Jurnal Pendidikan Progresif
e-ISSN: 2550-1313 | p-ISSN: 2087-9849
URL: https://jpp.fkip.unila.ac.id/index.php/jpp
Jounal's DOI: http://doi.org/10.23960/jpp

Vol. 16, No. 3, 2026
Article ID: 1866
Pages: 1679-1703

Fostering Algorithmic Thinking through AI-Supported STEM Activities: Experimental Study Using Cacao Disease Classification

Firma Nur Muttakin[1,*], Dafik[2], Zainur Rasyid Ridlo[3], & Udaya Sri Kakarla[4]
[1]Postgraduate Student of Science Education, Universitas Jember, Indonesia
[2]Department of Mathematics Education, Universitas Jember, Indonesia
[3]Department of Science Education, Universitas Jember, Indonesia
[4]KG Reddy College of Engineering and Technology, India

*Corresponding email: firmanur4032@gmail.com

Received: 15 June 2026 *Accepted: 13 August 2026* *Published: 26 August 2026*

Abstract: This study investigated the effect of AI-integrated STEM activities using cocoa leaf disease classification materials on undergraduate students' algorithmic thinking skills. This quasi-experimental study used a non-randomized pretest–posttest comparison design and involved 66 undergraduate students. Data were collected using an algorithmic thinking test and an instrument validation sheet. The instrument demonstrated acceptable content validity, with a mean validation score of 4.57, a feasibility percentage of 91.38%, Aiken's V coefficient of 0.892, and very high internal consistency (Experimental á = 0.942 and Control á = 0.938). The data were analyzed using assumption testing, Welch's t-tests, gain-score analysis, and ANCOVA. The experimental group achieved a substantially larger mean gain than the control group (d = 1.43) and a higher average normalized gain (0g 0 = 0.87 versus 0.52), and outperformed the control group on all seven algorithmic thinking indicators (d = 0.72 to 1.60). The ANCOVA indicated a significant group difference in posttest performance, F(1, 63) = 29.302, p < .001, with a higher adjusted mean in the experimental group (M = 76.8) than in the control group (M = 66.5). However, the pretest covariate explained virtually none of the posttest variance, so the covariance adjustment was negligible, and the two intact classes differed significantly at baseline. Because posttest variance was also heterogeneous, interpret the findings with appropriate caution and understand them as associational rather than causal. These results suggest that AI-integrated STEM activities based on cocoa leaf disease classification may support the development of undergraduate students' algorithmic thinking through structured, iterative, and evidence-based learning processes.

Keywords: algorithmic thinking, STEM activities, cacao diseases, classification, undergraduate students.

Article's DOI: https://doi.org/10.23960/jpp.v16i3.pp1679-1703

INTRODUCTION

Twenty-first-century science education requires learners to develop systematic, logical, and complex problem-solving skills to respond effectively to rapid advances in science, technology, and society. These demands have shifted educational priorities toward developing higher-order cognitive competencies, including problem-solving, critical thinking, and computational thinking, as essential foundations for preparing future educators and professionals (Ogegbo & Ramnarain, 2022; Rusmin *et al.*, 2024). In Indonesia, this shift is explicitly reflected in national education policy, where computational thinking is positioned as a core competency in primary and secondary education and is expected to be developed in science and STEM learning contexts (Ministry of Primary and Secondary Education, 2025). Within this policy framework, attention has increasingly focused on learning designs that operationalize computational thinking in authentic and contextually meaningful learning environments.

Computational thinking is a multidimensional cognitive competence that enables learners to formulate problems and develop solutions through systematic processes such as decomposition, pattern recognition, abstraction, and algorithmic reasoning (Wing, 2006; Ridlo, Ningsih, *et al*., 2025). Rather than being limited to programming, computational thinking involves applying structured problem-solving processes across diverse contexts, including science and STEM education. These processes enable learners to break down complex problems, identify relevant patterns, abstract essential information, and develop systematic solutions. One component of computational thinking that deserves attention is algorithmic thinking.

Algorithmic thinking refers to the ability to construct, organize, evaluate, and refine systematic procedures for solving problems through logical and sequential reasoning. As an essential component of computational thinking, algorithmic thinking has become increasingly important in science and STEM education because it enables learners to develop, test, and improve solutions to complex problems (Lehmann, 2024; Wong *et al*., 2024). However, previous studies show that many students still struggle to construct valid algorithms, debug errors, and refine procedural solutions during problem-solving activities (Tupouniua, 2023; Sun *et al*., 2024). These findings suggest that algorithmic thinking does not develop automatically through conventional instruction but requires learning experiences that explicitly engage students in designing, testing, evaluating, and improving solution procedures. In this regard, Research-Based Learning (RBL) provides a promising instructional approach by involving students in authentic inquiry, problem-solving, and iterative solution development, thereby creating meaningful opportunities to foster algorithmic thinking.

Research-Based Learning (RBL) is a pedagogical approach that engages students in authentic inquiry and systematic problem-solving. By positioning students as active researchers, RBL involves processes such as problem formulation, evidence collection, data analysis, solution development, evaluation, and communication of research findings (Dafik *et al*., 2023). These processes are highly relevant to algorithmic thinking because they require students to organize complex tasks into systematic procedures and iteratively refine solutions based on evidence. One approach that can be used with the research-based learning model is STEM (science, technology, mathematics, and engineering).

STEM education provides an interdisciplinary learning framework that connects science, technology, engineering, and mathematics through authentic and problem-oriented activities. Recent research indicates that integrated STEM learning can promote higher-order thinking by engaging students in problem identification, solution design, experimentation, data analysis, and reflection (Astawan *et al*., 2023; Le *et al.,* 2023). These processes closely relate to algorithmic thinking because students must organize procedures, develop systematic solutions, evaluate outcomes, and refine their reasoning. However, developing these competencies increasingly requires learning environments that give learners opportunities to engage with contemporary technologies and complex data-driven problems.

Artificial intelligence (AI) has emerged as a promising technological component for strengthening STEM learning by enabling students to engage with classification, prediction, and data-driven decision-making. Recent research has highlighted AI's potential to support inquiry, problem-solving, and higher-order thinking in science education by giving learners opportunities to interact with complex data and technological systems (Zhai *et al*., 2021; Almasri, 2024). In AI-supported classification activities, students must organize information, identify relevant features, evaluate outputs, and refine solutions based on evidence. These processes give students opportunities to engage in systematic, iterative reasoning rather than merely using AI as a passive technological tool. Therefore, the educational value of AI depends not only on the technology itself but also on the authenticity and cognitive richness of the problem in which it is embedded. This consideration highlights the importance of situating AI-supported STEM learning within meaningful scientific contexts.

Cocoa (*Theobroma cacao* L.) is an important plantation commodity in Indonesia and plays a significant role in the country's agricultural and economic activities (Ramadhani *et al*., 2023). East Java, specifically Jember Regency, provides a relevant local context for this study because cocoa production and related agricultural research have been conducted there (Thifany *et al*., 2020). Cocoa in Jember is highly diverse and has distinct leaf characteristics, as shown in Figure 1. This diversity may be associated with differences in disease susceptibility and resistance. Cocoa diseases present complex biological problems that require observing and classifying visual symptoms, comparing diagnostic features, and making evidence-based decisions. Research on cocoa disease management in Indonesia further highlights the importance of identifying and managing pest and disease problems to support cocoa production (Sari *et al*., 2025). Because disease symptoms are often manifested in visible leaf features, cocoa leaves provide a suitable image-based substrate for AI-assisted disease classification (Ferentinos, 2018; Mohanty *et al*., 2016). These characteristics make cocoa leaf disease classification a meaningful context for AI-assisted STEM learning because the process requires learners to organize information, apply classification criteria, evaluate diagnostic results, and refine their decisions.

Three important research gaps remain in the current literature. First, an instrument gap concerns assessing algorithmic thinking in undergraduate science education. Recent studies have developed specific tools for assessing algorithmic thinking, including game-based tests and digital or unplugged assessment environments (Zengin & Karal, 2024; Adorni et al., 2024). However, existing instruments differ in their conceptualizations, assessment formats, and target populations, and much recent work focuses on school-age learners and algorithm-construction tasks. Limited research has developed and psychometrically evaluated an instrument that operationalizes algorithmic thinking through multiple dimensions relevant to undergraduate science learning, including sequencing, control flow understanding, algorithm design, representation, debugging, optimization, and generalization. This limitation underscores the need for valid and reliable assessment instruments that can capture algorithmic thinking as a multidimensional competence in higher education science contexts.

