Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrating Machine Learning into K–12 Classrooms: Enhancing Critical Thinking Through Interactive AI Tasks

Merve AYDIN

merveaydin@trabzon.edu.tr

Trabzon University

Neriman TOKEL

Kastamonu University

Article

Keywords: Critical Thinking, Machine Learning, Machine Learning for Kids, Middle School Education, AI Literacy, Educational Technology

Posted Date: July 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9818579/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Integrating Machine Learning into K -12 Classrooms: Enhancing Critical Thinking Through Interactive AI Tasks

Abstract

Developing critical thinking (CT) is essential for helping students analyze complex problems and make informed decisions in data-intensive digital environments shaped by coding and artificial intelligence (AI). This study examines how machine learning (ML)–based instructional interventions shape middle school students' engagement in CT processes and their performance on CT indicators. Using an explanatory case study design, the intervention involved 15 fifth- and sixth-grade students who completed five ML activities (text-, image-, and number-based) over eight weeks using the Machine Learning for Kids platform. Data were collected through researcher-designed worksheets, screen recordings, reflection prompts, and semi-structured interviews. CT performance was assessed with a researcher-developed performance rubric rather than a standardized CT test, and cluster analysis was used to identify patterns in students' responses across ML phases (training, testing, and application). Results suggest that ML activities can support observable CT performance and CT-related behaviours, particularly goal-oriented reasoning, application-based evaluation, and inference. Higher-performing students more consistently identified relevant concepts during training and testing and engaged in analytic monitoring across the task cycle, whereas lower-scoring students struggled with selecting relevant concepts and completing the ML cycle, indicating a need for stronger scaffolding during learning phases. Overall, observed change across the five activities highlights the value of sequencing ML tasks and suggests that application-oriented tasks may elicit higher-order thinking in authentic contexts.

Keywords: Critical Thinking, Machine Learning, Machine Learning for Kids, Middle School Education, AI Literacy, Educational Technology

Introduction

The onset of the 21st century has marked a paradigmatic shift in educational systems, driven by the accelerated integration of digital technologies, algorithmic infrastructures, and artificial intelligence (AI) into pedagogical frameworks. The widespread use of digital technologies, especially coding and algorithm-based systems and AI applications, has led to radical changes in the way students access and process information (Halpern, 2014). In this context, the integration of technology into education is no longer just an effort to increase access and participation; it requires the development of high-level cognitive skills, especially CT.

CT, as conceptualized by Facione (1990), encompasses purposeful and self-regulatory judgment processes, including interpretation, analysis, evaluation, inference, and explanation. Rather than a peripheral skill, CT has become a central competency in contemporary education, equipping learners to navigate uncertainty, engage in reflective problem-solving, and make informed decisions. This increasing emphasis on CT is evident in how educational systems around the world are integrating it into curricula, not just as content, but through pedagogical practices. Strategies such as inquiry-based learning, structured argumentation, digital simulations, and authentic problem-solving tasks are being deployed to foster these competencies from the early stages of formal education (Alsaleh, 2020; Elder & Paul, 2020; Wing, 2006). These shifts reflect a growing recognition that equipping students with such metacognitive tools is essential in navigating the complexities of 21st-century knowledge societies.

Integrating ML into classroom practices has recently emerged as a promising pathway for fostering CT skills among students. As a subfield of AI, ML enables systems to identify patterns and make decisions based on data without explicit programming, closely mirroring cognitive processes such as inference, reflection, and feedback-based reasoning (Goodfellow et al., 2016). Platforms like Machine Learning for Kids provide young learners with accessible, visual environments to engage in classification, prediction, and model refinement tasks that inherently require metacognitive engagement and iterative problem-solving (Chung & Shamir, 2021; Shamir & Levin, 2021). Studies show that even middle school students can meaningfully participate in ML activities, thereby improving their computational and analytical thinking (Çakiroğlu & Selçuk, 2024; Martins & von Wangenheim, 2022).

ML activities also engage habits of mind associated with CT by requiring students to identify features, interpret educational outcomes, identify errors, and improve output, paralleling processes of questioning

knowledge. This makes ML not only a technical competency but also a powerful pedagogical framework for promoting CT skills in K-12 contexts (Irgens et al., 2022; Arndt et al., 2025). The cognitive demands of ML tasks—problem formulation through feature selection, model building and testing, systematic error control, and reasoned iteration—closely align with the core components of CT APA Delphi consensus, which are interpretation, analysis, evaluation, inference, explanation, and self-regulation (Facione, 1990). This alignment suggests that ML activities can make CT visible not merely as "content" but as a designed practice; particularly in situations where decision criteria are explained, evidence is weighed, and claims are reviewed, reinforcing students' habits of evidence-based reasoning and justification (Irgens et al., 2022). Classroom reports and conceptual frameworks suggest that ML's capacity to embed customization, real-time data analysis, and adaptive feedback into learning processes can make students' inference and evaluation behaviors in inquiry-based reasoning cycles more visible and measurable (Irgens et al., 2022). In parallel, mapping studies in K-12 computing/ML emphasize that CT measures are often limited and poorly aligned with the framework; therefore, there is a need for research where ML-integrated designs are clearly matched with CT outcomes (Arndt et al., 2025).

On the other hand, while the inquiry-based learning (IBL) literature reports positive effects on problem solving and CT indicators, a significant portion of studies focus on traditional IBL applications and do not sufficiently leverage the potential of advanced technologies such as ML (Sonsun et al., 2023; Sreejun & Chatwattana, 2023). Examining the potential ML-CT effect is critical not only to bridge the existing gap but also to (i) align data-driven feedback loops in IBL with CT components, (ii) test transfer in complex problem-solving contexts, and (iii) justify design principles (e.g., transparency, criterion-based assessment, pedagogical use of error analysis) through theoretical-empirical bridges (Arndt et al., 2025; Sonsun et al., 2023; Sreejun & Chatwattana, 2023).

Despite this advantage, empirical research investigating the relationship between ML education and CT development remains limited, particularly among middle school students. Existing studies focus on computational thinking or ML literacy as separate outcomes without fully exploring their connections to broader cognitive development (Hutchins & Biswas, 2024). Furthermore, ML is often perceived as too advanced for younger students, leading to its under-representation in K-12 curricula (Shrigoud & Agrawal, 2022). In the literature, it has been argued that ML learning may be "premature" for early age groups, particularly in terms of cognitive/abstraction load and data-ethical complexity; this view is based on assessments that rule-based traditional programming-centered approaches increase the need for abstraction at a young age and make pedagogical adaptation difficult. In contrast, recent studies show that ML offers meaningful opportunities for teaching higher-order cognitive skills such as CT at the early and middle school levels: design and implementation evidence indicates that in-class ML tasks directly activate CT components such as analysis, inference, evaluation, and explanation (Irgens et al., 2022; Arndt et al., 2025), and significant increases in CT subdimensions and related scores have been observed in high school-based AI/ML curricula (Liu et al., 2024). Moreover, in game-based tasks in early childhood contexts, CT processes have been shown to naturally drive children toward CT indicators such as evidence weighing, justification, and explanation (Granone et al., 2023). More broadly, current findings emphasize that the use of generative AI interactively enhances deep learning with CT, meaning that technological gains are grounded in students' CT skills (Zhao et al., 2025)

To address this gap, the current study investigates how hands-on ML-based activities are associated with students' CT performance and engagement in CT processes in secondary school students. Through an explanatory case study design that includes student activity performance measures, screen recordings, and qualitative interviews, the study investigates whether engagement with ML concepts supports reasoning, reflection, and evidence-based judgment. The findings aim to inform both pedagogical practices and curriculum design in AI-integrated learning environments.

Literature Review

1. Critical Thinking in Educational Contexts

CT is widely accepted as a fundamental 21st century competency to enable students to analyze arguments, evaluate evidence, and make sound decisions. Elder and Paul (2020) define CT as the

disciplined process of actively conceptualizing, applying, analyzing, synthesizing, and evaluating information. On the other hand, Halpern (2014) defines CT as the capacity to make inferences, evaluate alternatives, and systematically solve problems.
Multiple frameworks have emerged to structure CT development. Elder and Paul (2020) propose seven key dimensions: identifying goals, formulating questions, selecting relevant information, recognizing core concepts, reasoning aligned with objectives, offering practical insights, and drawing inferences. These dimensions serve as a foundational lens for evaluating how various instructional interventions, especially technology-mediated ones, support CT development in learners.

CT is not a single skill; it is a set of interconnected sub-skills and tendencies. Therefore, different categorizations of CT's "components" have been developed in the literature, and it is critically important for researchers reporting their work to clearly specify the framework they rely on for comparability. The most commonly referenced APA Delphi consensus defines CT through six core cognitive skills: interpretation, analysis, evaluation, inference, explanation, and self-regulation; these skills are considered alongside dispositions such as open-mindedness, fairness, reason-seeking tendency, flexibility, and desire for well-informed judgment (Facione, 1990, 2000). The Watson–Glaser approach presents a measurement-focused quintet: inference, assumption recognition, deduction, interpretation, and argument evaluation; in many experimental studies, "CT subtests" are scored based on this quintet (Watson & Glaser, 1994). The Halpern framework addresses CT as purposeful, reasoned, and goal-directed thinking; problem solving, decision making, probability/statistical reasoning, argument analysis, and logical inference, and emphasizes that these should be supported by metacognitive monitoring (consciously deciding what, when, and how to apply them) (Halpern, 1998; 2001). Another frequently used schema in educational practice is Bloom's higher-order cognitive levels (analysis–synthesis–evaluation), but its conceptual clarity is not as high as Delphi/Watson-Glaser (Bloom, 1956; Ennis, 1985). In recent years, frameworks such as PACIER (Problem solving, Analysis, Creativity, Interpretation, Evaluation, Reasoning) have proposed a six-faceted clustering that aligns well with the Delphi core by making creativity visible as a separate facet and is suitable for educational design.

Two examples are noteworthy in terms of how this diversity is reflected in current studies. Bellaera et al., (2021) systematically rated the CT skills that instructors in higher education contexts consider important and showed that sub-skills closely aligned with the APA/Delphi core, such as analysis, evaluation, and interpretation, are the most valued. This study treats CT as a set of independent sub-skills and bases its categorization on the Delphi tradition. Orhan and Çeviker Ay (2023) meanwhile, measured CT skills and CT tendencies together while comparing three teaching approaches: general, embedded, and blended; thus, aligning with the APA/Delphi line based on the "skill and disposition" duality and reporting the effect of the pedagogical approach on both performance (skills) and dispositions in a differentiated manner. These findings are consistent with the modern consensus that CT is not only about "what we can do" (cognitive processes) but also "how inclined we are to do it" (dispositions) (Facione, 2000; Paul & Elder, 2006).

