructions and data collection forms were made available online. We invited 20 experts and obtained a response rate of 80% (n = 16). The majority of the participants have experience and knowledge in machine learning and/or computing education, while also including experts from related areas such as mathematics, computer graphics, and psychology enabling the collection of feedback from different points of view (Fig. 3). Although most participants are researchers, four participants are K-12 teachers representing directly the target audience.




Practical experience in teaching computing subjects in K-12




Experience in creating a Machine Learning model




Number of participants per expertise knowledge area
(As some experts have expertise in more than one knowledge area, the sum of participants in different areas is bigger than the number of the participants)

Fig. 3. Demographic characteristics of the expert panel.

## 5.3. *Results of the Evaluation*

### 5.3.1. *What is the Evidence of Internal Consistency of the Performance-based Assessment Scoring Rubric?*

In order to evaluate if the design of the scoring rubric allows a reliable assessment (Moskal & Leydens, 2000), we analyzed inter-rater reliability among the responses of the experts' assessments of the two examples of ML learning outcomes. For the analysis, we used Fleiss Kappa (Fleiss *et al.*, 2003), a measure that extends Cohen's Kappa for the level of agreement between two or more assessors as in our case 16. Commonly, values below 0 indicate less than chance agreement, values between 0.01–0.20 indicate slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–0.99 almost perfect agreement (Landis & Koch, 1977).

**Inter-rater reliability.** Analyzing the 11 items of the scoring rubric using the assessment from 16 experts, we obtained a value of Fleiss kappa = 0.617 indicating substantial agreement. This is confirmed by the p-value ($p < 0.0001$), indicating that the kappa value is significantly different from zero. Individual kappa values for each of the performance levels separately were also computed and compared to all other categories together (Table 4).

Table 4
Computed Kappa per performance level

| Performance level | Poor – 0 pt. | Acceptable – 1 pt. | Good – 2 pt. |
|---|---|---|---|
| Fleiss Kappa | 0.692 | 0.544 | 0.614 |

Table 5
Computed Kappa for the assessment of exemplary weak learning outcome

| Performance level | Poor – 0 pt. | Acceptable – 1 pt. | Good – 2 pt. |
|---|---|---|---|
| Fleiss Kappa | 0.606 | 0.545 | 0.180 |

Table 6
Computed Kappa for the assessment of exemplary strong learning outcome

| Performance levels | Poor – 0 pt. | Acceptable – 1 pt. | Good – 2 pt. |
|---|---|---|---|
| Fleiss Kappa | 0.455 | 0.541 | 0.526 |

A substantial agreement between assessors can be observed on the lowest (i.e., poor) and highest (i.e., good) performance level, while on the other hand, only a moderate agreement on the intermediate performance level (i.e., acceptable) (Table 4). This points out that it seems easier to recognize very good or very poor performance, rather than an intermediate performance that may be classified as either good or poor respectively by some assessors.

We also computed individual kappa values for each of the performance levels separately for the exemplary weak and strong learning outcomes assessed by the experts.

Here, we can observe substantial agreement between assessors on the lower performance levels, while they demonstrated only a slight agreement on the highest performance level (Table 5). It seems rather surprising that even for a learning outcome that has been constructed as weak on purpose, some assessors still rated some criteria on the highest performance level. While in one case such a higher assessment may be related to the specific perception of some assessors, other criteria such as C2 and C4 demonstrated a variance across all three performance levels, indicating that the assessment of the relevance and categorization of the images can be difficult to judge.

Different from these results, higher performance levels are more consistently rated than the lowest level concerning the exemplary strong learning outcome (Table 6).

While criteria C2 and C4 have been rated much more uniform in this case, criteria C5, C8, and C10 presented variances across all performance levels. The divergence regarding C5 may be due to the low quality of the images presented to the experts making the identification of messy images difficult. The disagreement concerning the interpretation criteria may be due to different ML knowledge levels of the assessors, indicating a need for well-trained K-12 teachers and/or automated support for the assessment.