Second, a contextual gap concerns the scientific problems in which AI-supported STEM learning is situated. Existing studies have largely embedded artificial intelligence in generic, decontextualized, or pre-packaged datasets such as ready-made image libraries and simulated classification tasks, in which learners operate a model without confronting the ambiguity of authentic scientific evidence (Zhai et al., 2021; Almasri, 2024). Because the training data are already curated, students are rarely required to establish classification criteria, justify category boundaries, or account for the biological variability that generates misclassification in the first place. Authentic and locally meaningful scientific contexts, such as plant disease diagnosis, offer precisely these affordances, since disease symptoms are visually graded, overlapping, and variable across cultivars (Ferentinos, 2018; Mohanty et al., 2016). Nevertheless, such locally grounded agricultural contexts have seldom been converted into structured learning designs that make students responsible for constructing the dataset, the labels, and the decision rules. This gap indicates that the cognitive value of AI-supported STEM learning cannot be separated from the authenticity of the scientific problem in which the technology is embedded.

Third, an empirical gap concerns the evidential basis for claims about the effectiveness of AI-supported STEM learning at the undergraduate level. Reviews of computational thinking in science and STEM education indicate that the available evidence is dominated by K-12 populations, descriptive or small-scale designs, and composite measures of computational thinking in which algorithmic thinking is not reported as a distinct outcome (Ogegbo & Ramnarain, 2022; Tang et al., 2020; Li & Oon, 2024). Comparative studies that contrast an AI-supported STEM condition with an equivalent non-AI condition while statistically controlling for baseline differences remain scarce, particularly among preservice teachers and other undergraduate science learners. As a result, it is still unclear whether the gains attributed to AI-supported learning reflect the instructional design itself or merely pre-existing differences in prior ability. Addressing this gap requires quasi-experimental evidence which algorithmic thinking is measured with a validated multidimensional instrument and analyzed with procedures that account for non-equivalent groups.

Figure 1. Differences in varieties of cocoa leaves

Integrating cocoa disease classification into AI-supported RBL-STEM activities therefore provides a theoretically grounded response to the identified research gaps. Previous research shows that AI can support science learning by engaging students in data-driven inquiry, classification, and problem-solving processes (Zhai *et al*., 2021; Almasri, 2024). Research on integrated STEM education has also emphasized the importance of authentic, interdisciplinary, and problem-oriented learning experiences for developing higher-order thinking and applying knowledge across disciplinary boundaries (Thibaut *et al*., 2018; Li *et al*., 2020). In addition, research-based STEM learning can engage students in inquiry, data analysis, solution development, evaluation, and communication of evidence-based outcomes (Dafik et al., 2023). Together, these findings suggest that combining AI-supported classification with research-based inquiry and interdisciplinary STEM learning can provide a structured environment for students to engage in iterative and evidence-based reasoning. In the present study, this framework was contextualized through cocoa leaf disease classification. Through machine learning applications such as Google Teachable Machine, students engaged in activities that required them to organize data, evaluate classification outputs, identify inaccuracies, and refine solution strategies. The cocoa disease context provided an authentic scientific problem, while the RBL-STEM framework structured students' engagement with inquiry, technology, and problem-solving. Accordingly, AI was positioned not merely as a technological tool but as a component of a learning environment in which students could develop and apply algorithmic thinking through observable and revisable problem-solving processes. Based on the identified research gap, this addresses the following research questions:

RQ1. To what extent is the algorithmic thinking test valid for measuring undergraduate students' algorithmic thinking skills?
RQ2. To what extent does the algorithmic thinking test demonstrate internal consistency in measuring undergraduate students' algorithmic thinking skills?
RQ3. To what extent do AI-integrated STEM activities using cocoa leaf disease classification materials affect undergraduate students' algorithmic thinking skills?

METHOD

Research Design

This study used a quasi-experimental methodology based on a non-randomized pre-intervention and post-intervention comparison between two groups (Ridlo, Dafik, *et al*., 2025). Figure 2 shows the research flow. Within this framework, participants assigned to the experimental condition completed an initial assessment prior to the instructional intervention and a final assessment after the treatment was implemented. In parallel, a comparison group with no treatment exposure completed the same sequence of assessments without receiving the instructional intervention. The overall structure of this design is summarised below.

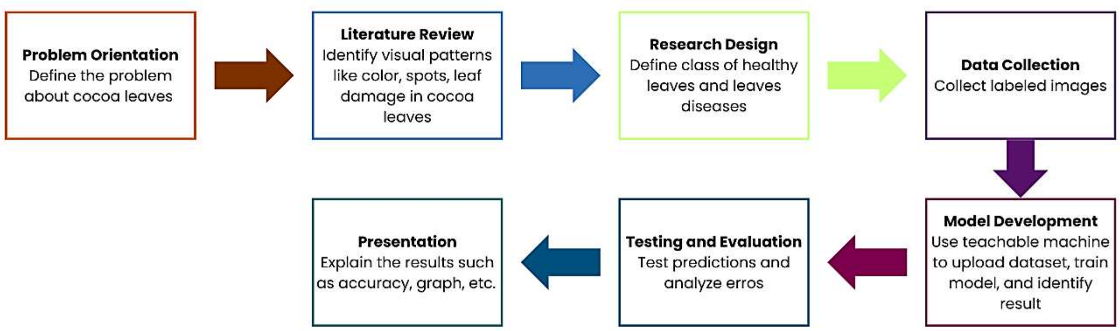

Figure 2. Research flowchart

Formally, the design comprised two non-equivalent intact groups measured on two occasions. The experimental group completed an initial assessment (O1), received the STEM-based instructional intervention using cacao disease materials (T1), and then completed a final assessment (O2). In parallel, the comparison group completed an initial assessment (O3), followed a contextual learning approach applied to the same cacao disease materials without AI support (T2), and then completed the same final assessment (O4). Both groups were therefore assessed with the identical instrument at both time points, and two conditions differed only in the presence of AI-supported classification activities.

Before the instructional intervention began, both the experimental and control groups completed the same algorithmic thinking pretest to establish baseline equivalence. The intervention was conducted over eight weeks, with one instructional session of 2 × 50 minutes per week. Both groups studied the same topic, cocoa leaf diseases, and were taught by the same teacher using equivalent learning objectives, instructional duration, classroom organization, and assessment procedures.

In week 1 (phase 1: orientation and problem identification), students in the experimental group were introduced to authentic cases of cocoa leaf diseases found in local agricultural environments and identified plant-health problems. This phase emphasized the science component of STEM and primarily developed the generalization indicator of algorithmic thinking by encouraging students to recognize patterns across different disease cases. In the control group, the teacher introduced cocoa leaf diseases through contextual examples from the local environment and guided students through question-and-answer activities to connect the content with their daily experiences.

In week 2 (phase 2: literature exploration and observation), experimental-group students explored scientific references, observed healthy and diseased cocoa leaves, and recorded distinguishing characteristics of each disease. The activities integrated science through systematic observation and data organization and targeted the representation indicator of algorithmic thinking. In the control group, students observed printed images and teacher-provided cocoa leaf samples and completed observation worksheets based on textbook descriptions and classroom explanations.

During week 3 (phase 3: research planning), students in the experimental group worked collaboratively to design procedures for collecting cocoa leaf images and planned a simple investigation for disease identification. This phase reflected the engineering aspect of STEM and developed the sequencing indicator by requiring students to organize and arrange investigation steps logically. In contrast, the control group followed a predetermined procedure demonstrated by the teacher and practiced the sequence of disease-identification activities provided in the learning materials.

In week 4 (phase 4: data collection and STEM investigation), experimental-group students collected cocoa leaf image data, organized their observations, and prepared datasets for subsequent analysis. This phase integrated science and technology and strengthened control flow understanding as students managed ordered procedures during data collection and preparation. In the control group, students conducted observation and classification exercises using teacher-provided materials and organized observational data according to the teacher's instructions.

In week 5 (phase 5: AI-supported analysis and solution development), students in the experimental group used Google Teachable Machine to train and test a simple image-classification model for cocoa leaf diseases. They interpreted prediction results, analyzed classification errors, and discussed possible improvements to the model, integrating technology and engineering while developing algorithm design and debugging/error-detection skills. In the control group, students manually compared leaf characteristics with disease-identification criteria provided in textbooks and worksheets and discussed possible diagnoses without using AI-based image classification tools.

In week 6 (phase 6: optimization and application), experimental-group students refined image inputs, evaluated model performance, and proposed improvements to both the disease-identification procedure and the model. This phase integrated engineering and mathematics and focused on the optimization indicator of algorithmic thinking. The control group reviewed the results of manual disease identification, corrected classification errors based on teacher feedback, and revised worksheet responses through guided classroom discussion.

In week 7 (phase 7: presentation and scientific communication), students in the experimental group presented their investigation process, AI classification outcomes, and evidence-based recommendations for cocoa disease management. This phase integrated science and technology and further developed representation and generalization skills through scientific communication and justification of findings. The control group presented the results of manual disease-identification activities and discussed appropriate treatment recommendations based on classroom learning materials.