In summary, the subdimensions of CT are conceptualized in different but largely overlapping clusters, such as (i) six Delphi-based skills, (ii) five Watson–Glaser-based subtests, (iii) Halpern's strategy and metacognition focus, (iv) Bloom's higher-order cognition, and (v) PACIER's six facets. Clearly stating which categorization is adopted in research reports makes the operationalization of variables transparent, highlights the emphasis on criteria/standards, and strengthens the synthesis across studies (Ennis, 1985; Bailin, 2002; Case, 2005). In this study, we also accept the Delphi-based framework, which addresses cognitive sub-skills (analysis, evaluation, inference, etc.) and dispositions together, as the main axis; when necessary, we refer to the Watson–Glaser and Halpern classifications when the measurement of sub-dimensions and alignment with instructional design is required. While computational thinking (CTh) is often cited alongside CT in technology-focused curricula, the two address distinct cognitive processes. CTh emphasizes abstraction, decomposition, and algorithmic logic (Wing, 2006), whereas CT focuses on metacognitive awareness, analytical reasoning, and evaluative judgment (Kestel, 2022). However, many educational studies conflate these concepts or overemphasize CTh due to its closer alignment with coding and programming activities (Sun et al., 2022; Kules, 2016).

1.1. Developing CT: Approaches and Evidence

CT is not a single course content but a multi-component competency that must be embedded in instructional design. Four main approaches stand out in the literature: general (stand-alone) education, immersion (implicit within the subject course), infusion (explicit and systematic CT instruction alongside subject content), and hybrid (general and infusion). Extensive meta-analyses show that CT yields meaningful gains when taught purposefully and explicitly; specifically, infusion and blended models are more effective than the immersion approach (Abrami et al., 2008). A meta-analysis conducted in the Turkish context similarly distinguishes between skill-based (general), content-based (immersion), and blended approaches, reporting that the blended/immersion line achieves higher effects on CT skills. This study shows that gains become apparent when CT sub-skills are aligned with open education goals. (Çeviker Ay & Orhan, 2020)

In terms of program development and assessment, the ACER CT Framework operationalizes teaching around three process strands: Constructing knowledge, evaluating ideas, making decisions. This framework provides a strong foundation for both immersion designs and blended models (e.g., task analyses, criterion-based rubrics, explicit reasoning evidence) because it links instructional objectives to sub-skills and ties them to observable performance indicators. (Heard et al., 2025) Current curriculum reviews also show that it is both possible and necessary to align subject-specific learning outcomes with CT sub-skills such as explanation, problem solving, data analysis, and inference; this finding supports an immersion/blended approach with explicit objectives rather than immersion. (Jamil et al., 2024)

In practical terms, the literature indicates that (i) teaching CT with explicit objectives, (ii) aligning subject-specific tasks with CT sub-skills (immersion), (iii) providing in-service training for teachers on CT pedagogy and assessment criteria, and (iv) monitoring performance using operationalizable frameworks such as ACER form a coherent strategy set for fostering CT. Accordingly, clearly specifying both the instructional approach (general, immersion, infusion, blended/hybrid) and the CT sub-dimension scheme in empirical studies enhances the comparability of findings and the validity of interpretations (Abrami et al., 2015).

2- ML in Educational Contexts

The integration of ML into educational contexts has followed two main trajectories: (1) the use of ML as an analytic tool to predict student performance, behavior, or needs; and (2) the use of ML as a pedagogical tool to support cognitive development, including critical and computational thinking. The latter, though emerging, remains significantly underexplored, especially at the K–12 level. Numerous studies have employed ML to enhance student profiling, predict academic outcomes, and support early intervention systems. For instance, Shrigoud and Agrawal (2022) and Kong and Shen (2024) applied ML algorithms to predict student performance and cognitive engagement by analyzing behavioral and demographic data. Similarly, Ge et al. (2024) utilized ML for early warning systems, while Balaji et al. (2021) emphasized its capacity to identify at-risk students. These studies, however, primarily focus on data analytics rather than student cognition or reasoning processes. Recent research has begun to investigate ML as a learning subject in itself—particularly in ways that engage students in data-driven reasoning and reflective learning.

Computational thinking (CTh) and critical thinking (CT) are two modes of thinking that do not fully overlap but are meaningfully related. CTh organizes the solution-generation process through abstraction, decomposition, algorithmic design, and iteration, while CT ensures evidence- and criterion-based justification of these solutions and the questioning of assumptions. Therefore, CTh-based tasks (e.g., data decomposition, algorithm selection, debugging) often simultaneously trigger CT sub-skills such as analysis, evaluation, and inference; conversely, high-quality CT outputs nurture the systematic and organized problem-solving habits associated with CTh. The literature indicates that CTh applications, especially in educational contexts, are frequently reported alongside CT indicators and that the two constructs contextually reinforce each other, although one does not replace the other: CTh emphasizes "how to do it," while CT emphasizes "why this way and according to what criteria it is better" (Wing, 2006; Cansu & Cansu, 2019; Granone et al., 2023).

Çakıroğlu and Selçuk's (2024) exploratory mixed-methods study with 20 fifth-grade students (aged 11–12) examined the development of CTh sub-dimensions (abstraction, decomposition, pattern recognition, algorithm design) through 60 hours of ML activities (12 visual, text, audio, and numerical activities) on the Machine Learning for Kids platform. Using a scoring rubric informed by semi-structured interviews and screen recordings, the authors reported increases to the "sufficient–advanced" level across all activity types, particularly in algorithm design and pattern recognition, with relatively lower gains in decomposition and the lowest averages in number-based activities (and the highest in visual activities). The study argues that in-class ML activities make CTh skills visible through concrete behavioral patterns and support a transition from traditional CTh teaching to data-driven reasoning, while at the same time providing a foundation for CT processes by strengthening data-driven reasoning, error analysis, and reasoned decision-making in elementary school students.

Complementary findings in other contexts suggest that teaching ML to younger learners is a viable pedagogical approach. Su and Zhong (2022) and Banadaki (2020) report that early childhood and secondary learners can engage meaningfully with ML concepts when supported by visual platforms and real-world examples. User-friendly environments such as Machine Learning for Kids and ScratchThAI (Katchapakirin et al., 2022) simplify the technical aspects of ML, allowing students to concentrate on developing key cognitive skills, including recognizing patterns, experimenting with models, and iteratively improving their designs. Karalekas et al. (2023) examined the use of robotics-based activities to introduce fundamental AI ideas to younger students, while Sanusi et al. (2023) presented evidence that engaging with ML tasks can foster CTh skills in elementary and middle school learners. Some studies extend ML use to examine cognitive and emotional patterns in children; for example, Rani et al. (2021) used ML to model behavioral traits, highlighting ML's diagnostic potential rather than its instructional value. While informative, such applications do not address how engaging in ML tasks can develop CT in learners. Despite promising developments, only a few studies, such as Lamb et al. (2021) and Çakıroğlu and Selçuk (2024), have directly explored the impact of ML activities on CT. Most educational ML studies continue to emphasize computational skill acquisition rather than reflective, evaluative reasoning, and the middle school age group—a cognitively crucial developmental period—remains underrepresented in this literature. A summary of key studies in this domain is presented in Table 1.

Table 1. Summary of Selected Studies on ML in Education

Author(s) and Year	**Study Focus**	**Target Age Group**
Khan et al. (2018)	Diagnostic system for learning disabilities	Children
Marques et al. (2020)	ML in early childhood computing education	Early Childhood
Rani et al. (2021)	ML model for emotion and behavior analysis	Children
Lamb et al. (2021)	ML algorithms and their impact on CT in science	Secondary School Students
Katchapakirin et al. (2022)	Speech-based ML learning support (ScratchThAI)	K–12 Students
Pacheco et al. (2023)	ML activities and computational thinking development	Primary and Middle School Students
Sanusi et al. (2023)	ML activities	Middle School Students
Karalekas et al. (2023)	ML and robotics for teaching AI concepts	K–12 Students
Banadaki (2020)	ML-intensive STEM projects to improve computing and analytical skills	University Students
Jin & Cutumisu (2023)	Predicting pre-service teachers' CTh using ML	University Students
Kong & Shen (2024)	Predicting CTh understanding via ML based on learner traits	Secondary and High School Students
Çakiroğlu & Selçuk (2024)	Hands-on ML activities to enhance computational thinking	Secondary School Students

As summarized in Table 1, research on ML in K–12 education is expanding but remains fragmented. Most studies focus on technical outcomes (e.g., classification accuracy, CTh skills), while only a handful explore ML's potential to engage students in reflective and evaluative reasoning. This study seeks to address that gap by investigating how middle school students' participation in hands-on ML tasks supports CT performance and engagement in CT processes, particularly reasoning, inference, and judgment application across ML stages (training, testing, application).

Across K–12 settings, studies position ML in education along two trajectories: an analytic tool for prediction, profiling, and early-warning systems, and a pedagogical tool that engages learners in data-driven reasoning and iterative problem solving. The first trajectory is mature but largely system-facing, emphasizing technical metrics over students' reasoning processes, whereas the second is emerging and remains conceptually fragmented. Many classroom studies using visual or block-based ML environments (e.g., Machine Learning for Kids, ScratchThAI) report gains in computational routines and, in a smaller subset, CT indicators such as evaluation, inference, and explanation. However, few papers analyze which ML phase (training, testing, application) elicits which CT dimension, or how learners' CT profiles cluster across multiple ML tasks. This pattern reveals a clear gap: we lack studies that (a) map CT dimensions to specific ML phases within classroom tasks, (b) measure CT performance directly (beyond self-report), and (c) differentiate learner profiles to inform targeted scaffolding. Accordingly, the present study uses a multi-method design (worksheets, screen recordings, interviews) and a two-dimensional CT measure, coupled with cluster analysis, to capture phase-sensitive CT performance and identify instructional levers that can guide ML-integrated learning environments.

Research Questions

1. What is the impact of ML activities on the development of CT skills among middle school students?
2. How do students' experiences with different stages of the ML process (e.g., training, testing, application) reflect structural components of CT?
3. What critical-thinking performance profiles emerge across the ML task sequence, and how do these profiles differ in observed CT indicators across the training, testing, and application phases?

Method

This study employed an explanatory case study design, combining both quantitative and qualitative approaches to investigate the impact of ML activities on middle school students' CT skills. The quantitative data were used to assess the level students' CT performance across rubric-based indicators, while qualitative data provided a deeper understanding of the learning processes and student experiences associated with the ML activities. Perspective taking was treated as part of reasoning and justification within the adopted CT framework (Elder & Paul, 2020).