Yet, for most items, the majority of assessors agreed on the same rating. Initial results, thus, demonstrate that in general, a substantial agreement can be achieved. However, larger-scale studies are required to study the differences observed on a broader variety of learning outcomes.

### 5.3.2. *What is the Evidence of Content Validity of the Performance-based Assessment Scoring Rubric?*

Most participants considered the criteria and performance levels in general as correct (88%), complete (75%), and clear (63%).

**Correctness:** Concerning correctness, three experts observed that the criteria related to the interpretation of the confusion matrix may not admit the possibility that object classification errors are not identified correctly, while the interpretation of the model is correct. As criterion C10 combines these two aspects, either additional performance levels have to be added to comprehensively represent all combinations or the criterion needs to be split into two separate ones. Another suggestion is also related to a more detailed refinement of performance levels, e.g., by dividing the highest performance level of criterion *C7. Tests with new objects* into several ones.

**Relevance:** All items of the rubric have been considered most essential on a 3-point ordinal scale ranging from irrelevant to essential, with few experts considering some criteria as only desirable (Fig. 4). None of the criteria has been considered to be irrelevant.

Analyzing the content validity ratio defined as CVR = (Ne-N/2) / (N/2), in which Ne is the number of experts marking essential and N is the total number of experts, the adequacy of the rubric was also confirmed. Only criteria *C3. Distribution of the dataset* has a CVR = 0.38 below the threshold of 0.49 (Lawshe, 1975), as in this case four experts considered the criteria only desirable but not essential. Consequently, this criterion could be excluded to minimize the assessment effort.

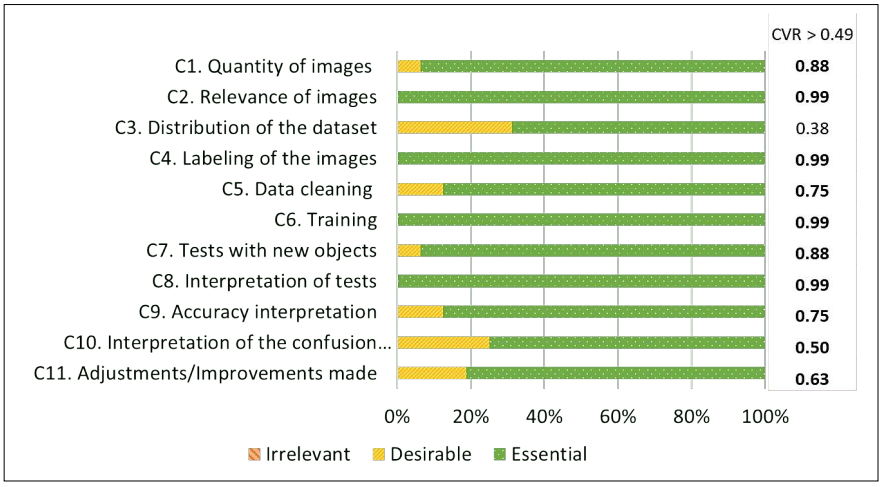



Fig. 4. Relevance of the criteria.

**Completeness:** Some participants suggested new criteria, i.e., on checking if the student did not reuse images of the training set to test the model and/or if the tests included at least one test for each image category. Other suggestions requiring the inclusion of further learning content and/or reporting by the student, include the assessment of reflections of the student his/her tests to be sufficient, and further data pre-processing activities.

**Clarity:** Some items are quite subjective compared to the other ones, which can cause uncertainty and inaccuracies from the point of view of the assessor. This may also be the reason for a lack of higher inter-rater agreement. Yet, observing a disagreement even on quantifiable criteria, such as *C1. Quantity of images* may also indicate other factors. Some participants suggested revising the wording to be less technical in order to be more easily understood by non-computing K-12 teachers. Furthermore, as some criteria depend on the specific ML model developed by the students, e.g., C9 and C10 aiming at assessing if the student correctly identified categories with low accuracy and correctly interpreted the evaluation results, substantial ML knowledge and effort from the assessor is required. As this may not be given currently in the context of K-12 education on a larger scale, a possible solution would be to automatize the assessment, refining the criteria through fine-grained rules and/or adopting Machine Learning techniques to assess subjective criteria.