Finally, in week 8 (phase 8: conclusion and evaluation), students in the experimental group synthesized the findings from the research process, formulated evidence-based conclusions about the effectiveness of AI-assisted cocoa leaf disease identification, and completed the posttest. The control group summarized the main concepts learned through contextual instruction, formulated conclusions about cocoa leaf disease identification, and completed the same posttest. The parallel instructional structure ensured that both groups received equivalent instructional time and learning content, while differing in the implementation of the STEM framework with AI support in the experimental group and the contextual learning approach without AI support in the control group.

Participant and Sampling Technique

This study involved 66 undergraduate students from the Mathematics Education Program at the University of Jember. The participants were selected using criterion-based purposive sampling with an intact-group selection approach. This sampling technique was used because the study required participants who had relevant prior knowledge and learning experiences related to STEM-oriented instruction. Specifically, the selected students were enrolled in a science learning innovation course that emphasized applying STEM-based instructional approaches. Two existing groups were selected, each consisting of 33 students. One group was designated as the treatment group, while the other was assigned as the comparison group. Participants in the treatment group received the designated instructional intervention involving STEM-based learning integrated with artificial intelligence-based cacao leaf disease classification, whereas participants in the comparison group followed a contextual learning approach using the same cocoa leaf disease materials without AI support; the parallel activities of each phase are detailed in the Research Design section. The study was implemented across one academic term using a hybrid instructional format that integrated in-person sessions with online learning activities, in accordance with curricular requirements and institutional timetables.

Ethical approval for this study was granted by the Institutional Review Board of the Faculty of Teacher Training and Education, Universitas Jember (Reference Number 9754/DST/UN25.B5/SP/2026) prior to data collection. All participants were informed about the purpose, procedures, potential benefits, and voluntary nature of the study before participation. Written informed consent was obtained from all participants. Participants were informed that participation was voluntary, that they could withdraw at any time without academic consequences, and that their identities and individual responses would be kept confidential. The data were analyzed and reported in an anonymized form and were used solely for research purposes.

Research Instrument

This study used two instruments: an algorithmic thinking test and an expert validation sheet. The algorithmic thinking test was developed to measure undergraduate students' algorithmic thinking skills before and after the instructional intervention. The test consisted of 28 items distributed across seven indicators: sequencing, control flow understanding, algorithm design, algorithm representation, debugging and error detection, algorithm optimization, and generalization of algorithmic procedures. Each indicator was represented by four problem-based items embedded in cocoa leaf disease classification contexts. The items were designed to assess students' ability to organize procedures, apply conditional rules, construct and represent algorithms, identify procedural errors, optimize solutions, and transfer algorithmic reasoning to new conditions. Each item was scored on a four-point rubric ranging from 0 to 3, giving a maximum of 12 points per indicator and 84 points for the whole instrument. The same test instrument was administered as a pretest and posttest to measure changes in students' algorithmic thinking performance.

An expert validation sheet was used to evaluate the content and feasibility of the algorithmic thinking test. The validation sheet employed a five-point rating scale ranging from 1 (very inappropriate) to 5 (very appropriate). The assessment covered the relevance of the items to the indicators of algorithmic thinking, the clarity of instructions and item statements, the suitability of the cognitive demand, the contextual relevance of the cocoa disease classification scenarios, and the overall feasibility of the instrument. An expert reviewed the instrument before it was used in the main study. The validation results were then used to determine the instrument's content feasibility.

The algorithmic thinking test was further evaluated for internal consistency using Cronbach's alpha based on study participants' responses. The reliability analysis was conducted to determine the consistency of the items in measuring the intended construct. The resulting coefficient was interpreted as evidence of the instrument's internal consistency.

Research Procedures

The learning treatment applied to the experimental group consisted of AI-integrated STEM activities centered on classifying cocoa plant diseases, as shown in Figure 3. These activities utilized Google Teachable Machine as a machine learning platform to support image-based classification tasks. Students engaged in a structured learning sequence that included: (1) organizing step-by-step procedures for identifying cacao leaf diseases, (2) designing decision branches for disease classification, (3) constructing diagnostic algorithms, (4) representing algorithms in the form of flowcharts or pseudocode, (5) testing and debugging prediction errors generated by the model, (6) optimizing model performance through iterative refinement, and (7) testing model generalization using new images and different cacao varieties. Through this process, AI functioned as an epistemic scaffold that externalized students' algorithmic reasoning into explicit and observable procedures. The instructional syntax used in this study follows an RBL-STEM structure in which students engage in research-oriented activities, beginning with problem formulation and ending with project reporting and presentation. The syntax connects scientific problem identification, engineering-based machine learning preparation, technological data collection and Android application development, and mathematical analysis of classification performance.

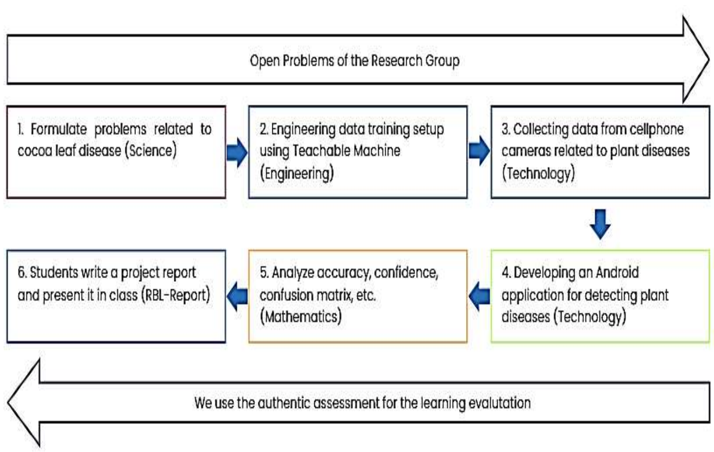

Figure 3. Syntax of RBL-STEM integrated with CNN and computer vision

Students' algorithmic thinking was assessed using a cognitive-based test developed to measure algorithmic thinking as a distinct construct within a STEM learning context. The instrument operationalized algorithmic thinking into seven indicators: (1) sequencing, (2) control flow understanding, (3) algorithm design, (4) representation of algorithms, (5) debugging and error detection, (6) algorithm optimization, and (7) generalization of algorithmic procedures. These indicators were derived from established theoretical frameworks of computational and algorithmic thinking.

Each indicator was measured using problem-based test items embedded in cacao disease classification scenarios. For example, sequencing items required students to arrange diagnostic steps in a logical order; debugging items asked students to identify and correct errors in a given classification procedure; and optimization items required students to propose improvements to increase classification accuracy or efficiency. Table 1 presents the mapping between students' algorithmic thinking indicators and sample test items. Measurements were conducted at the outset and at the conclusion of the instructional period using the same assessment instrument in order to determine changes in students' algorithmic thinking.

Table 1. Algorithmic thinking indicators and conceptual description of test items

Algorithmic Thinking Indicator	Cognitive Focus	Conceptual Description of Item Content	Number of Items
Sequencing	Logical ordering	Items require students to determine the correct order of procedural steps involved in identifying cacao leaf diseases, emphasizing step-by-step reasoning rather than factual recall	4
Control Flow Understanding	Conditional logic	Items assess students' ability to apply conditional decision rules (e.g., if–then relationships) when differentiating cacao disease types based on observable	4
Algorithm Design	Procedure constructio n	Items focus on students' ability to formulate a complete and coherent diagnostic procedure for cacao disease classification, including inputs, decision rules, and outputs	4
Representation of Algorithm	Symbolic representati on	Items measure students' capacity to represent diagnostic algorithms using structured formats such as flowcharts or pseudocode that clearly reflect logical decision paths.	4
Debugging and Error Detection	Error analysis	Items require students to identify inconsistencies or errors in existing diagnostic procedures and to explain how such errors affect classification outcomes	4
Algorithm Optimization	Refinement and efficiency	Items assess students' ability to refine diagnostic procedures by improving efficiency, accuracy, or clarity through modification of decision steps or criteria.	4
Generalization of Algorithmic Procedures	Transfer of reasoning	Items evaluate students' ability to apply established diagnostic procedures to new cacao varieties or altered conditions while maintaining algorithmic consistency	4

Data Analysis

Data on students' algorithmic thinking were examined through both summary-based and inferential statistical techniques. Measures of central tendency and score dispersion were calculated to describe performance patterns before and after the instructional intervention. Prior to inferential analysis, assumption testing was conducted to examine data normality and variance homogeneity. Considering the non-randomized grouping and observed differences in baseline performance, a covariance-based analytical framework was initially selected. Because the assumption checks reported in the Results indicated violations of normality and of posttest variance homogeneity, Welch's t-tests on gain scores were subsequently adopted as the primary inferential analysis, with the covariance-based analysis retained in a supporting role. Outcome scores obtained after instruction were treated as the response variable, group membership as the predictor, and baseline performance as a control factor. This analytical strategy was intended to provide a more precise estimation of instructional impact by accounting for initial disparities between groups; as reported in the Results, however, the pretest covariate ultimately explained little of the posttest variance. Normalized gain was computed at class level following Hake (1998). All computations were carried out using the jamovi statistical platform.