Participant

The participants were 15 voluntary middle school students (7 girls and 8 boys) from 5th and 6th grades in a public school in Türkiye, enrolled in an Information and Communication Technologies (ICT) course. Participants were selected through purposeful sampling, ensuring that all students had prior basic training in block-based programming. Activities were conducted in a computer lab, and each student was assigned an individual workstation. The students engaged in weekly ML activities as part of their curriculum, designed to align with their developmental and cognitive readiness.

The study was conducted as part of an 8-week "Learning Machine Learning" course designed as an extracurricular learning opportunity for middle school students. The course was a continuation of an 8-week 'Block-Based Coding' program previously conducted with the same group of students. Students who successfully completed this first program, which served as a prerequisite, were required to score at least 70 points on a standardized proficiency test in order to proceed to the second stage. This ensured that participants had a similar level of basic proficiency in block-based programming. The applications were conducted in the computer lab of a state middle school, once a week (≈90 minutes), with the guidance of an experienced instructor and a research assistant; the class size was kept constant at n=15.

Procedure

The ML activities were conducted using the ML for Kids platform over five weeks, with four instructional hours per week. The research process is presented in detail in Figure 1.

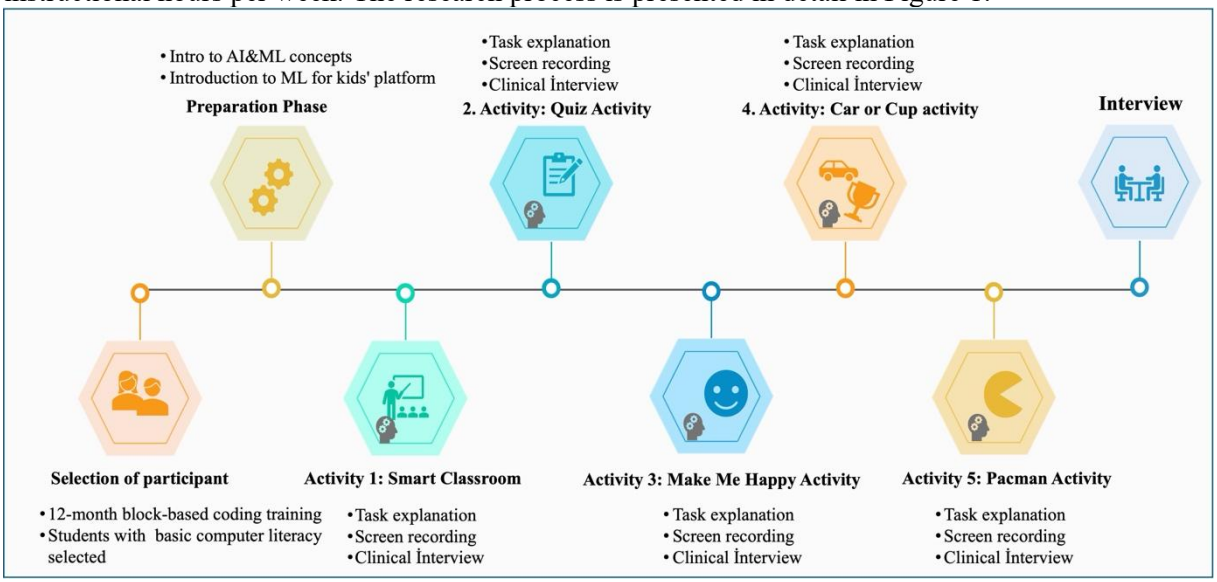


Figure 1.The research process

A series of five instructional activities were implemented to support the development of students' CT skills through structured interaction with ML concepts. Activities were selected and structured based on students' cognitive readiness, with guidance from domain experts.

The selection and structuring process was carried out under the guidance of three field experts. Two of the experts were academics with doctoral degrees in Computer and Instructional Technology Education, whose work focuses on the use of block-based programming and ML in educational contexts. The third expert was the Information Technologies teacher at the participating school, who also taught the ICT course to the students in the study. Based on a literature review and existing online platforms, the research team initially created a pool of 20 activities. The two PhD-level experts evaluated these activities in terms of alignment with the targeted learning outcomes, cognitive difficulty level, compatibility with classroom time, and conceptual progression. Activities deemed unsuitable in terms of duration and content were eliminated, reducing the pool to nine. Subsequently, the Information Technologies teacher, who was familiar with the students' prior academic performance and areas of interest, re-examined these nine activities with respect to the student profile and selected five activities that were most suitable for the classroom dynamics. Following expert feedback, minor adjustments were made to the guidelines of the selected activities to ensure that tasks were clearly defined and structured to support step-by-step progression.

In total, the activity set consisted of five tasks: three text-based, one image-based, and one number-based. The activities targeted various ML modalities (e.g., classification, supervised learning, pattern recognition) and were aligned with eight core dimensions of CT. A summary of these activities, including their instructional type, structural features, and pedagogical objectives, is provided in Table 2.

Table 2. Activities, Including Their Instructional Type, Structural Features, Pedagogical Objectives

Activity Name	**Activity Type**	**Description**	**Activity Visual**	**Learning Objective**
Smart Classroom	Text-based	Two components are defined, and descriptions are based on these		Students design an intelligent assistant in Scratch to control virtual devices. Key guiding questions include: - How can computers be trained to understand the purpose behind writing?

		components.		- How do virtual assistants like Siri or Alexa work?
Quiz Bot	Text-based	A single component is defined and used as the basis for classification.		Students train a machine to answer quiz questions. Key guiding questions include: - How can computers be trained to find the correct answers to questions? - What is question answering as a field of ML?
Make Me Happy	Text-based	Two components are defined, and descriptions are based on sentiment analysis.		Students train a machine to recognize positive and negative expressions. Key guiding questions include: - How can computers be trained to detect emotional tone? - How does supervised learning handle unexpected inputs?
Car or Cup?	Image-based	Two components are visually defined and used for classification tasks.		Students train machines to distinguish between car and cup images. Key guiding questions include: - How can computers be trained to recognize images? - What is the role of diversity in ML training datasets?
Pac-Man Learner	Number-based	Requires understanding of two-dimensional coordinates; the more the machine plays, the better it learns.		Students teach a machine to play games. Key guiding questions include: - How can machines be taught to play games? - How is decision tree learning used in training machines to play?

During each session, students received introductory explanations from the researcher, engaged with the ML tasks, and responded to reflection prompts. All sessions were recorded, and post-activity interviews were conducted individually. These activities required students to perform the core steps of supervised ML: labeling data, training models, testing predictions, and refining outputs. Each activity was aligned with specific CT dimensions, including goal setting, identifying key information, reasoning, and drawing conclusions. Following expert consultation and a review of relevant literature, five ML activities were carefully selected and implemented over a five-week period, with students engaging in the tasks for four hours each week. The design of these activities considered students' age and readiness levels, ensuring an appropriate cognitive load. Specifically, the intervention included three text-based,

one image-based, and one number-based activity, each aligned with key dimensions of CT. These tasks were structured to prompt cognitive processes such as goal definition, question generation, information selection, reasoning, application of practical ideas, and inference-making. Through this structured approach, the study aimed to foster the systematic development of students' CT abilities. Visual examples from the implementation of these activities are provided in Figure 2.

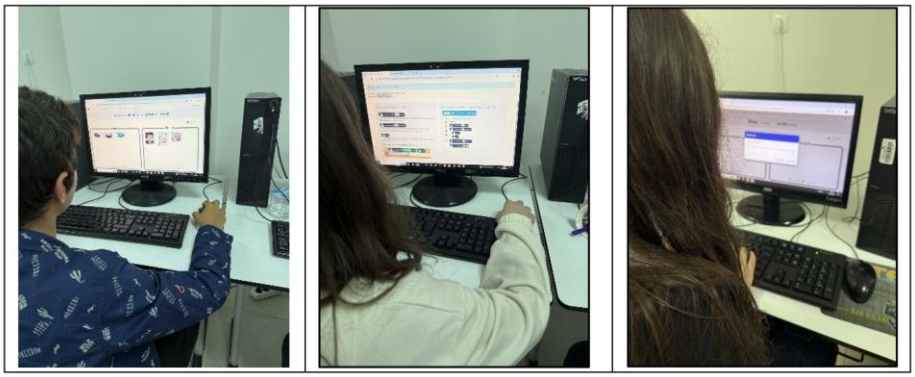


Figure 2. Visual Examples from the Implementation of Activities

The overall research process, including session durations and activity breakdowns, is presented in Figure 1.

Data Collection Tools

The data collection instruments utilized to examine the contributions of ML activities to students' CT performance and engagement in CT processes, as well as the corresponding data analysis techniques employed for each instrument, are summarized in Table 3.

Table 3. Data instruments, are summarized

Data collection tool	**Data analysis**
Screen Recording (to determine how students use "Machine Learning for Kids" platform ML activities output (to reveal the development of students' ML and **CT**)	ML and CT performance rubric
Clinical interviews	ML and CT performance rubric
Interviews (to explain the role of ML activities in CT performance	Content analysis
Rubric scores (ML–CT rubric) → Descriptive statistics / graphs	Rubric scores (ML–CT rubric) → Descriptive statistics / graphs
Rubric total/phase scores → Cluster analysis (RQ3 profiling)	Rubric total/phase scores → Cluster analysis (RQ3 profiling)

Screen Recording

Students' behavioral patterns and verbal expressions during their engagement with the ML activities on the Machine Learning for Kids platform served as a rich source of data for understanding how they engaged in CT processes and how their CT performance evolved during tasks. To capture these processes in situ, all student interactions were systematically recorded using screen capture software during each session, which lasted approximately 30–35 minutes. These screen recordings enabled the researchers to trace the specific steps students followed in the ML workflow (e.g., training, testing, application), and to analyze how students articulated reasoning, evaluated outputs, and revised their approaches, key indicators of CT. Screen recordings were later transcribed and coded for qualitative

content analysis. The timing and sequence of actions were synchronized with rubric-based CT dimensions.

Clinical interviews

Clinical interviews were conducted to gain deeper insights into how students engaged with each stage of the ML process and to identify the specific phases they completed successfully or found challenging.