All experts considered the rubric applicable in K-12 education, taking into consideration a careful definition of the specific learning objectives and strategies, as well as its complementation by other types of assessments to obtain a more comprehensive understanding of the students' learning performance.

## 6. Discussion

Considering the importance of innovative educational opportunities for young people to gain a better understanding of ML concepts and practices to succeed in the 21st century, we aim to teach ML primarily in middle and high school by proposing a scoring rubric for the performance-based assessment of ML learning outcomes.

In this regard, the presented research aims at advancing the current state of the art like, different from most other approaches using quizzes or tests, proposes a scoring rubric based on the ML artifacts created by the students. The few other scoring rubrics for this kind of assessment encountered during the literature review either aim at the assessment of the end result in a more abstract way or on the presentation of the result. For example, the rubric to be used by judges in the AI Family Challenge (Technovation Families, 2019a) focuses on a general assessment of the end result in the challenge taking students to create their own intelligent solutions. It includes criteria on the ideation such as "How well does the team's invention solve the problem in their community?", project development, pitch and communication, and overall expression (i.e., How much does the submission stand out from others?). Regarding project development, the rubric only includes three criteria (How well does the invention use AI or other technologies?; How well thought out is the team's prototype or plan to create a prototype?; Does the

invention solve the problem in a unique way?) not covering in a more detailed way the ML concepts to be learned. Another example of a scoring rubric is given as part of the Alternate Curriculum Unit: AI (Exploring Computer Science, 2019). This rubric is used for the assessment of the presentation of the final results of the student including only criteria related to presentations such as content quality, presentation quality, image and video presentation, use of English conventions. In this way, the proposed scoring rubric represents the first step for a more detailed assessment of the ML artifacts created as learning outcomes.

Furthermore, to date, no rigorous information on any kind of evaluation of the reliability and validity of the proposed assessment approaches in literature has been explicitly encountered, thus our research also stands out by conducting an initial evaluation through an expert panel. Results of this initial evaluation confirmed mostly the definition of the established criteria and performance level descriptors as a part of the proposed scale. Using the assessment from 16 experts, a value of Fleiss kappa = 0.617 showed that a substantial inter-rater agreement can be obtained using the scale. It seems however easier to recognize very good or very poor performance, rather than an intermediate performance that may be classified as either good or poor depending on the assessor. Observed inconsistencies between the assessments of different assessors may also be related to the specific perception of some assessors and their proficiency level in Machine Learning, pointing out the need for well-trained K-12 teachers to enable reliable assessments.

Some criteria such as the one related to the inclusion of messy images may also be difficult to be assed manually, indicating an opportunity for the automation of this kind of performance assessment to facilitate the assessment and achieve more consistent results, while also reducing effort related to the assessment in practice.

Regarding content validity, all experts considered the scoring rubric applicable in K-12 education, taking into consideration a careful definition of the specific learning objectives and strategies, as well as its complementation by other assessments to obtain a more comprehensive understanding of the students' learning performance. This is also confirmed based on the results of the analysis of the content validity ratio. Only one exception, criteria *C3. Distribution of the dataset* with a CVR = 0.38 demonstrated results below the expected threshold as four experts considered the criteria desirable but not essential. Yet, as it is more probable to achieve acceptable performance of an ML model using a well-balanced dataset with more or less the same quantities of images for each of the categories, we still consider this an important assessment criterion to be revised in further studies.