The AI-supported STEM learning design implemented in this study was grounded in supervised machine learning to externalize and structure students' algorithmic reasoning within an authentic scientific context. Google Teachable Machine was employed not merely as a technological platform, but as an instructional medium through which core principles of convolutional neural networks (CNNs) were operationalized into observable learning activities. The learning design emphasized students' algorithmic processes underlying image-based classification rather than the technical implementation of deep learning models.

Data collection constituted the initial stage of the AI-supported learning process. Students collected digital images of cacao leaves representing different disease conditions and healthy samples. This activity required students to apply systematic observation and classification criteria when selecting representative images, thereby introducing the concept of structured input data as a prerequisite for students' algorithmic processing. Through this stage, students learned that algorithmic outcomes depend on the quality, diversity, and relevance of input data.

After data collection, students labeled features by assigning disease categories to the collected cacao leaf images. Although feature extraction in CNNs occurs automatically through convolutional layers, the labeling process required students to explicitly identify salient visual characteristics, such as leaf color changes, lesion patterns, and spot distribution. This step functioned as a cognitive bridge between human perceptual reasoning and algorithmic representation, reinforcing the role of rule-based categorization in students' algorithmic thinking.

The training–testing cycle represented the core algorithmic phase of the learning design. Students trained the classification model using labeled image data and then tested it on unseen samples. Through iterative training and testing, students observed how changes in data composition and labeling affected model performance. This cycle engaged students in key algorithmic thinking processes, including sequencing, debugging, and optimization, as they refined inputs and evaluated outputs based on performance feedback.

Decision boundaries emerged implicitly through the model's classification behavior during the testing phase. Students analyzed misclassification cases to infer how the model differentiated between disease categories, thereby engaging in reflective reasoning about decision rules and conditional logic. This analysis required students to articulate why certain samples were classified correctly or incorrectly and how algorithmic decisions could be improved. In this way, decision boundaries served as a concrete representation of students' algorithmic reasoning, transforming abstract if–then logic into observable classification outcomes.

RESULT AND DISCUSSION

Algorithmic thinking in this study consisted of seven indicators: sequencing, control flow understanding, algorithm design, representation of algorithms, debugging and error detection, algorithm optimization, and generalization of algorithmic procedures. To foster these indicators, STEM activities were implemented using cacao plant disease classification materials supported by Google Teachable Machine. The learning activities were organized into sequential stages, beginning with arranging step-by-step procedures for identifying cacao leaf diseases, followed by designing decision branches for disease identification, constructing diagnostic algorithms, representing algorithms in flowcharts or pseudocode, testing and correcting prediction errors using the Teachable Machine model, improving model accuracy and efficiency, and finally testing algorithm generalization under new conditions and across different cacao varieties.

The content validity of the algorithmic thinking test instrument was evaluated through expert judgment by three independent validators with expertise in science education and educational assessment. The validators assessed 28 essay items across five dimensions: (1) relevance to the algorithmic thinking indicators, (2) accuracy of science concepts and the cocoa leaf disease context, (3) clarity of item construction, (4) language and readability, and (5) suitability of the scoring rubric. A five-point validation scale was employed, ranging from 1 (not valid) to 5 (very valid). The overall validation of Items 1–28 yielded a mean score of 4.57 and an Aiken's V coefficient of 0.892, indicating high content validity. Based on these results, the instrument was considered appropriate for measuring students' algorithmic thinking skills after minor revisions based on the validators' comments (Table 2, Panel A).

The internal consistency reliability of the algorithmic thinking test instrument was examined using Cronbach's Alpha separately for the experimental and control groups based on pretest responses. The analysis included 28 essay items completed by students in each group, and only the scores for Items 1–28 were included in the reliability analysis. Separate reliability analyses were performed because the two groups exhibited different baseline achievement levels, and estimating Cronbach's Alpha within each group provides a more accurate assessment of internal consistency by avoiding inflation of the reliability coefficient due to between-group variance. The results showed high internal consistency across the seven algorithmic thinking indicators and the overall instrument in both groups, indicating that the instrument was reliable for measuring students' algorithmic thinking skills before the intervention (Table 2, Panel B).

Table 2. Content validity and internal consistency of the algorithmic thinking test instrument

Panel A. Content validity (expert judgment of 28 items)

Validation Aspect	Mean Score	Feasibility (%)	Aiken's V	Category
Relevance to algorithmic thinking indicators	4.57	91.43%	0.893	Very valid
Accuracy of science concepts and cocoa leaf disease context	4.65	93.10%	0.914	Very valid
Clarity of item construction	4.51	90.24%	0.878	Very valid
Language and readability	4.67	93.33%	0.917	Very valid
Suitability of the scoring rubric	4.44	88.81%	0.860	Very valid
Overall	4.57	91.38%	0.892	Very valid

Panel B. Internal consistency (N = 66 students)					
Indicator	**Item Number**	**Number of Items**	**Experimental α (n=33)**	**Control α (n=33)**	**Category (Exp / Ctrl)**
Sequencing	1–4	4	0.838	0.765	High / High
Control flow	5–8	4	0.748	0.829	High / High
Algorithm design	9–12	4	0.935	0.851	Very high / High
Representation	13–16	4	0.685	0.838	Acceptable / High
Debugging	17–20	4	0.866	0.955	High / Very high
Optimization	21–24	4	0.958	0.950	Very high / Very high
Generalization	25–28	4	0.883	0.955	High / Very high
Total	1–28	28	0.942	0.938	Very high / Very high

The reliability analysis showed that the overall Cronbach's Alpha coefficient was 0.942 for the experimental group and 0.938 for the control group, indicating very high internal consistency in both groups. Across the seven algorithmic thinking indicators, Cronbach's Alpha values ranged from 0.685 to 0.958 in the experimental group and from 0.765 to 0.955 in the control group, demonstrating that all indicators exhibited acceptable to excellent reliability. These results indicate that the 28-item cognitive test instrument possessed strong internal consistency and was therefore considered reliable for measuring students' algorithmic thinking skills in the context of AI-integrated STEM learning.

To complement the quantitative findings, a process-oriented analysis was conducted to explain how the AI-integrated STEM activities may have contributed to the observed improvement in students' algorithmic thinking. Table 3 maps each algorithmic thinking indicator onto the corresponding learning task and reports the change actually observed on that indicator in both groups. Because each indicator was measured by four items scored from 0 to 3, indicator scores range from 0 to 12, and gains are expressed on that scale. This allows the process-based account to be tested rather than merely asserted. If the AI-supported activities operate through the mechanisms proposed, the indicators most directly exercised by those activities should show the largest between-group differences.

The experimental group gained more than the control group on every indicator, and all seven between-group differences were statistically

Table 3. Learning tasks and observed change in each algorithmic thinking indicator (maximum score per indicator = 12)

Algorithmic Thinking Indicator	**Student Task in the STEM-AI Activity**	**Control Gain M (SD)**	**Experimental Gain M (SD)**	**Welch t**	**Cohen's d**
Sequencing	Organizing step-by-step procedures for identifying cacao leaf diseases	1.03 (2.46)	3.55 (3.32)	3.50	0.86
Control flow understanding	Designing conditional (if–then) rules to differentiate disease categories	2.18 (3.30)	5.52 (3.69)	3.87	0.95
Algorithm design	Developing a complete classification procedure with multiple decision points	1.12 (3.13)	5.70 (5.25)	4.30	1.06
Representation of algorithms	Converting diagnostic procedures into flowcharts or pseudocode	2.58 (3.86)	7.97 (2.78)	6.52	1.60
Debugging and error detection	Identifying misclassified samples and analyzing causes of prediction errors	3.36 (4.44)	7.30 (4.07)	3.76	0.93

Algorithm optimization	Adjusting training parameters to improve model performance	3.48 (5.32)	8.61 (3.97)	4.43	1.09
Generalization of procedures	Testing the trained model on new images and different cacao varieties	5.18 (5.39)	8.48 (3.68)	2.91	0.72

Note: Gain = posttest “ pretest for each student; n = 33 per group. Welch’s t compares the gain scores of the two groups; all comparisons were statistically significant ($p < .01$). Cohen’s d is the standardized between-group difference in gain

significant. The effects were not uniform, however, and their ordering is informative. The largest difference occurred for representation of algorithms (d = 1.60), followed by algorithm optimization (d = 1.09) and algorithm design (d = 1.06). Control flow understanding (d = 0.95), debugging and error detection (d = 0.93), and sequencing (d = 0.86) formed an intermediate group, while generalization showed the smallest advantage (d = 0.72). The indicators that benefited most were therefore those requiring students to externalize a procedure in an explicit symbolic form and modify it deliberately, rather than those requiring them to recognize or repeat an ordered sequence.