- *What is my primary goal in this task? (Purpose of thinking)*
- *What key questions am I trying to answer? (Defining the central problem*
- *What information do I need to answer these questions?" (Identifying relevant data)*
- *What are the core concepts in this problem? (Clarifying fundamental ideas)*
- *What assumptions underlie my reasoning? (Recognizing implicit beliefs)*
- *What is my perspective on this issue? (Considering point of view)*
- *What conclusions am I drawing? (Drawing inferences)*
- *If my reasoning is valid, what are the possible implications? (Exploring consequences)*

These targeted prompts allowed for a deeper examination of students' logical structuring, reasoning strategies, and cognitive self-awareness, key indicators of higher-order thinking and critical reasoning development.

Interviews

To gain deeper insights into how students perceived the ML activities and how these activities contributed to their CT development, semi-structured interviews were conducted individually with each participant at the end of every activity. The interview protocol, developed specifically for this study by the first and second authors, consisted of 10 open-ended questions aimed at eliciting students' reflections, cognitive strategies, and evaluative judgments. In developing the protocol, relevant literature on ML education and CT was reviewed, and a preliminary version of the questions was examined by two experts in Computer and Instructional Technologies to ensure content relevance and clarity. The protocol focused on four main dimensions: (a) students' perceptions of the activities, (b) perceived difficulties and cognitive challenges, (c) perceived contribution to CT, and (d) suggestions for improving the instructional design. Example prompts included: *"Was there any new concept you learned during the activity? How do you think you can apply it in real life?"*, *"Which stage of the ML activity did you find most challenging, and why?"*, *"Do you believe these activities contributed to your* CT *skills? If so, how?"* and *"If you were to explain this activity to another student, which parts would you emphasize?"*. In addition to assessing cognitive engagement, the interviews also sought students' recommendations for improving the activities; accordingly, reflective prompts such as *"Could different versions of this activity be developed?"* and *"If you had the chance to redo this activity, what changes would you make?"* were included. Each interview was conducted in a quiet environment, lasted approximately 45 minutes, and was audio-recorded with prior consent. All recordings were transcribed verbatim for qualitative analysis, and the semi-structured format allowed for follow-up questions based on students' responses, thereby enhancing the depth and trustworthiness of the data.

Data Analyses

ML and CT performance rubric

The assessment rubric developed for this study was designed to establish a structured relationship between the key phases of the ML process, namely training, learning, testing, and application, and various dimensions of CT. The CT dimensions integrated into the rubric include: goal setting, formulating guiding questions, selecting relevant information and core concepts, reasoning, articulating application-based judgments, and making inferences. Each dimension was operationalized through performance indicators that reflect the expected competencies students should demonstrate during each

ML phase. Drawing upon relevant literature and expert consultation, these indicators were categorized into three distinct performance levels: Insufficient (0), Needs Improvement (1), and Proficient (2). The content validity and reliability of the rubric were ensured through iterative reviews and feedback from field experts specializing in educational measurement and cognitive skills development. The finalized version of the rubric, which guided both the qualitative and quantitative analysis of student responses, is presented in Table 4.

Table 4. Finalized Version of the Rubric

CT Dimension	Training (0-1-2)	Learning & Testing (0-1-2)	Application (0-1-2)
Goal Identification	☐☐☐	☐☐☐	☐☐☐
Question Formulation	☐☐☐	☐☐☐	☐☐☐
Identification of Relevant Information	☐☐☐	☐☐☐	☐☐☐
Identification of Core Concepts	☐☐☐	☐☐☐	☐☐☐
Logical Reasoning	☐☐☐	☐☐☐	☐☐☐
Application-Oriented Judgment	☐☐☐	☐☐☐	☐☐☐
Drawing Inferences	☐☐☐	☐☐☐	☐☐☐
Drawing Conclusions	☐☐☐	☐☐☐	☐☐☐

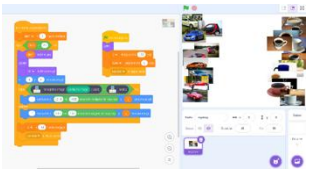
The specific indicators that define student performance at each rubric level across all CT dimensions and ML phases are included in the appendix for reference and transparency. Interrater reliability was computed on 30% of student–activity scoring units (n = 23) independently scored by both raters. Reliability was evaluated using quadratic-weighted Cohen's κ for ordinal ratings, and disagreements were resolved through discussion to reach consensus for the final dataset used in subsequent analyses.

Illustrative Data Analysis

Summary of student responses and their alignment with ML phases and CT components. Clinical interviews explored students' understanding of ML processes through guided prompts. Responses were analyzed in relation to their corresponding CT indicators and supported by behavioral evidence from screen recordings Table 5.

Table 5. Analyzed in relation to their corresponding CT indicators and supported by behavioral evidence from screen recordings

Interview Prompt	ML Process Stage	CT Component	Student Response (Translated)	Screenshot	Analysis Summary
What should a computer pay attention to when distinguishing between images of cups and cars?	Training	Question Formulation Concept Clarification	*"The more varied photos we train the computer with, the better it can tell the difference. More examples*		The student understood the logic of the training phase in ML and added multiple visual samples to reinforce model differentiation.

			help it learn."		
What block codes were used by the computer to distinguish between a cup and a car?	Application	Drawing Inferences Reasoning	*"I used conditional structures in this step to help separate cups and cars."*		The student applied conditional logic effectively during the application phase, indicating an understanding of programming principles and reasoning processes.

This table illustrates how students' verbal responses during clinical interviews were mapped to specific ML steps (e.g., training, application) and associated with key dimensions of CT (e.g., inference, reasoning, concept clarification). Student reasoning was triangulated with behavioral data from screen recordings and evaluated using researcher-developed rubric. To enhance the reliability of the rubrics, performance data was first recorded synchronously by the research assistant present in the classroom during the application. Following the application, screen recordings of all sessions were independently re-examined by two researchers and compared with the research assistant's rubric scores. This comparison demonstrated a high level of consistency between the research assistant and the researchers overall.

Cluster Analysis

To classify students based on their CT performance during the ML activities, a K-means clustering analysis was conducted using the K-means algorithm in the RStudio environment. The analysis identified three distinct clusters, corresponding to low (Cluster 1), medium (Cluster 2), and high (Cluster 3) levels of CT engagement.

Qualitative Data Analysis and Trustworthiness Procedures

The semi-structured interviews were analyzed using a directed content analysis approach. A preliminary coding framework was developed deductively, based on the predefined ML process stages (e.g., data collection, training, testing, application) and the CT components targeted in this study (e.g., question formulation, concept clarification, inference, reasoning, evaluation). The unit of analysis was each meaningful statement in which students described how they understood, justified, or evaluated the activity. Two researchers first independently coded a subset of the transcripts to refine the codebook and ensure a common understanding of the categories; discrepancies were discussed until consensus was reached. The remaining transcripts were then coded using the agreed-upon framework, and coded segments were mapped onto the corresponding ML process stages and CT components, as illustrated in Appendix 2.

To enhance the trustworthiness of the qualitative findings, several strategies were employed. Data triangulation was achieved by integrating three complementary sources: students' interview responses, behavioral evidence from screen recordings, and performance scores from the ML–CT rubric, allowing us to examine whether students' reported reasoning aligned with their on-screen actions. Researcher triangulation was also implemented: during the activities, the performance rubric was completed in real time by a research assistant, and after the intervention, all screen recordings were independently reviewed by two researchers, who compared their judgments with the assistant's scores and resolved any discrepancies through discussion. In addition, an audit trail documenting the evolution of the codebook and key analytic decisions was maintained. Collectively, these procedures supported the credibility, dependability, and confirmability of the qualitative results.

Finding

This section presents the findings obtained in line with the three research questions guiding this study. For the first two research questions, data were collected through systematic observations of 15 middle school students during ML activities, including their interactions and on-screen behaviors. These observations were evaluated using a structured rubric, quantitatively coded, and visualized through graphs to reveal patterns related to students' CT performance and engagement in CT processes. In addressing the third research question, a cluster analysis was conducted to categorize students into low, medium, and high CT profiles. Subsequently, qualitative data were gathered through clinical interviews and semi-structured interviews with representative students from each cluster. These data were analyzed to explore how students' CT levels influenced their performance across different stages of the ML tasks. The findings are organized and reported under subheadings corresponding to each research question.

Impact Of Each ML Activity on Students' CT **Performance**

The impact of each ML activity on students' CT performance was examined in detail to understand how different task designs influenced specific cognitive dimensions. Using a rubric aligned with eight key components of CT, total scores were calculated for each activity and aggregated across all participants. Figure 3 illustrates the extent to which each of the five ML activities supported the development of these skills.

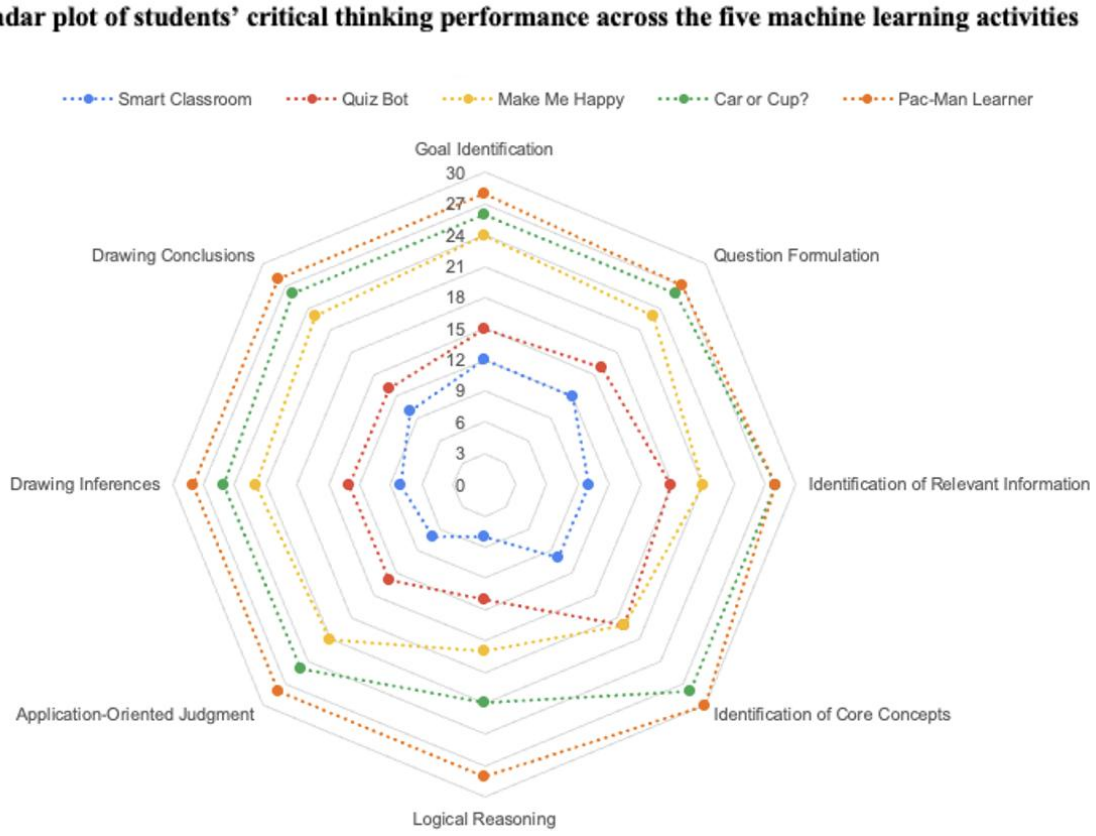


Figure 3. Comparison of students' critical thinking scores across five ML activities by dimension.