With exception of one criterion related to the interpretation of the confusion matrix that seems to combine two different criteria and should therefore be separated into different criteria and a suggestion to refine the criteria related to the number of tests performed with new objects into more performance levels, all experts considered the criteria correct.

Regarding completeness, some new criteria have been suggested such as checking if the student did not reuse images used during training for testing and/or if the tests included at least one test for each image category. Concerning clarity of the definition of

the criteria and performance level descriptors, some participants suggested revising the wording to be less technical to be more easily understood by non-computing K-12 teachers, again pointing out the need for teachers to be well-trained in Machine Learning. Therefore, some criteria might need to be further refined as well as additional criteria to be added for a more comprehensive assessment.

**Threats to validity.** To mitigate threats related to the research design, we systematically developed the scoring rubric based on an analysis of the educational context adopting methods for rubric definition and conducted an initial evaluation in the form of an expert panel. Another threat is related to the diversity and sample size of the participants in the evaluation. Although primarily including experts from the computing area, the panel does cover diverse areas of interest, representing diverse points of view, as well as the target audience including K-12 teachers. In terms of size, there is also evidence that 16 experts are sufficient to draw results (Lawshe, 1975). To reduce threats associated with the data analysis, we conducted a statistical evaluation following Lawshe (1975), Moskal and Leydens (2000), and Rubio *et al.* (2003). We followed the methodology proposed by Lawshe (1975) to minimize any impact of bias due to the subjectiveness of the experts' feedback. Representing only an initial evaluation, larger-scale studies are necessary to confirm the results and analyze open issues.

## 7. Conclusion

Based on the domain analysis and modeling regarding the teaching and learning of basic ML concepts, we propose a scoring rubric as part of a scale for the performance-based assessment of ML learning outcomes in the context of K-12 computing education. Results of an initial evaluation confirmed mostly the definition of the established criteria and performance level descriptors as a part of the proposed scale, indicating a substantial inter-rater agreement as well as content validity in terms of correctness, relevance, completeness, and clarity.

Thus, based on first positive feedback, the proposed scoring rubric presents a first step for the assessment of open-ended ML learning activities regarding image recognition with supervised learning, which can be used by instructional designers and researchers to evolve support for the assessment in the context of teaching Machine Learning in K-12 as well as by instructors to assess the outcomes of students in this educational context. Of course, although our focus in this article is on the proposal of a performance-based assessment based on learning outcomes, in educational practice this kind of assessment should be completed by other types of assessment such as observations or interviews.

Based on these results, we are currently revising the initial scale. We are also planning further studies to collect data based on learning outcomes created by students aiming at the development of a measurement model using Item Response Theory as a part of the evidence model that gives information about the connection between the student model variables and observable variables.

## 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. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. This work was also partially supported by the AI.R-NISTH AI for Social Good Research Grant 2021, Nanyang Technological University in Singapore.

## References

Allen, S., Knight, J. (2009). A Method for Collaboratively Developing and Validating a Rubric. *International Journal for the Scholarship of Teaching and Learning*, 3(2), article 10.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., Zimmermann, T. (2019). Software engineering for machine learning: A case study. In Proc. of the *IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice*, IEEE Press, 291–300.

Apps for Good. (2019). *Apps for Good Machine Learning*. Apps for Good. `https://www.appsforgood.org/courses/machine-learning`

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

Brookhart, S.M. (1999). *The Art and Science of Classroom Assessment*. ASHE-ERIC Higher Education Report, 27(1).

Carney, M., Webster, B., Alvarado, I., Phillips, K., Howell, N., Griffith, J., Jongejan, J., Pitaru, A., Chen, A. (2020). Teachable Machine: Approachable Web-based tool for exploring machine learning classification. In Proc. of the *CHI Conference on Human Factors in Computing Systems, ACM*, 1–8.

Code.org. (2019). *Lesson 9: AI For Oceans*. `https://curriculum.code.org/hoc/plugged/9/`

Computer Science Teachers Association. (2017). *K–12 Computer Science Standards*. K12cs.Org. `http://k12cs.org`

D'Ignazio, C. (2017). Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Information Design Journal*, 23(1), 6–18.