Representation of algorithms also began at the lowest baseline in both groups (M = 1.73 of 12 in the experimental group and M = 3.36 in the control group), indicating that translating diagnostic reasoning into flowcharts or pseudocode was the competence least available to students before instruction. The control group remained comparatively weak on this indicator after instruction (M = 5.94 of 12, or 49% of the maximum), whereas the experimental group reached M = 9.70 (81%). This is the clearest point of divergence between the two conditions, and it is consistent with the requirement, specific to the AI-supported activities, that students commit their classification logic to an explicit representation before a model could be trained and evaluated. Conversely, the smallest difference for generalization arose because the control group also improved substantially on this indicator (gain = 5.18), suggesting that both instructional conditions supported transferring a procedure to new cases rather than AI in particular.The development of sequencing and control-flow understanding was supported during the initial stages of the AI-based classification activity. Students were required to organize the procedure for identifying cocoa leaf diseases, beginning with image collection and observation, followed by feature identification, labeling, model training, testing, and classification evaluation. This sequential structure required students to determine not only what action to perform, but also when to perform it. In addition, students formulated conditional decision rules based on observable characteristics of cocoa leaves, such as symptom patterns, discoloration, or lesion distribution. Thus, the AI classification task transformed the abstract notion of procedural order and if–then reasoning into a concrete classification workflow. The repeated use of this workflow may explain why students demonstrated more structured reasoning in tasks involving ordering procedures and applying conditional rules.

The algorithm design and algorithm representation indicators were developed when students transformed the disease classification problem into an explicit computational procedure. Students were required to identify the input data, define the classification process, establish decision points, and determine the expected output. Converting this procedure into flowcharts or pseudocode required students to externalize their reasoning and make relationships between successive steps and conditional branches visible. This activity was important because students did not merely operate an existing AI tool; they were required to conceptualize the logic underlying the classification process. Consequently, the AI-supported activity provided a concrete context in which students could construct and represent algorithmic procedures rather than simply memorize algorithmic concepts. This requirement is the most likely reason why representation of algorithms and algorithm design showed the two largest between-group differences reported in Table 3, since neither competence can be bypassed if a working classifier is to be produced.

Debugging and error detection clearly illustrate the mechanism, although the empirical advantage on this indicator was moderate rather than the largest observed (d = 0.93). During model testing, students encountered misclassified or incorrectly predicted cocoa leaf images. These prediction errors created an immediate feedback loop: students compared the model output with the expected disease category, identified possible sources of error, and then examined the image data, labels, or classification procedure. For example, when visually similar disease symptoms led to incorrect predictions, students had to reconsider whether the training images were sufficiently representative or whether the classification criteria needed refinement. This process transformed an incorrect AI output into an opportunity for systematic error analysis. Therefore, the improvement in algorithmic thinking can be explained not simply by students' use of AI, but by their repeated engagement in the cycle of prediction, error identification, cause analysis, procedural revision, and retesting. Such a cycle directly operationalizes debugging as a cognitive process.

Similarly, students developed algorithm optimization through iterative refinement of the classification model. Students compared model performance after modifying aspects of the training process, such as the composition and quality of the training data and the training configuration. Rather than accepting the first classification result as final, students were required to evaluate whether the model produced sufficiently accurate and efficient outcomes and to consider how to improve the procedure. This iterative process encouraged students to compare alternative solutions, evaluate the consequences of procedural changes, and select more effective strategies. The activity therefore provided an authentic context for understanding optimization as a process of systematic improvement rather than as a one-time correction.

The generalization indicator was supported when students tested the trained classification procedure using images that were not included in the original training examples, including images representing different conditions or cocoa varieties. This activity required students to determine whether the classification procedure remained applicable when the input conditions changed. When the model produced different results for unfamiliar images, students were encouraged to consider the limitations of the existing procedure and how far the classification rules could be transferred to new cases. This process developed algorithmic transfer because students had to evaluate whether a procedure constructed from known examples could be applied beyond the original training conditions.

Taken together, these learning processes provide a plausible explanation for the quantitative improvement observed in the experimental group. The AI-integrated STEM activities did not treat algorithmic thinking as an isolated theoretical construct. Instead, the seven indicators were embedded in a sequence of authentic scientific and technological activities: students organized procedures, constructed conditional rules, designed and represented algorithms, debugged prediction errors, optimized solutions, and tested the transferability of procedures to new cases. The AI system was particularly important because it provided immediate empirical feedback through classification outputs, enabling students to observe the consequences of their decisions and revise their procedures accordingly. In this sense, AI functioned as an interactive epistemic scaffold that connected students' algorithmic reasoning with observable model behavior. This process-oriented evidence helps explain why the experimental group achieved a higher posttest mean than the control group, and the ordering of effects across the seven indicators is broadly consistent with the mechanisms proposed. However, the correspondence is not exact: generalization, which the activities also targeted explicitly, showed the smallest between-group difference.

Prior to evaluating the effect of the AI-integrated STEM activities using cocoa disease classification materials on students' algorithmic thinking, the assumptions underlying the subsequent ANCOVA were examined. The Shapiro–Wilk test indicated that neither distribution met the normality assumption: $W = 0.945$, $p = 0.006$ for the pretest and $W = 0.903$, $p < 0.001$ for the posttest. The assumption was not satisfied for the ANCOVA residuals either ($W = 0.960$, $p = 0.032$). Within groups, the picture was more favorable, and the departure was confined to the experimental pretest scores ($W = 0.894$, $p = 0.004$), which were compressed toward the lower end of the scale; the remaining three distributions did not depart significantly from normality ($W = 0.955$ to 0.967, $p = 0.189$ to 0.399). Levene's test indicated that the pretest variances were homogeneous, $F(1,64) = 0.09$, $p = 0.766$, whereas the posttest variances were markedly heterogeneous, $F(1,64) = 15.20$, $p < 0.001$. The source of this heterogeneity is visible in the descriptive data. Posttest scores in the experimental group clustered near the upper limit of the instrument, averaging 91.3% of the maximum attainable score of 84 and ranging from 71 to 83, whereas the control group averaged 79.3% and ranged from 42 to 81. The experimental distribution was therefore compressed against the test ceiling, reducing its posttest variance to approximately one-seventh of that observed in the control group. Two of the three assumptions required for ANCOVA were therefore violated, and the third, independence of the covariate from the treatment, was also compromised by the significant baseline difference between groups reported below. For this reason, the analytic strategy was reordered: Welch's t-tests on gain scores, which require neither variance homogeneity nor an effective covariate, were treated as the primary inferential evidence, and the ANCOVA was retained as a supporting analysis whose results are reported for completeness and interpreted with corresponding caution.

To complement the inferential analysis, the distribution of students' algorithmic thinking scores was visualized using paired boxplots for the pretest and posttest assessments in both the experimental and control groups (Figure 4). The boxplots display the median, interquartile range (IQR), whiskers, and outliers, allowing direct comparison of score distributions before and after the intervention.

As shown in Figure 4, the pretest distributions indicate a noticeable baseline imbalance between the two groups. The experimental group was concentrated at lower pretest scores, with a median of approximately 23, whereas the control group had a substantially higher median of approximately 47. The experimental group also exhibited a wider spread of low scores, while the control group showed a broader distribution across the middle score range. The posttest boxplots reveal a markedly different pattern. The experimental group shifted upward to a median of approximately 77 and displayed a relatively narrow interquartile range, indicating that most students achieved consistently high scores. This compression of scores into a narrow upper range suggests a ceiling-effect tendency in the experimental group. In contrast, the control group also improved, with a median