The radar plot displays the aggregated CT scores that students achieved in each of the five activities across eight components: Goal Identification, Question Formulation, Identification of Relevant Information, Identification of Core Concepts, Logical Reasoning, Application-Oriented Judgment, Drawing Inferences, and Drawing Conclusions. Each spoke represents one CT component, and the concentric circles indicate the possible score range (0–30 points). The coloured polygons correspond to the five activities; polygons that extend further from the centre indicate higher performance on that component for the corresponding activity.

In this scoring system, each CT component had a maximum of 30 points per activity. For example, in the Goal Identification dimension, students obtained a total of 12 points in Activity 1, whereas in

Activity 5 they reached 28 points, approaching the maximum possible score. Thus, the relatively smaller, inner polygon for Activity 1 reflects a more limited demonstration of CT across components, while the larger outer polygon for Activity 5 indicates more advanced and consistent engagement in identifying goals, formulating questions, selecting relevant information, and drawing logically grounded inferences and conclusions. Overall, the visual expansion of the polygons from Activity 1 to Activity 5 illustrates a progressive strengthening of students' CT performance as they engaged with increasingly complex ML tasks.

As illustrated in Figure 3, there was a progressive observed change in students' CT performance across the five ML activities. Activity 1 showed the lowest scores across all eight dimensions, particularly in Logical Reasoning and Application-Oriented Judgment, indicating students' initial difficulty in engaging with the analytical and metacognitive aspects of the tasks. Activity 2 demonstrated modest gains, especially in Goal Identification and Question Formulation, suggesting that students began to orient their thinking more effectively. Notable change were observed in Activity 3, with consistent increases across nearly all dimensions, particularly in Identification of Core Concepts and Identification of Relevant Information. Activity 4 introduced more complex problem-solving, and while it maintained high scores, there was a slight plateau in Logical Reasoning and Application-Oriented Judgment. In contrast, Activity 5 achieved the highest scores in all eight dimensions, indicating a well-rounded and integrated application of CT. The dimensions of Drawing Inferences and Drawing Conclusions peaked notably in this final activity, reflecting students' increased capacity to synthesise information and justify their decisions based on prior experience.

Total Scores by CT **Dimensions Across the Phases of ML Activities**

To evaluate how students' CT skills manifested across different phases of the ML activities, total scores were analyzed for each of the eight dimensions of CT clarifying purpose, question formulation, information selection, concept identification, purposeful reasoning, reflection for practical application, and drawing conclusions. These scores were examined separately for the training, learning/testing, and application stages, as illustrated in the radar graph below (Figure 4).

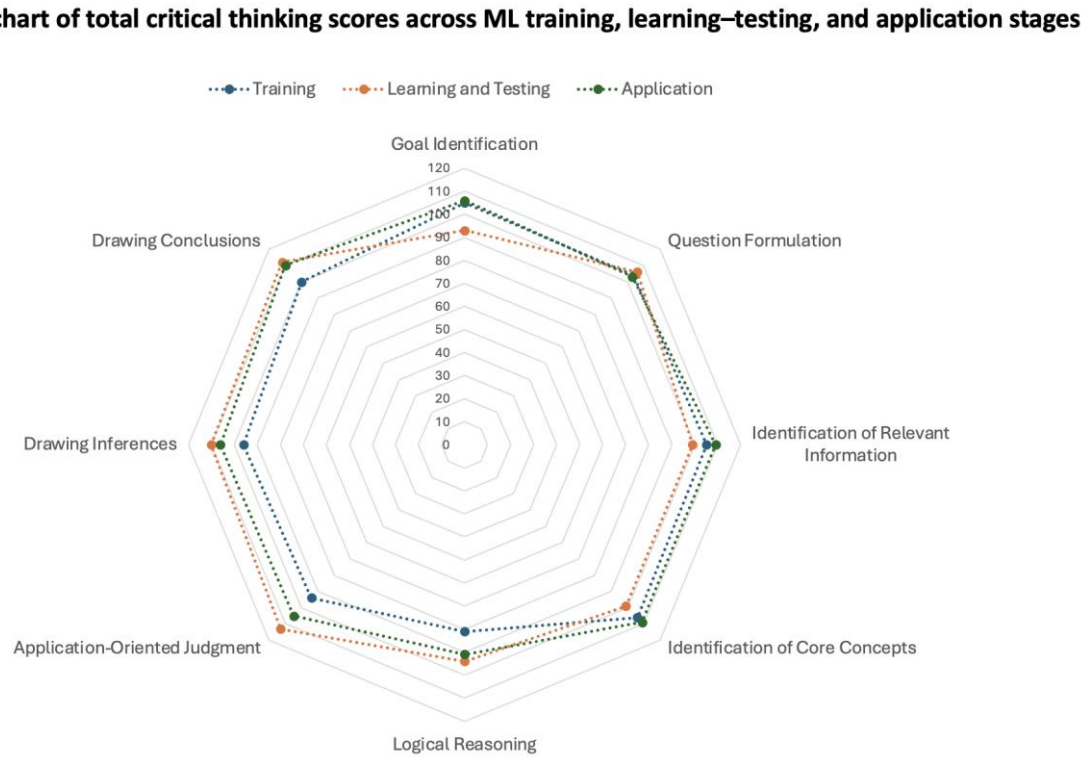


Figure 4. Total Scores by *CT* Dimensions Across the Training, Testing, and Application Phases of ML Activities

Figure 4 presents a radar chart of total CT scores for the three ML stages—Training, Learning & Testing, and Application—across eight components: Goal Identification, Question Formulation, Identification of Relevant Information, Identification of Core Concepts, Logical Reasoning, Application-Oriented Judgment, Drawing Inferences, and Drawing Conclusions. Each subtask comprised these eight components, which were scored between 0 and 2 points for 15 students across five activities (75 scoring instances), yielding a possible total of 0–150 points per component. Each spoke represents one CT component, the concentric circles indicate this 0–150 scale, and the coloured lines correspond to the three ML stages; lines further from the centre denote higher performance.

As shown in Figure 4, CT performance differed systematically across the three stages. The Application stage consistently produced the highest scores on all dimensions, followed by Learning & Testing, while the Training stage yielded the lowest scores. Among the eight CT components, Goal Identification, Question Formulation, and Drawing Conclusions emerged as the strongest, with all three stages exceeding 100 points in the Application phase, suggesting that students were able to clearly define goals, pose meaningful questions, and derive coherent outcomes in authentic ML tasks. By contrast, Logical Reasoning and Application-Oriented Judgment were weakest during the Training phase, remaining below 75 points, yet they observed change markedly in the Application stage, indicating growth in metacognitive and evaluative skills over time. This pattern may reflect increased familiarity with the workflow and/or the fact that application tasks generate more observable evidence aligned with rubric indicators. Therefore, the stage differences should be interpreted as observed performance patterns rather than causal effects.

When students' scores from the ML activities were examined according to the core stages of the ML process—Training, Learning & Testing, and Application—distinct performance patterns emerged. Total scores for each stage were calculated using a rubric based on eight CT dimensions and then aggregated across all 15 participants. Figure 5 illustrates the distribution of these scores across the five ML activities, enabling a comparative analysis of how students performed at each stage of the process.

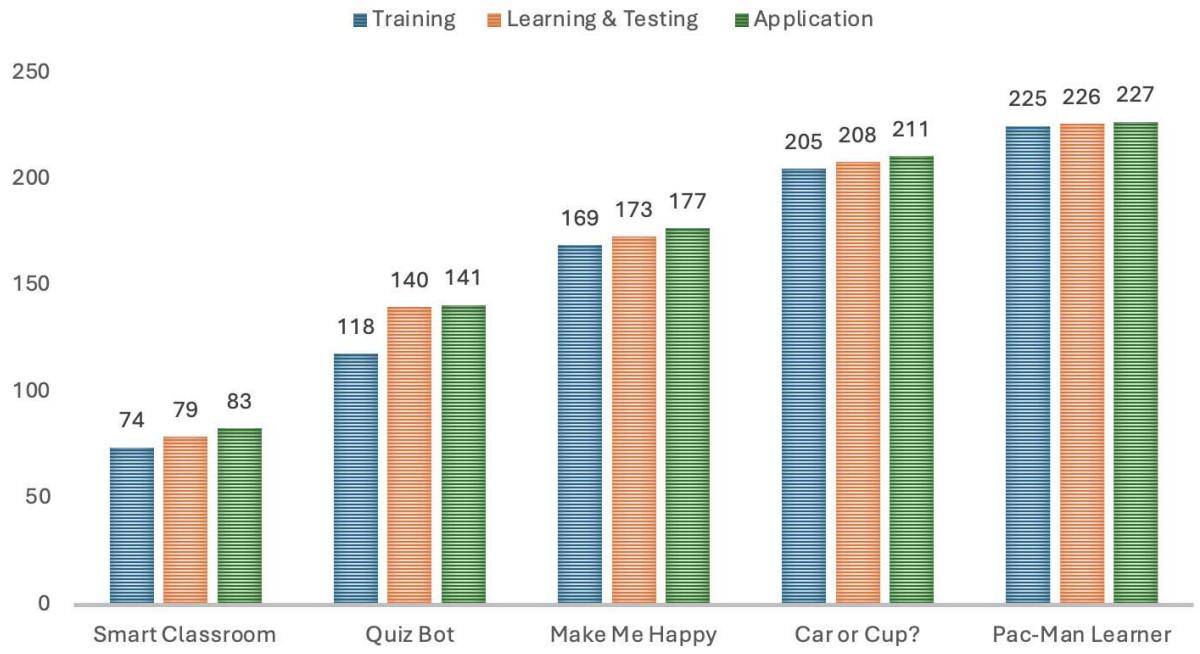


Figure 5. Total scores of students in each ML process stage (training, learning and testing, application) across five activities.