Elements of AI. (2019). *Machine learning*. `https://course.elementsofai.com/4`

Evangelista, I., Blesio, G., Benatti, E. (2018). Why Are We Not Teaching Machine Learning at High School? A Proposal. In Proc. of the *World Engineering Education Forum*, Albuquerque, NM, USA, 1–6.

Exploring Computer Science. (2019). *Artificial Intelligence—Alternate Curriculum Unit*. Exploring Computer Science. `http://www.exploringcs.org/for-teachers-districts/artificial-intelligence`

Fleiss, J.L., Levin, B., Paik, M.C. (2003). *Statistical Methods for Rates and Proportions* (3rd ed). John Wiley & Sons, Inc.

Google. (2020). *Google Teachable Machine*. `https://teachablemachine.withgoogle.com/`

Gresse von Wangenheim, C., Hauck, J.C.R., Pacheco, F.S., Bertonceli Bueno, M.F. (2021). Visual Tools for Teaching Machine Learning in K-12: A Ten-Year Systematic Mapping. *Education and Information Technologies*, 26(5), 5733–5778.

Gresse von Wangenheim, C., Marques, L.S., Hauck, J.C.R. (2020) Machine Learning for All – Introducing Machine Learning in K-12. SocArXiv. DOI: 10.31235/osf.io/wj5ne.

Hattie, J., Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.

Hitron, T., Orlev, Y., Wald, I., Shamir, A., Erel, H., Zuckerman, O. (2019). Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes. In Proc. of the *Conference on Human Factors in Computing Systems*, ACM, Paper 415, 1–11.

Ho, J.W., Scadding, M. (2019). Classroom Activities for Teaching Artificial Intelligence to Primary School Students. In Proc. of the *Int. Conference on Computational Thinking*, Hong Kong, China, 157.

Holzinger A., Kieseberg P., Weippl E., Tjoa A.M. (2018). Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. In Proc. of the *Int. Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Springer.

House of Lords of the UK Parliament. (2018). *AI in the UK: Ready, Willing and Able?,* Report, HL Paper 100, UK. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

Hubwieser, P., Giannakos, M.N., Berges, M., Brinda, T., Diethelm, I., Magenheim, J., Pal, Y., Jackova, J., Jasute, E. (2015). A global snapshot of computer science education in K-12 schools. In Proc. *ITiCSE on Working Group Reports*, Vilnius, Lithuania, 65–83.

Jordan, M.I. and Mitchell, T.M. (2015). Machine Learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260.

Kahn, K., Lu, Y., Zhang, J., Winters, N., Gao, M. (2020). Deep learning programming by all. In Proc. of the *Conference on Constructionism*, Dublin, Ireland.

Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., Huber, P. (2016). Artificial intelligence and computer science in education: From kindergarten to university. In Proc. of *2016 IEEE Frontiers in Education Conference*, Erie, PA, USA, 1–9.

Landis, J.R., Koch, G.G. (1977). The measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563–575.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature,* 521**,** 436–444.

Lee, I., Ali, S., Zhang, H., DiPaola, D., Breazeal, C. (2021). Developing Middle School Students' AI Literacy. In Proc. of *52nd ACM Technical Symposium on Computer Science Education*, Association for Computing Machinery, New York, NY, USA, 191–197.

Long, D., Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In Proc. of *2020 CHI Conference on Human Factors in Computing Systems*, ACM, 1–16.

Lwakatare, L.E., Raj, A., Bosch, J., Olsson, H.H., Crnkovic, I. (2019). A Taxonomy of Software Engineering Challenges for Machine Learning Systems: An empirical Investigation. In: Kruchten P. *et al.* (Eds) Agile Processes in Software Engineering and Extreme Programming. Lecture Notes in *Business Information Processing*, 355, Springer, Cham.