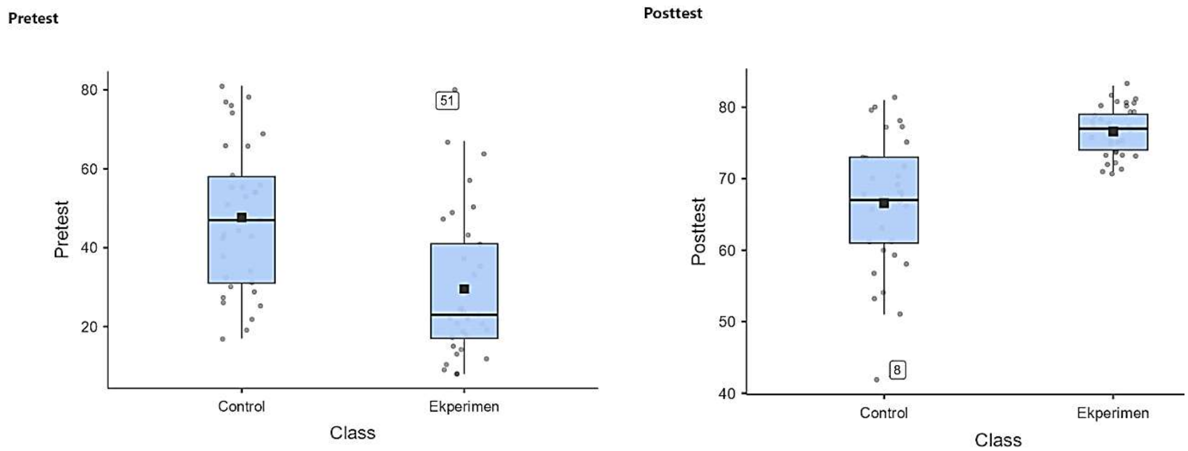

Figure 4. Distribution of students' algorithmic thinking scores in the pretest and posttest by group

of approximately 67, but retained a wider interquartile range and greater overall variability. The control group additionally showed a lower outlier, indicating that some students remained substantially below the main distribution after instruction. Taken together, the boxplots show that although both groups improved from pretest to posttest, the experimental group achieved higher posttest scores with considerably lower variability. The figure also makes the initial baseline imbalance visible. It therefore serves as a descriptive visualization of score distributions that complements the inferential analyses reported below rather than constituting evidence of an instructional effect in its own right.

Because the posttest scores violated both the normality and the homogeneity-of-variance assumptions, Welch's t-tests, which do not require equal variances, were used as the primary inferential analysis. These were conducted on pretest scores and on gain scores.

The Welch's t-test indicated a statistically significant difference in pretest scores between the control and experimental groups, $t(64) = 3.93$, $p < .001$, Cohen's $d = 0.968$. The control group obtained a higher mean pretest score ($M = 47.6$, $SD = 18.8$) than the experimental group ($M = 29.5$, $SD = 18.6$), indicating a substantial baseline imbalance between the two groups. The analysis of gain scores also revealed a statistically significant difference between groups, $t(64) =$ "5.81, $p < .001$, Cohen's $d = 1.43$ in favor of the experimental group. The experimental group demonstrated a substantially greater mean gain ($M = 47.1$, $SD = 19.8$) than the control group ($M = 18.9$, $SD = 19.6$). This result provides descriptive evidence of greater improvement in algorithmic thinking among students who participated in the AI-integrated STEM activities. Because the groups differed significantly at baseline, a covariance-adjusted analysis was subsequently conducted to account for initial pretest differences when comparing posttest performance. Before interpreting the adjusted group effect, we examined the homogeneity-of-regression-slopes assumption by including the interaction between class and pretest scores in the model. The interaction effect was not statistically significant, $F(1, 62) = 1.739$, $p = .192$, indicating that the relationship between pretest and posttest scores did not differ significantly between the two groups. Therefore, the ANCOVA model without the interaction term was considered appropriate for estimating the adjusted group effect.

As presented in Table 4, the ANCOVA indicated a statistically significant effect of class on posttest algorithmic thinking scores, $F(1, 63) = 29.302$, $p < .001$. The pretest covariate, however, was not statistically significant, $F(1, 63)$

Table 4. Ancova - posttest

	Sum of Squares	df	Mean Square	F	P	η^2	$\eta^2 p$	ω^2
Overall model	1683.34	2	841.67	17.4645	< 0.001			
Class	1412.18	1	1412.18	29.3023	<0.001	0.299	0.317	0.286
Pretest	3.21	1	3.21	0.0666	0.797	0.001	0.001	-0.009
Residuals	3036.19	63	48.19					

= 0.067, p = .797, partial ç² = 0.001. The pooled within-group correlation between pretest and posttest scores was r = .03, and the corresponding within-group coefficients were r = .15 for the control group and r = ".27 for the experimental group. Two consequences follow and are reported here explicitly. First, because the covariate accounted for almost none of the variance in posttest scores, the covariance adjustment was numerically negligible: the adjusted means (66.5 and 76.8) are effectively identical to the unadjusted posttest means (66.6 and 76.7). The ANCOVA therefore did not materially correct for the baseline imbalance between the two groups, and the group effect reported above is best understood as an essentially unadjusted comparison of posttest performance. Second, a near-zero association between two administrations of the same instrument is itself unexpected, since scores on a stable measure would ordinarily be expected to correlate across occasions; the principal reason is a ceiling effect in the posttest scores of the experimental group, whose values fell within a narrow band close to the maximum of 84 (range 71–83, SD = 3.42) compared with the far wider spread of the control group (range 42–81, SD = 9.13). At the indicator level, the same compression is evident: experimental posttest means reached 97% of the maximum for sequencing and 95% for control flow understanding and generalization. Because almost all students in the experimental group converged on similar posttest scores regardless of their starting point, there was little residual variance for the pretest to explain. This same disparity in dispersion, a posttest variance approximately 7.1 times larger in the control group, is also the source of the significant Levene's test reported earlier. Because the posttest homogeneity-of-variance assumption was also violated, the ANCOVA results were interpreted alongside complementary analyses. The Welch's independent-samples t-test showed a statistically significant difference in gain scores between the control group (M = 18.9) and the experimental group (M = 47.1), t(64) = "5.81, p < .001, with a very large effect size (d = 1.43). This pattern was also reflected in the analysis of normalized gain (Hake, 1998), computed at class level as ⟨g⟩ = (mean posttest " mean pretest) / (maximum score " mean pretest), with the maximum attainable score being 84 (28 items scored 0–3). The experimental group achieved ⟨g⟩= 0.87, which falls in the high category, whereas the control group achieved ⟨g⟩= 0.52, which falls in the medium category. The class-level form of the index was used rather than the mean of individual normalized gains because the latter is unstable when a student's pretest score is close to the maximum: the denominator becomes very small, so a decline of a few points produces an extreme negative value. Seven students in the control group scored lower on the posttest than on the pretest, and three of them had high pretest scores, which distorted the mean of the individual index without affecting the class-level estimate. Taken together, the higher normalized gain and absolute gain observed in the experimental group, together with the significant covariance-adjusted group difference, indicate a consistent pattern of greater improvement in algorithmic thinking among students who participated in the AI-integrated STEM activities. Nevertheless, these findings should be interpreted cautiously because the posttest homogeneity-of-variance assumption was violated.

These findings are consistent with the process-oriented learning activities implemented in the intervention. Through the AI-supported classification tasks, students repeatedly organized procedures, constructed decision rules, evaluated model outputs, identified classification errors, and refined their solutions based on feedback. This iterative interaction with AI-generated outputs gave students opportunities to externalize and revise their reasoning, particularly in sequencing, control flow, debugging, and algorithm optimization. Thus, the observed improvement in algorithmic thinking can be understood not only as a difference in final test scores but also as the result of sustained engagement with structured, iterative, and evidence-based algorithmic processes during the STEM learning activities (Otto *et al*., 2025).

The results indicate that the two instructional conditions were associated with substantially different posttest outcomes. Given the negligible covariance adjustment reported above, however, this difference should not be described as one that persists after controlling for initial ability, since no effective statistical control was in fact achieved. The analysis does establish that students who participated in the AI-integrated STEM activities improved substantially more than their peers, and that this pattern is consistent across three indices: absolute gain, normalized gain, and posttest standing. It cannot establish that the improvement is attributable to the intervention alone. In relation to the third research question, the findings therefore support the conclusion that AI-integrated STEM activities using cocoa disease classification were accompanied by markedly greater improvement in undergraduate students' algorithmic thinking, while stopping short of a causal claim.

After establishing a significant main effect of group in the ANCOVA analysis, the adjusted posttest means were examined to clarify the direction and magnitude of the group difference. In principle, the adjusted means represent posttest students' algorithmic thinking scores corrected for initial differences between groups; in this dataset, however, the correction was negligible because the pretest covariate explained almost none of the posttest variance, so the values below are effectively unadjusted group means (Table 5).

Table 5. Adjusted means of algorithmic thinking

Group	Mean	SE	95% Confidence Interval	
			Lower	Upper
Control Group	66.5	1.21	64.1	68.9
Experimental Group	76.8	1.21	74.4	79.2

The estimated post-intervention mean for the intervention group exceeded that of the comparison group. Students exposed to AI-enhanced STEM learning therefore achieved stronger algorithmic thinking outcomes than peers who experienced contextually oriented instruction without artificial intelligence. The absence of overlap between the confidence intervals indicates a distinct and practically meaningful separation in post-intervention performance between the two groups. This separation should nonetheless be read as a difference in outcome rather than as an isolated estimate of the instructional effect, because the covariance adjustment did not materially offset the baseline difference between the classes.