This bar chart presents the total CT scores achieved in each ML activity—Smart Classroom, Quiz Bot, Make Me Happy, Car or Cup, and Pac-Man Learner—for the three ML stages: Training, Learning & Testing, and Application. In this scoring system, each stage comprises eight CT components, each

scored between 0 and 2 points for 15 students, yielding a maximum possible score of 240 points per stage (2 × 8 × 15). Higher bars indicate stronger overall performance in the corresponding ML stage. As shown, total scores increase progressively from the first activity to the last, with the Application stage consistently displaying the highest performance, followed by Learning & Testing and Training. This pattern is consistent with improved performance over time; however, without a comparison group, it should be interpreted cautiously. Differences may also stem from task characteristics and the visibility of rubric evidence across stages.

The results reveal a consistent upward trend across all stages from Activity 1 to Activity 5. Training scores, which began at the lowest level in Activity 1, increased steadily and peaked in Activity 5. A similar pattern is observed for the Learning & Testing and Application stages, both of which show incremental gains across activities. In Activity 1, all three stages yielded relatively low scores, reflecting students' initial exposure to ML-based problem solving. From Activity 2 onward, performance observed change markedly, with the most pronounced increase occurring between Activities 2 and 3. By Activity 4, the total scores for each stage had converged, indicating more consistent performance across the ML process. In Activity 5, students achieved their highest scores in all three stages, with Application slightly surpassing Training and Learning & Testing, suggesting strong mastery in deploying ML models and navigating the full ML cycle

Cluster-Based Analysis of CT Performance

To classify students based on their CT performance during the ML activities, a K-means clustering analysis was conducted using the kmeans algorithm in the RStudio environment. The analysis identified three distinct clusters, corresponding to low (Cluster 1), medium (Cluster 2), and high (Cluster 3) levels of CT engagement.

- Cluster 1 (Low) included students S2, S5, S8, and S9.
- Cluster 2 (Medium) comprised students S4 and S7.
- Cluster 3 (High) contained students S1, S3, S6, S10, S11, S12, S13, S14, and S15.

These clusters were not only derived from quantitative rubric scores but were also qualitatively validated through the analysis of students' verbal and behavioral data collected during ML tasks. The resulting classification provided meaningful insight into the variations in students' engagement with CT dimensions such as reasoning, inference, reflection, and goal setting. This mixed-method triangulation enabled a deeper understanding of how different learners interacted with ML processes and how these interactions mapped onto their demonstrated CT competencies.

Based on the K-means clustering analysis, students' CT performance throughout the ML activities was categorized into three levels: low (Cluster 1)**,** medium (Cluster 2)**,** and high (Cluster 3)**.** The clusters were then interpreted in conjunction with qualitative data derived from students' task-related verbal reflections, allowing for a more comprehensive evaluation of how each group engaged with the CT dimensions. Because Cluster 2 contained only two students, the 'medium' profile should be viewed as exploratory. Cluster interpretations are presented as preliminary patterns rather than generalizable learner categories.

Cluster 3 – High Level

Students in this group (Ö1, Ö3, Ö6, Ö10, Ö11, Ö12, Ö13, Ö14, Ö15) demonstrated strong CTskills across all sub-dimensions, including clarifying purpose**,** question formulation**,** information selection**,** inference, and reasoning. For example, student Ö1 stated: *"While doing the training and testing parts, I realized that if the computer isn't trained properly, the activity won't work. So I went back, explained the activity again, retrained it, and then it worked."* This response indicates that Ö1 was able to identify the problem and generate a solution, thus effectively applying reasoning and achieving the drawing conclusions dimension. The student actively revised their approach based on observed outcomes, showcasing reflective and goal-oriented reasoning within the learning process. Similarly, student Ö6 shared: *"First, I asked myself how the machine would think and how it could respond to the question. Then I gave it examples to train it."* This indicates effective engagement

in question formulation and information selection within the training phase. Furthermore, Ö6 demonstrated awareness of how to align the instructional content with the machine's learning process. Both students Ö1 and Ö6 illustrated clear cognitive strategies by defining questions beforehand and presenting targeted examples during the training phase. Their responses affirm the potential of ML activities to foster higher-level CT competencies.

Cluster 2 – Medium Level

This cluster includes students Ö4 and Ö7, who demonstrated partial engagement with CTskills. Although they showed some awareness of task complexity, their ability to resolve problems or refine their strategies remained limited. Student Ö4 noted: *"Even though I uploaded pictures of a cup and a car and labeled them, the machine couldn't differentiate. Maybe the images weren't enough. I think I need to add more."* This indicates some application of the drawing conclusions dimension within the testing stage. However, the student did not implement a structured solution approach, resulting in an overall moderate performance. Likewise, student Ö7 stated: *"Before doing the activity, I thought about what makes me happy or sad and then uploaded those examples."* This demonstrates an attempt to engage in information selection; however, the limited number of examples resulted in classification errors by the model. Although both students were able to reflect on task challenges, they were not able to fully translate those reflections into effective strategies or corrections.

Cluster 1 – Low Level

Students in this cluster (Ö2, Ö5, Ö8, Ö9) struggled to meaningfully apply most of the CTdimensions during the ML activities. Student Ö2 reported: *"I had trouble while using the blocks, especially in the application part. I got confused and didn't understand how to do it."* This indicates an inability to engage in drawing conclusions during the implementation stage. Student Ö5 said: *"I considered what other possible outcomes could be, but I didn't test them during the training phase. I thought one option was enough."* This shows a lack of inference, as the student did not explore alternative pathways or evaluate the model's performance critically. Student Ö8 commented: *"I couldn't use the ML blocks. They confused me. I was supposed to use them with operators, but I forgot."* Similarly, Ö9 stated: *"I had difficulty with the Pac-Man activity. We had to train the computer by playing, but the coding part was hard. I really struggled to write the code."* Both students failed to apply CTin the application phase, especially in terms of reaching meaningful conclusions. In sum, the students in Cluster 1 expressed difficulties primarily during the application and inference stages. They often failed to develop strategic solutions and were limited to single-solution thinking. These challenges appeared to stem from a lack of understanding of the activity goals and limited coding competence, which hindered their ability to meaningfully engage in key CT dimensions.

Conclusion

The current study provides strong support for the proposition that integrating carefully designed ML activities into the middle school curriculum can significantly enhance students' engagement in CT abilities. Through goal-oriented reasoning, iterative problem solving, and reflective decision making, students engaged in cognitively challenging processes that closely aligned with established CT frameworks. The gradual observed change across the five ML activities demonstrates the pedagogical importance of sequencing ML tasks to support cumulative learning outcomes. The findings indicated that different phases of the ML process (training, testing, and practice) differentially tapped various dimensions of critical thinking. Specifically, the practice phase was the most effective in eliciting higher-order thinking, suggesting that the delivery of authentic, real-world tasks enhanced students' ability to synthesize information, draw inferences, and meaningfully evaluate results. Furthermore, the study's cluster analysis revealed significant variability in students' CT engagement, suggesting the need for differentiated instructional strategies. High-performing students exhibited metacognitive behaviors such as model revision and evidence-based reasoning, while low-performing students struggled to identify concepts and required more explicit support. Furthermore, the variety of ML activity types (text-, image,

and number-based) uniquely contributed to the development of CT subskills. In conclusion, this research highlights the potential of ML to be more than just a technical subject; it is clear that it can be a dynamic pedagogical tool that can develop core cognitive competencies in K-12 education. To maximize its impact, future applications should include structured thinking prompts, adaptive support mechanisms, and real-world applications that empower students to think critically, creatively, and ethically in AI-mediated environments.

Discussion

This study aimed to investigate how structured ML activities support the development of middle school students' CT skills. Both rubric-based assessments and qualitative reflections revealed that the integration of interactive, iterative ML tasks created meaningful opportunities for students to engage in higher order thinking processes. The findings support the emerging perspective that ML is not only a technical subject but also a cognitively rich learning environment with the potential to foster metacognitive engagement and reasoning.

One of the most significant findings was the overall enhancement of CT dimensions such as goal setting, information selection, logical reasoning, and inference-making. This improvement appears to be linked to the iterative and hands-on structure of ML tasks, which inherently involve hypothesis testing, feedback loops, and decision making under uncertainty. These processes align closely with metacognitive strategies that facilitate CT development (Halpern, 2014; Facione, 1990). Similar outcomes were reported by Çakiroğlu and Selçuk (2024), who found that secondary students engaged in ML activities demonstrated substantial gains in analytical reasoning. Likewise, Sanusi et al. (2023) emphasized that contextualized, student driven ML tasks encouraged reflective thinking and inquiry-based learning. The current study extends these findings to a younger cohort, showing that even 5th and 6th-grade students can develop CT skills through appropriately scaffolded ML experiences. This contrasts with Lamb et al. (2021), who observed limited CT development in older students when tasks lacked authentic application contexts.

Another important contribution of this study lies in demonstrating the cumulative benefits of sequencing ML activities over time. Students' CT performance improved progressively across the five tasks, with notable gains in purposeful reasoning and application-oriented judgment. This supports theories of scaffolded learning (Vygotsky, 1978; Halpern, 2014), which emphasize the importance of structured exposure to complex tasks for the gradual internalization of cognitive strategies. Martins and von Wangenheim (2022) similarly highlighted that repeated engagement with ML tasks improves both computational and analytical skills. Research on cognitive apprenticeship (Naidu et al., 2020) also reinforces this view, suggesting that learners benefit from gradually increasing task complexity paired with consistent modeling and reflection. In contrast, studies using one-off ML activities without sequencing such as Shrigoud and Agrawal (2022) tend to report weaker CT outcomes, pointing to the value of curriculum designs that build progressively on prior knowledge.

The study also revealed that different phases of the ML process activated different CT dimensions. Among training, testing, and application phases, the application stage yielded the highest CT scores, particularly in goal clarity, question formulation, and inference-making. This supports the theory of situated cognition (Brown et al., 1989), which posits that learning is most effective when embedded in meaningful, real-world contexts. Lamb et al. (2021) observed similar results in students applying ML algorithms to real data, where authentic application prompted more integrated reasoning. Sanusi et al. (2023) also emphasized that real-world ML design tasks support students' epistemic agency. Meanwhile, lower performance in training and testing phases may be attributed to the abstract, technical nature of those tasks, which without sufficient scaffolding can overwhelm novice learners, as noted by Jin and Cutumisu (2023).

Framed through the three ML stages used in this study (training, learning & testing, application) and the eight CT dimensions in the rubric, the findings suggest a phase-sensitive pattern of CT activation.