Lye, S.Y., Koh, J.H.L. (2014). Review on teaching and learning of computational thinking through programming: What is next for K-12?. *Computers in Human Behavior*, 41(C), 51–61.

Lytle, N., Cateté, V., Boulden, D., Dong, Y., Houchins, J., Milliken, A., Isvik, A., Bounajim, D., Wiebe, E., Barnes, T. (2019). Use, modify, create: Comparing computational thinking lesson progressions for stem classes. Proc. of the *ACM Conference on Innovation and Technology in Computer Science Education*, 395–401.

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

Mathewson, K.W. (2019). A Human-Centered Approach to Interactive Machine Learning. arXiv:1905.06289v1 [cs.HC].

McCauley, R. (2003). Rubrics as assessment guides. *Newsletter ACM SIGCSE Bulletin*, 35(4).

McGovern, A., Tidwell, Z., Rushing, D. (2011). Teaching Introductory Artificial Intelligence through Java-Based Games. In Proc. of *Second AAAI Symposium on Educational Advances in Artificial Intelligence,* San Francisco, CA, USA.

Messick, S. (1996). Validity of performance assessments. In G. Phillips (Ed.), *Technical Issues in Large-scale Performance Assessment*. Washington, DC: National Center for Education Statistic

Mislevy, R.J., Almond, R.G., Lukas, J.F. (2003). *A Brief Introduction to Evidence-Centered Design*. Research Report RR-03-16, Research & Development Division Princeton, NJ, USA.

MIT App Inventor. (2019a). *Introduction to Machine Learning: Image Classification*. http://appinventor.mit.edu/explore/resources/ai/image-classification-look-extension

MIT App Inventor. (2019b). *Personal Image Classifier*. http://appinventor.mit.edu/explore/resources/ai/personal-image-classifier

Moskal, B.M. (2000). Scoring rubrics: What, when and how?. *Practical Assessment, Research, and Evaluation*, 7(1), article 3.

Moskal, B.M., Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, and Evaluation*, 7(1), 10.

Pedró, F., Subosa, M., Rivas, A., and Valverde, P. (2019). *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development*. UNESCO.

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

Ramos, G., Meek, C., Simard, P., Suh, J., Ghorashi, S. (2020). Interactive machine teaching: A human-centered approach to building machine-learned models. *Human–Computer Interaction*, 35(5–6), 413–451.

ReadyAI. (2019). *AI + ME*. https://edu.readyai.org/courses/aime/

Resnick, M., Berg, R., Eisenberg, M. (2000). Beyond black boxes: Bringing transparency and aesthetics back to scientific investigation. *The Journal of the Learning Sciences*, 9(1), 7–30.

Rodríguez García, J.D., Moreno-León, J., Román-González, M., Robles, G. (2020). LearningML: A Tool to Foster Computational Thinking Skills Through Practical Artificial Intelligence *Distance Education Journal*, 20(63), article 7.

Royal Society, (2017). *Machine Learning: The Power and Promise of Computers that Learn by Example*. Technical Report.

Rubio, D.M., Berg-Weger, M., Tebb, S.S., Lee, E.S., Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, 27(2), 94–104.

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, P., Surareungchai, W., Pataranutaporn, P., Subsoontorn, P. (2018). Kids making AI: Integrating machine learning, gamification, and social context in STEM education. In Proc. of the *IEEE International Conference on Teaching, Assessment, and Learning for Engineering*, IEEE Press, 1005–1010.

Salvador, G. de C., Gresse von Wangenheim, C., Rauber, M.F., Garcia, A., Borgatto, A.F. (2021). Avaliação de Aprendizagem de Machine Learning na Educação Básica: Um Mapeamento da Literatura. In Proc. of the 29th *Workshop sobre Educação em Computação,* online, Brazil.

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

Sulmont, E., Patitsas, E., Cooperstock, J.R. (2019). Can You Teach Me To Machine Learn? In Proc. of *50th ACM Technical Symposium on Computer Science Education*, ACM, 948–954.