To provide a visual comparison of the adjusted posttest performance, the estimated marginal means of students' algorithmic thinking scores were plotted for the control and experimental groups. The adjusted means nominally represent posttest performance adjusted for differences in students' pretest scores, although this adjustment was negligible in the present data. Error bars represent 95% confidence intervals (Figure 5).

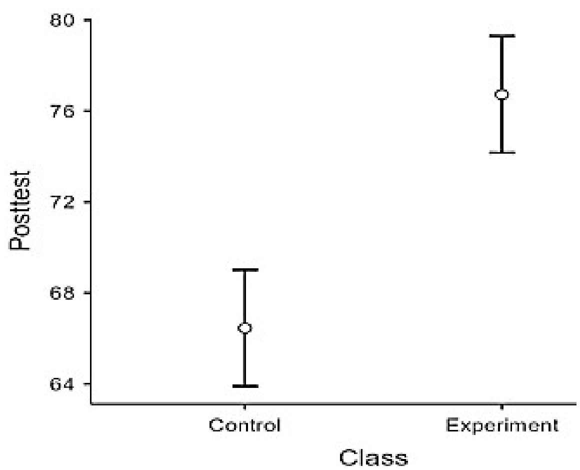

Figure 5. Estimated marginal means of posttest algorithmic thinking scores by instructional group

The figure shows a clear difference in adjusted posttest performance between the two groups. The experimental group exhibited a higher estimated marginal mean than the control group. This visual pattern is consistent with the significant group effect identified in the ANCOVA, bearing in mind that the covariance adjustment underlying these means was negligible in the present data. However, interpret the adjusted means alongside the assumption tests, Welch's t-test results, gain-score analysis, and individual score distributions because the posttest homogeneity-of-variance assumption was not satisfied.

The findings demonstrate that AI-integrated STEM activities based on cocoa disease classification were associated with substantially greater improvement in students' algorithmic thinking than the comparison learning condition. The experimental group achieved significantly higher posttest scores and a substantially larger gain, and the covariance-adjusted analysis showed a significant between-group effect, although that adjustment was numerically negligible. This finding is consistent with recent evidence that STEM-integrated learning can support the development and transfer of computational forms of reasoning through interdisciplinary and problem-oriented learning experiences (Li & Oon, 2024). The large gain-score effect observed in the present study further suggests that the intervention produced a meaningful educational effect rather than a trivial difference between groups.

The observed improvement can be interpreted as an outcome of learning experiences that actively engaged students in structured and procedural reasoning. STEM activities based on cacao disease contexts required learners to systematically analyze symptoms, organize observations into ordered steps, and apply explicit decision rules to reach conclusions. These cognitive operations align closely with the fundamental dimensions of algorithmic thinking, including problem decomposition, sequencing, rule-based decision-making, and cycles of evaluative refinement (Kanaki & Kalogiannakis, 2022; Wong *et al*., 2024). Similar patterns have been reported in recent studies showing that STEM learning environments emphasizing procedural and rule-based reasoning significantly enhance students' algorithmic thinking skills (Tang *et al*., 2020; Kong & Lai, 2022; Cheng *et al*., 2023).

The negligible contribution of pretest performance to the posttest outcome requires careful interpretation and, in this study, constitutes a limitation rather than a strength. A non-significant covariate does not demonstrate that differences in initial ability have been neutralized; it indicates the opposite, namely that the statistical mechanism intended to adjust for those differences was ineffective in this sample. Nor should it be read as evidence that prior knowledge is generally irrelevant to learning outcomes. This limitation is compounded by the direction of the baseline imbalance. The control group began with a substantially higher pretest mean (M = 47.6) than the experimental group (M = 29.5), yet finished with a lower posttest mean. This is precisely the pattern that regression to the mean would be expected to produce even in the absence of a treatment effect, because groups that are selected or observed at the extremes of a distribution tend to move toward the overall mean when they are measured again.

The comparison of non-equivalent groups on pretest-adjusted outcomes is also subject to Lord's Paradox, in which gain-score and covariance-adjusted analyses can support divergent conclusions without either being demonstrably correct. The convergence between the gain-score, normalized-gain, and covariance-adjusted results reported above should therefore not be treated as three independent confirmations of the same conclusion, since all three rest on the same two intact classes and are susceptible to the same source of bias. At the same time, the magnitude of the observed difference (d = 1.43) is considerably larger than regression artifacts alone would ordinarily generate, which makes a genuine instructional contribution plausible. The most defensible reading of the evidence is thus an intermediate one: participation in the AI-integrated STEM activities was accompanied by markedly greater improvement in algorithmic thinking, and this pattern is consistent with research emphasising structured learning environments and explicit opportunities for computational and algorithmic reasoning (Li & Oon, 2024; Wong et al., 2024), but the present design cannot rule out regression to the mean, pre-existing differences between the intact classes, or the novelty of the technology as partial explanations.

An important insight offered by this research is a clearer understanding of how artificial intelligence facilitates the cultivation of algorithmic thinking. By engaging students in machine learning–driven classification activities using Google Teachable Machine, algorithmic reasoning was made explicit through concrete decision rules, allowing abstract computational processes to be transformed into observable and empirically examinable procedures. Recent studies have highlighted that AI-supported learning environments facilitate this externalization process by making learners' reasoning visible through data-driven feedback and iterative model refinement (Zhai *et al*., 2021; Chen *et al*., 2023). Park et al. (2023) reported similar conclusions, highlighting that incorporating artificial intelligence into science instruction supports systematic reasoning by promoting data-informed decision processes. By interacting with a convolutional neural network (CNN)-based image classification system, students were required to examine feature representations, interpret classification outcomes, identify prediction errors, and refine decision criteria through repeated training–testing cycles. This interaction aligns with recent findings indicating that CNN-based visual classification tasks can support algorithmic reasoning by linking perceptual features with structured decision-making processes (Li *et al*., 2020).

Beyond the conceptual contribution of CNN-driven classification, the instructional process was further enhanced by learners' engagement with core training parameters provided within Google Teachable Machine, including epochs, batch size, learning rate, and model accuracy. By adjusting the number of epochs, students observed how repeated training cycles influenced model convergence and performance stability, reinforcing their algorithmic understanding of iteration. This is consistent with Chen *et al*. (2023), who found that engagement with machine learning model evaluation promotes deeper algorithmic reasoning. Variations in batch size allowed students to experience trade-offs between training speed and prediction consistency. At the same time, changes in learning rate provided concrete insights into optimization processes, including underfitting and overfitting behavior. Model accuracy functioned as a transparent feedback indicator, allowing learners to assess the quality of their algorithmic choices and to iteratively improve their classification strategies. Through engagement with these parameters, abstract notions of optimization, evaluation, and refinement in students' algorithmic thinking became observable and measurable processes, supporting students' understanding of how algorithmic performance is shaped by systematic parameter tuning rather than trial and error alone.From a pedagogical perspective, AI in this learning design did not function merely as a technological tool but as an epistemic scaffold that supported the construction, evaluation, and optimization of students' algorithmic thinking. Studies over the last five years have emphasized that AI-enhanced STEM learning environments promote deeper reasoning when students actively engage in data labeling, model evaluation, and error analysis, as these activities require explicit articulation of algorithmic logic and conditional reasoning (Zhai *et al*., 2021). This explains why students exposed to AI-integrated STEM activities demonstrated stronger students' algorithmic thinking outcomes than those engaged in contextual learning without AI support.

Previous theoretical frameworks further support these findings. Algorithmic thinking is characterized by the ability to formulate executable procedures and refine them through iterative evaluation (Kanaki *et al*., 2025). Recent empirical work has shown that learning designs emphasizing iteration, debugging, and optimization are particularly effective in strengthening this form of reasoning. Within science education, situating these processes in biologically meaningful settings has been associated with heightened learner engagement and deeper cognitive processing, particularly when students work with real-world classification and diagnostic challenges (Le et al., 2023; Almasri, 2024).