Training tasks primarily elicited information identification and basic concept recognition, whereas learning and testing phases required students to monitor evidence, compare model outputs, and adjust decisions in response to feedback. The application phase, in turn, mobilized more complex CT processes such as goal clarification, question formulation, integration of multiple information sources, and application-oriented judgment. This pattern indicates that different segments of the ML workflow lend themselves to different CT sub-skills and that a balanced ML curriculum should intentionally orchestrate phases to ensure that all dimensions of CT are systematically engaged rather than assuming that "higher-order thinking" will emerge automatically.

Furthermore, the cluster analysis offered nuanced insights into student variability in CT engagement. High-performing students demonstrated traits of self-regulated learning (Zimmerman, 2002), including model revision, reflective thinking, and adaptive strategy use. These behaviors are strongly associated with deep CT engagement (Halpern, 2014). Medium-level learners showed partial understanding but struggled to apply structured problem-solving, echoing findings by Karalekas et al. (2023) who identified a similar need for targeted scaffolding in this group. The low-performing cluster characterized by limited goal clarity and difficulty executing ML tasks likely faced challenges related to cognitive load and insufficient prior experience (Sweller, 1998). This finding aligns with Shrigoud and Agrawal (2022), who stressed that novices often require explicit support to move beyond surface-level engagement.

Taken together, these findings reinforce the view that the pedagogical value of ML activities is not limited to their technical complexity but lies in how meaningfully they engage students in reflective, goal-oriented learning. Application oriented phases, in particular, afford students opportunities to synthesize information, evaluate outcomes, and make evidence-based decisions core competencies of critical thinking. To optimize the cognitive benefits of ML integration, instructional designs should incorporate sequenced tasks, authentic application, and differentiated scaffolding. High-performing learners may thrive in open-ended, inquiry-driven environments, whereas lower-performing students benefit from guided questioning, visual modeling, and structured decision pathways.

This study is among the first to systematically examine the interplay between ML task structure, cognitive phase engagement, and CT performance in early adolescence. The findings highlight the importance of developmentally appropriate, scaffolded ML experiences that promote not only technical fluency but also the cognitive dispositions required for navigating AI-mediated learning environments in the 21st century.

Implications and Limitations

This study offers several pedagogical implications for integrating ML into middle school curricula. ML activities that incorporate authentic datasets such as images, text, and numerical data—hold strong potential to enhance students' CT by fostering logical reasoning, reflective analysis, and iterative problem-solving. To fully harness this potential, teachers require professional development that addresses both the technical use of ML tools and the facilitation of CT-focused dialogue and reflection during classroom instruction.

In addition, the CT assessment rubric developed in this study provides a practical framework for evaluating students' cognitive engagement during ML activities and may be adapted to related domains, including ethical reasoning and AI literacy. Designing inclusive ML activities is also essential to accommodate students' diverse prior experiences; differentiated instruction should therefore be incorporated to ensure equitable access to meaningful ML learning opportunities.

At the curriculum level, we recommend sequencing ML tasks from highly guided to increasingly autonomous across units, explicitly aligning them with CT objectives and rubric criteria. Professional development initiatives should support teachers in designing such scaffolded sequences and in using analytic rubrics to make students' CT processes visible, discussable, and assessable in everyday classroom practice. These implications collectively highlight the promise of ML-enhanced instruction for strengthening middle school students' CT skills and AI literacy.

Furthermore, due to the exploratory case-study design, the small sample size, and unbalanced cluster sizes, we deliberately refrained from conducting inferential statistical tests between clusters; future studies with larger and more balanced samples should employ ANOVA or non-parametric equivalents to more rigorously examine between-group differences.

Despite its contributions, this study has several limitations. The small sample size (n = 15) and the use of a single-group intervention design without a control or comparison group restrict the generalizability of the findings and limit causal interpretations. Without a comparison condition, alternative explanations for improvements in CT such as concurrent curricular influences or general effects of teacher guidance cannot be ruled out. Moreover, the short-term implementation does not allow for conclusions about the sustainability of CT gains or their transfer to broader STEM learning.

Future research should therefore examine this instructional model with larger and more diverse populations, incorporate control or comparison groups ideally through quasi-experimental or experimental designs and employ longitudinal approaches to assess long-term cognitive outcomes. Such studies would more rigorously isolate the specific contribution of ML-based activities to CT development and provide stronger evidence for integrating ML into middle school curricula.

Author Contributions: The author solely contributed to the conceptualization, methodology, investigation, formal analysis, data curation, writing original draft preparation, writing review and editing, and visualization of this study. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical approval for this study was obtained from the Social and Human Sciences Research and Publication Ethics Committee of Trabzon University (Approval No. E-81614018-000-526). Permission to conduct the study was also obtained from the relevant school administration prior to data collection. All procedures involving human participants were conducted in accordance with the ethical principles of the Declaration of Helsinki and in compliance with relevant institutional and national regulations.

Informed Consent Statement: Informed consent was obtained from the parents/legal guardians of all participating students prior to data collection.

Data Availability Statement: The anonymized dataset, survey instruments, coding framework, and supporting materials used in this study are available for peer review and have been provided to the journal for editorial and reviewer access. All participant-identifying information has been removed to ensure confidentiality and privacy.

Acknowledgments: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for Teaching Students to Think Critically: A Meta-Analysis: A Meta-Analysis. Review of Educational Research, 85(2), 275-314. https://doi.org/10.3102/0034654314551063

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional Interventions Affecting Critical Thinking Skills and Dispositions: A Stage 1 Meta-Analysis. *Review of Educational Research*, *78*(4), 1102-1134. https://doi.org/10.3102/0034654308326084

Alsaleh, Nada J. (2020). Teaching Critical Thinking Skills: Literature Review. *Turkish Online Journal of Educational Technology – TOJET*, 19(1), 21-39.https://files.eric.ed.gov/fulltext/EJ1239945.pdf

Arndt, D. M., Martins, R. M., & Hauck, J. C. R. (2025). Critical Thinking Assessment in K-12 Computing Education: A Systematic Mapping. Informatics in Education, 24(1), 1-44.https://doi.org/10.15388/infedu.2025.02

Bailin, S. (2002). Critical thinking and science education. *Science & Education, 11*(4), 361–375. https://doi.org/10.1023/A:1016042608621

Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of machine learning models towards student academic performance prediction: a systematic review. *Applied Sciences*, *11*(21), 10007. https://doi.org/10.3390/app112110007

Banadaki, Y. M. (2020). Enhancing the role of machine learning in STEM disciplines through supervised undergraduate research experiences. *Literacy Information and Computer Education Journal (LICEJ)*,*11*(2). https://doi.org/10.20533/licej.2040.2589.2020.0449

Bellaera, L., Weinstein-Jones, Y., Ilie, S., & Baker, S. T. (2021). Critical thinking in practice: The priorities and practices of instructors teaching in higher education. *Thinking Skills and Creativity*, *41*, 100856.https://doi.org/10.1016/j.tsc.2021.100856

Bloom, B. S. (1956). Taxonomy of. *Educational Objectives*, *250*.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32-42. https://doi.org/10.3102/0013189X018001032

Çakiroğlu, Ü., & Selçuk, V. (2024). Machine learning meets secondary school classrooms: using hands-on activities to advance computational thinking. *Education and Information Technologies*, 1-25. https://doi.org/10.1007/s10639-024-13196-8

Cansu, F. K., & Cansu, S. K. (2019). An Overview of Computational Thinking. *International Journal of Computer Science Education in Schools*, *3*(1), 17–30. https://doi.org/10.21585/ijcses.v3i1.53

Case, R. (2005). Moving critical thinking to the main stage. *Education Canada*, *45*(2), 45-49. https://eric.ed.gov/?id=EJ736084

Çeviker Ay, Ş., & Orhan, A. (2020). The Effect of Different Critical Thinking Teaching Approaches on Critical Thinking Skills: A Meta-Analysis Study. Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 49, 88-111. https://doi.org/10.9779/pauefd.561742

Chung, C., & Shamir, L. (2021). Introducing machine learning with Scratch and robots as a pilot program for K–12 computer science education. *International Journal of Learning and Teaching*, 7(3), 181–186. https://doi.org/10.18178/ijlt.7.3.181-186

Elder, L., & Paul, R. (2020). *Critical thinking: Tools for taking charge of your learning and your life*. Rowman & Littlefield.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational leadership*, *43*(2), 44-48. https://jgregorymcverry.com/readings/ennis1985assessingcriticalthinking.pdf

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. The Delphi Report. American Philosophical Association. https://eric.ed.gov/?id=ED315423

Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill. *Informal logic*, *20*(1). https://doi.org/10.22329/il.v20i1.2254

Ge, X., Fang, C., Li, X., Sun, W., Wu, D., Zhai, J., ... & Chen, Z. (2024). Machine learning for actionable warning identification: A comprehensive survey. *ACM Computing Surveys*, *57*(2), 1-35. https://doi.org/10.1145/3696352

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.

Granone, F., Kirsti Lie Reikerås, E., Pollarolo, E., & Kamola, M. (2023). Critical Thinking, Problem-Solving and Computational Thinking: Related but Distinct? An Analysis of Similarities and Differences Based on an Example of a Play Situation in an Early Childhood Education Setting. IntechOpen. https://doi.org/10.5772/intechopen.110795

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*(4), 449–455. https://doi.org/10.1037/0003-066X.53.4.449

Halpern, D. F. (2001). Assessing The Effectiveness Of Critical Thinking Instruction. *The Journal of General Education*, *50*(4), 270–286. http://www.jstor.org/stable/27797889

Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking* (5th ed.). Psychology Press.