Tang, D., Utsumi, Y., Lao, N. (2019). PIC: A Personal Image Classification Webtool for High School Students. In Proc. of the *IJCAI EduAI Workshop*, Macao, China.

Technovation Families. (2019a). *AI Families Challenge*. `https://www.curiositymachine.org/about/`

Technovation Families. (2019b). *Build a Neural Network*. `https://www.curiositymachine.org/challenges/126/`

Tissenbaum, M., Sheldon, J., Abelson, H. (2019). From computational thinking to computational action. *Communications of the ACM*, 62(3), 34–36.

Torrey, L.A. (2012). Teaching problem-solving in algorithms and AI. In Proc. *Third AAAI Symposium on Educational Advances in Artificial Intelligence*. Toronto, Ontario, Canada.

Torrance, H. (1995). *Evaluating Authentic Assessment: Problems and Possibilities in New Approaches to Assessment*. Buckingham: Open University Press.

Touretzky, D., Gardner-McCune, C., Martin, F., Seehorn, D. (2019). Envisioning AI for k-12: What should every child know about AI? In Proc. of *AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA, 9795–9799.

Tuomi, I. (2018). *The Impact of Artificial Intelligence on Learning, Teaching, and Education.* Report EUR 29442 EN, Publications Office of the European Union, Luxembourg.

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

Webb, M.E., Fluck, A., Magenheim, J., Malyn-Smith, J., Waters, J., Deschênes, M., Zagami, J. (2021). Machine learning for human learners: opportunities, issues, tensions and threats. *Education Technology Research Development*, 69, 2109–2130.

Yang, X. (2019). *Accelerated move for AI education in China*. ECNU Review of Education, 2(3), 347–352.

Whittaker, C.R., Salend, S.J., Duhaney, D. (2001). Creating instructional rubrics for inclusive classrooms. *Teaching Exceptional Children*, 34(2), 8–13.

Wong, G.K., Ma, X., Dillenbourg, P., Huan, J. (2020). Broadening artificial intelligence education in K-12: Where to start? *ACM Inroads*, 11(1), 20–29.

Yasar, O., Veronesi, P., Maliekal, J., Little, L., Vattana, S.E., & Yeter, I.H. (2016). Computational pedagogy: Fostering a new method of teaching. *Computers in Education Journal, 7*(3), 51–72.

**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 that aims at bringing computing education to schools in Brazil. She received the Dipl.-Inform. and Dr. rer. nat. degree 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. She is also PMP – Project Management Professional and MPS.BR Assessor and Implementor.

**M.F. Rauber** 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 is a full professor in the Informatics area at the Instituto Federal Catarinense (IFC), Camboriú. 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.

**N. da Cruz Alves** 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. She received her BSc (2017) and MSc (2019) in Computer Science from the Federal University of Santa Catarina. Her main research interests are computing education, creativity, and assessment.

**J.C.R. Hauck** holds a PhD in Knowledge Engineering and a Master's Degree in Computer Science from the Federal University of Santa Catarina (UFSC) and a degree in Computer Science from the University of Vale do Itajaí (UNIVALI). He was a visiting researcher at the Regulated Software Research Center – Dundalk Institute of Technology – Ireland. He is currently a professor in the Department of Informatics and Statistics at the Federal University of Santa Catarina, member of the Software Quality Group (GQS) and member of the initiative Computing at Schools.

**I.H. Yeter**, Ph.D., is an Assistant Professor in the National Institute of Education (NIE) at Nanyang Technological University (NTU) in Singapore. Previously, he was appointed as a Postdoctoral Research Fellow at Harvard University and Purdue University. He is currently the Director of the World MOON Project and an affiliated faculty member of CRADLE@NTU. He has worked on several international and national education-focused projects including computational thinking, engineering, artificial intelligence, and machine learning.