Heard, J., Scoular, C., Duckworth, D., Ramalingam, D., & Teo, I. (2025). Critical thinking: Skill development framework. 2nd edition. Australian Council for Educational Research. https://doi.org/10.37517/978-1-74286-752-6

Hutchins, N. M., & Biswas, G. (2024). Co-designing teacher support technology for problem-based learning in middle school science. *British Journal of Educational Technology, 55*(3), 802-822. https://doi.org/10.1111/bjet.13363

Irgens, G. A., Vega, H., Adisa, I., & Bailey, C. (2022). Characterizing children's conceptual knowledge and computational practices in a critical machine learning educational program. *International Journal of Child-Computer Interaction*, 34, 100541. https://doi.org/10.1016/j.ijcci.2022.100541

Jamil, M., Bokhari, T. B., & Rafiq, M. (2024). Critical thinking skills development for 21st century: An analysis of Biology curriculum (2006). Voyage Journal of Educational Studies, 4(1), 127-138. https://doi.org/10.58622/vjes.v4i1.132

Jin, H. Y., & Cutumisu, M. (2023). Predicting pre-service teachers' computational thinking skills using machine learning classifiers. *Education and Information Technologies*, *28*(9), 11447-11467. https://doi.org/10.1007/s10639-023-11642-7

Karalekas, G., Vologiannidis, S., & Kalomiros, J. (2023). Teaching Machine Learning in K–12 Using Robotics. *Education Sciences*, *13*(1), 67. https://doi.org/10.3390/educsci13010067

Katchapakirin, K., Anutariya, C., & Supnithi, T. (2022). ScratchThAI: A conversation-based learning support framework for computational thinking development. *Education and Information Technologies*, *27*(6), 8533-8560. https://doi.org/10.1007/s10639-021-10870-z

Kestel, M. (2022). *Critical thinking skills: A meta-analysis study* (Doctoral dissertation, Necmettin Erbakan University, Turkey). https://hdl.handle.net/20.500.12452/8304

Khan, R. U., Cheng, J. L. A., & Bee, O. Y. (2018). Machine Learning and Dyslexia: Diagnostic and Classification System (DCS) for Kids with Learning Disabilities. *International Journal of Engineering & Technology*, *7*(3.18), 97-100. https://doi.org/10.14419/ijet.v7i3.18.19022

Kong, S. C., & Shen, W. (2024). Using students' cognitive, affective, and demographic characteristics to predict their understanding of computational thinking concepts: a machine learning-based approach. *Interactive Learning Environments*, *32*(10), 7647–7660. https://doi.org/10.1080/10494820.2024.2331148

Kules, B. (2016). Computational thinking is critical thinking: Connecting to university discourse, goals, and learning outcomes. *Proceedings of the association for information science and technology*, *53*(1), 1-6. https://doi.org/10.1002/pra2.2016.14505301092

Lamb, R., Hand, B., & Kavner, A. (2021). Computational modeling of the effects of the science writing heuristic on student critical thinking in science using machine learning. *Journal of Science Education and Technology*, *30*(2), 283-297. https://doi.org/10.1007/s10956-020-09871-3

Liu, J., Ma, J. & Li, S. Research on school-based AI curriculum design and practice for cultivating computational thinking in high school students. *Educ Inf Technol* **30**, 7949–7993 (2025). https://doi.org/10.1007/s10639-024-13115-x

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. doi:10.15388/infedu.2020.14

Orhan, A., & Çeviker Ay, Ş. (2023). How to teach critical thinking: An experimental study with three different approaches. *Learning Environments Research, 26*, 199–217. https://doi.org/10.1007/s10984-022-09413-1

Pacheco, J., Ferreira, J., Tavares, H., & Miranda, M. (2023). *Machine learning tool for kids: A contribution to teaching computational thinking in schools*. ResearchGate. https://www.researchgate.net/publication/373924038_Machine_Learning_Tool_for_Kids_A_Contribution_to_Teaching_Computational_Thinking_in_Schools

Paul, R., & Elder, L. (2006). *Critical thinking competency standards*. Dillon Beach, CA: Foundation for Critical Thinking.

Rani, S., Bhambri, P., & Chauhan, M. (2021). A machine learning model for kids' behavior analysis from facial emotions using principal component analysis. In *2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT)* (pp. 522-525). IEEE. 10.1109/ACAIT53529.2021.9731203

Sanusi, I. T., Oyelere, S. S., Vartiainen, H., Suhonen, J., & Tukiainen, M. (2023). Developing middle school students' understanding of machine learning in an African school. *Computers and Education: Artificial Intelligence*, *5*, 100155. https://doi.org/10.1016/j.caeai.2023.100155

Shamir, G., & Levin, I. (2021). Teaching machine learning in elementary school. *International Journal of Child-Computer Interaction*, 31, 100415. https://doi.org/10.1016/j.ijcci.2021.100415

Shrigoud, S., & Agrawal, S. (2022) Student Performance Prediction using Machine Learning Algorithms: A Review. *International Journal of Computer Applications*, *32*(184), 8887. https://doi.org/10.5120/ijca2022922396

Sonsun, P., Hemtasin, C., & Thongsuk, T. (2023). Development of Science Learning Activities Using Inquiry-Based Learning Management to Improve the Academic Achievement of Secondary School Students. *Journal of Education and Learning*, *12*(3), 86-92. https://eric.ed.gov/?id=EJ1385197

Sreejun, S., & Chatwattana, P. (2023). The Imagineering Learning Model with Inquiry-Based Learning via Augmented Reality to Enhance Creative Products and Digital Empathy. *Journal of Education and Learning*, *12*(2), 52-59. https://eric.ed.gov/?id=EJ1385143

Su, J., & Zhong, Y. (2022). Artificial Intelligence (AI) in early childhood education: Curriculum design and future directions. *Computers and Education: Artificial Intelligence, 3*, 100072. https://doi.org/10.1016/j.caeai.2022.100072

Sun, L., Hu, L., & Zhou, D. (2022). Programming attitudes predict computational thinking: Analysis of differences in gender and programming experience. *Computers & Education*, *181*, 104457. https://doi.org/10.1016/j.compedu.2022.104457

Sweller, J. (1998). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Watson G., Glaser E. M. (1994). *Watson–Glaser Critical Thinking Appraisal, Form S manual*. San Antonio, TX: The Psychological Corporation.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3), 33–35. https://doi.org/10.1145/1118178.1118215

Zhao, Y., Liu, Y., Pásztor, A., & Molnár, G. (2025). Relationships Among Components of Critical Thinking Disposition in Primary School Students: An Investigation Based on Hungarian Context. *Psychology in the Schools*. https://doi.org/10.1002/pits.23445

Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Appendix 1

Critical Thinking Component	Process Phase	Proficient (2 pts)	Developing (1 pt)	Insufficient (0 pt)
Goal Identification	Training	Clearly defines the project goal.	Partially defines the project goal.	Does not define the project goal.
Question Formulation	Training	Develops appropriate questions aligned with the goal.	Partially develops goal-oriented questions.	Cannot formulate relevant questions.
Relevant Information Selection	Training	Selects information relevant to the project goal.	Partially selects relevant information.	Fails to identify relevant information.
Criterion Definition and Testing	Learning & Testing	Defines and tests criteria appropriate to the project goal.	Partially defines and tests relevant criteria.	Fails to define or test relevant criteria.
Question Testing	Learning & Testing	Teachers and tests project questions effectively.	Partially teaches and tests project questions.	Fails to teach or test questions.
Information Application in Testing	Learning & Testing	Identifies and tests relevant information for implementation.	Partially identifies and tests relevant information.	Fails to identify or test information.
Application Development	Application	Develops an application aligned with the project goal.	Partially develops an appropriate application.	Fails to develop a suitable application.
Question Integration in Application	Application	Integrates questions into the application process.	Partially integrates questions.	Fails to integrate questions.
Information Use in Application	Application	Use relevant information effectively in the application.	Partially uses relevant information.	Does not use relevant information.
Concept Identification	Training	Clearly identify key project concepts.	Partially identifies key concepts.	Fails to identify concepts.
Reasoning and Assumptions	Training	Identifies reasoning and assumptions in the process.	Partially identifies assumptions.	Fails to identify reasoning and assumptions.
Justification and Perspective Sharing	Training	Clearly presents appropriate views related to the goal.	Partially presents appropriate views.	Fails to articulate a coherent perspective.
Concept Teaching	Learning & Testing	Effectively teaches concepts through the application.	Partially teaches key concepts.	Fails to teach core concepts.
Assumption Testing	Learning & Testing	Tests assumptions within the application process.	Partially tests assumptions.	Fails to test assumptions.
Perspective Testing	Learning & Testing	Tests and reflect on personal viewpoints within application.	Partially tests personal viewpoints.	Fails to test personal perspectives.
Concept Application	Application	Apply identified concepts in practice.	Partially applies concepts.	Fails to apply concepts.
Assumption Application	Application	Integrates identified assumptions into application use.	Partially integrates assumptions.	Fails to integrate assumptions.
Viewpoint Application	Application	Applies and reflects on personal views in application.	Partially applies views.	Does not apply personal views.
Inference Making	Training	Clearly articulates inferences from the project.	Partially articulates inferences.	Fails to articulate inferences.
Conclusion Selection	Training	Selects the most appropriate project outcomes.	Selects reasonably appropriate outcomes.	Fails to identify appropriate conclusions.
Inference Testing	Learning & Testing	Tests inferences within the application.	Partially test inferences.	Does not test inferences.
Conclusion Testing	Learning & Testing	Tests selected conclusions in the application.	Partially test conclusions.	Does not test conclusions.
Inference Application	Application	Applies tested inferences effectively.	Partially applies inferences.	Does not apply inferences.
Conclusion Application	Application	Apply appropriate conclusions in implementation.	Partially applies conclusions.	Fails to apply conclusions.

Appendix 2

Actions involved in machine learning activities	Student Statements	Matching critical thinking steps
Training	*"I paid attention to the details."* ***(S11)***	Identifying fundamental concepts appropriate to the purpose
Training	*"When creating the project, I first focused on the project's purpose. Then I created the tags according to the project's purpose."* ***(S3)***	Goal setting Identifying fundamental concepts appropriate to the purpose
Learning & Testing	*"After completing the labels, I generally decided what could make me happy in my own life. Then I thought about how I could teach this to the machine, and I taught the machine by giving it many examples. Because if I give it many examples, the machine will learn easily and won't make mistakes"* ***(S10)***	Goal setting Purposeful reasoning
Application	*"I wrote down my thoughts on the paper the teacher gave us about what the results of the activity might be. Then I did the activity. At the end of the activity, I wrote down what I had written on the paper."* ***(S14)***	Achieving Results Deduction Determination
Training	*"When working on the project, I first tried to think like a machine. That is, I thought about how I could answer whatever questions might be asked, and I determined the questions before creating the tags."* ***(S6)***	Identifying the questions
Training Learning & Testing Application	*"Teacher, there were parts of the activities where I needed to determine information. I thought about what information I needed to teach the computer, and I taught it to the computer. Then I did my application with this information."* ***(S13)***	identifying information relevant to the purpose
Training Application	*"I determined my views on the activity. When determining my views, I first paid attention to the purpose of the activity, then thought about how I could implement it, and wrote down my thoughts while paying attention to these points during implementation."* ***(S1)***	Expressing an opinion regarding the application Goal setting
Training Application	*"After my teacher determined the activities, I determined what the results might be. Then, after the application, I saw what I wrote after the application at the end of the activity."* ***(S5)***	Deduction Determination Achieving Results