CONCLUSION

This study addressed three research questions concerning the validity, reliability, and effectiveness of an AI-integrated STEM learning intervention for developing undergraduate students' algorithmic thinking. First, the algorithmic thinking test demonstrated high content validity, with an overall mean validation score of 4.57, a feasibility percentage of 91.38%, and an Aiken's V coefficient of 0.892, indicating strong agreement among the three independent expert validators on the instrument's relevance and quality. Second, the instrument demonstrated very high internal consistency, with Cronbach's alpha coefficients of 0.942 for the experimental group and 0.938 for the control group across the seven targeted algorithmic thinking indicators, indicating that the instrument was reliable for measuring students' algorithmic thinking skills prior to the intervention. Third, the findings provide evidence of a consistent pattern of greater improvement in algorithmic thinking among students who participated in the AI-integrated STEM activities based on cocoa leaf disease classification. The experimental group demonstrated a higher class-level normalized gain (0g 0 = 0.87, high category) than the control group (0g 0 = 0.52, medium category), and the gain-score comparison indicated a very large between-group effect (d = 1.43). The advantage was observed on all seven indicators, with between-group effect sizes for gain ranging from d = 0.72 for generalization to d = 1.60 for representation of algorithms. The covariance-adjusted analysis showed a significant group difference in posttest performance. However, the pretest covariate explained virtually none of the posttest variance, so that this analysis functioned in practice as an essentially unadjusted comparison. Together, these findings suggest that integrating AI-supported image classification into STEM learning can contribute to the development of students' algorithmic thinking by engaging them in sequencing procedures, applying conditional logic, designing and representing algorithms, debugging errors, optimizing solutions, and generalizing procedures to new cases. However, the findings should be interpreted with appropriate caution, and four limitations constrain the strength of these conclusions. First, the normality and homogeneity-of-variance assumptions were both violated; the latter was attributable to a ceiling effect: experimental posttest scores averaged 91.3% of the maximum attainable score, which compressed their distribution and meant the instrument could not register further growth in that group. Second, the two intact classes differed substantially at baseline, and the direction of that difference coincides with the observed effect, so regression to the mean cannot be excluded as a partial explanation. Third, the pretest covariate did not function as an effective statistical control, which left the study without a working mechanism for adjusting the baseline imbalance. Fourth, although three independent experts corroborated the instrument's content validity, inter-rater reliability for scoring the open-ended responses was not established, so scoring objectivity remains a limitation. Taken together, these limitations mean that the results are best understood as evidence that participation in the AI-integrated STEM activities was accompanied by substantially greater improvement in algorithmic thinking, rather than as a demonstration that the intervention caused that improvement. The non-randomized design and variance heterogeneity limit the extent to which causal conclusions can be generalized. Future research should replicate the intervention with larger, more diverse samples; randomized or more closely matched groups; multiple expert reviewers and independent scorers; a delayed posttest to examine retention; an instrument with sufficient headroom to avoid ceiling effects; and statistical approaches specifically designed to address heteroscedasticity.

■ DECLARATION OF GENERATIE AI USAGE IN THE WRITING PROCESS

During the drafting of this manuscript, the authors used Grammarly and ChatGPT to refine sentence structure and translate text. After using these tools, the authors reviewed and revised the content as needed and accept full responsibility for the final content of the article.

■ DATA AVAILABILITY STATEMENT

The datasets generated and analyzed during this study, including the raw pretest and posttest scores, the 28-item algorithmic thinking test, the expert validation sheet, and the scoring rubric, are not publicly archived because they contain information relating to student participants and are subject to institutional confidentiality requirements. The data are available from the corresponding author upon reasonable request, provided that the intended use is consistent with the conditions of the ethical approval granted for this study and that participant anonymity is preserved. The complete dataset and supporting research instruments are also available at the following Google Drive link: https://drive.google.com/drive/folders/1eIn0G7U2Jj-dm1EwaXibY3Zk2HvBUTiG?usp=drive_link.

■ ACKNOWLEDGMENT

We gratefully acknowledge the support provided by the PMDSU Scholarship (Program

Magister Menuju Doktor untuk Sarjana Unggul) year 2026 (Project No. 308/C3/DT.05.00/PL-PMDSU2026), initiated by the Directorate of Educators and Education Personnel, Directorate General of Higher Education (Ditjen Dikti), Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia. This support has been essential for my study and for the successful completion of this research.

REFERENCES

Adorni, G., Piatti, S., & Karpenko, V. (2024). Virtual CAT: A multi-interface educational platform for algorithmic thinking assessment. SoftwareX, 27, 101737. doi:10.1016/j.softx.2024.101737

Almasri, F. (2024). Exploring the impact of artificial intelligence in teaching and learning of science: A systematic review of empirical research. Research in Science Education, 54(5), 977–997. doi: 10.1007/s11165-024-10176-3

Astawan, I. G., Suarjana, I. M., Werang, B. R., Asaloei, S. I., Sianturi, M., & Elele, E. C. (2023). Stem-based scientific learning and its impact on students' critical and creative thinking skills: an empirical study. Jurnal Pendidikan IPA Indonesia, 12(3), 482–492. doi:10.15294/jpii.v12i3.46882

Chen, P., Yang, D., Metwally, A. H. S., Lavonen, J., & Wang, X. (2023). Fostering computational thinking through unplugged activities: A systematic literature review and meta-analysis. International Journal of STEM Education, 10(1), 47. doi:10.1186/s40594-023-00434-7

Cheng, L., Wang, X., & Ritzhaupt, A. D. (2023). The effects of computational thinking integration in STEM on students' learning performance in K-12 education: A meta-analysis. Journal of Educational Computing Research, 61(2), 416–443. doi:10.1177/07356331221114183

Dafik, Kadir, Maryati, T. K., Sufirman, & Ridlo, Z. R. (2023). The analysis of the implementation of RBL-STEM learning materials in improving student's meta-literacy ability to solve wallpaper decoration problems using local antimagic graph coloring techniques. Heliyon, 9(6), e17433. doi:10.1016/j.heliyon.2023.e17433

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145, 311–318. doi:10.1016/j.compag.2018.01.009

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics, 66(1), 64–74. doi:10.1119/1.18809

Kanaki, K., & Kalogiannakis, M. (2022). Assessing algorithmic thinking skills in relation to age in early childhood STEM education. Education Sciences, 12(6), 380. doi: 10.3390/educsci12060380

Kanaki, K., Chatzakis, S., & Kalogiannakis, M. (2025). Fostering algorithmic thinking and environmental awareness via Bee-Bot activities in early childhood education. Sustainability, 17(9), 4208. doi:10.3390/su17094208

Kong, S. C., & Lai, M. (2022). Validating a computational thinking concepts test for primary education using item response theory: An analysis of students' responses. Computers & Education, 187, 104562. doi:10.1016/j.compedu.2022.104562

Le, H. C., Nguyen, V. H., & Nguyen, T. L. (2023). Integrated STEM approaches and associated outcomes of K-12 student learning: A systematic review. Education Sciences, 13(3), 297. doi: 10.3390/educsci13030297

Lehmann, T. H. (2024). How current perspectives on algorithmic thinking can be applied to students' engagement in algorithmatizing tasks. Mathematics Education Research Journal, 36(3), 609–643. doi: 10.1007/s13394-023-00462-0

Li, Y., Schoenfeld, A. H., DiSessa, A. A., Graesser, A. C., Benson, L. C., English, L. D., & Duschl, R. A. (2020). On computational thinking and STEM education. Journal for STEM Education Research, 3, 147–166. doi:10.1007/s41979-020-00044-w

Li, Z., & Oon, P. T. (2024). The transfer effect of computational thinking (CT)-STEM: A systematic literature review and meta-analysis. International Journal of STEM Education, 11, 44. doi:10.1186/s40594-024-00498-z

Ministry of Primary and Secondary Education. (2025). *Naskah akademik pembelajaran koding dan kecerdasan artifisial pada pendidikan dasar dan menengah. Badan Standar, Kurikulum, dan Asesmen Pendidikan*. Kementerian Pendidikan Dasar dan Menengah Republik Indonesia. https://kurikulum.kemendikdasmen.go.id/service/download.php?kategori=rujukan&id=117

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419. doi:10.3389/fpls.2016.01419

Ogegbo, A. A., & Ramnarain, U. (2022). A systematic review of computational thinking in science classrooms. Studies in Science Education, 58(2), 203–230. doi:10.1080/03057267.2021.1963580

Otto, S., Lavi, R., & Bertel, L. B. (2025). Human–GenAI interaction for active learning in STEM education: State-of-the-art and future directions. Computers & Education, 239, 105444. doi:10.1016/j.compedu.2025.105444

Park, J., Teo, T., Teo, A., Chang, J., Huang, J., & Koo, S. (2023). Integrating artificial intelligence into science lessons: Teachers' experiences and views. International Journal of STEM Education, 10, 61. doi:10.1186/s40594-023-00454-3

Ramadhani, R. A., Rifin, A., & Novianti, T. (2023). *Analisis daya saing dan kebijakan bea keluar pada komoditas kakao (Theobroma cacao) Indonesia* [Analysis of competitiveness and export duty policy on Indonesian cocoa (Theobroma cacao) commodities]. *Analisis Kebijakan Pertanian, 21*(2):171–186. doi: 10.21082/akp.v21i2.171-186

Ridlo, Z. R., Dafik, Ningsih, S. P. A., & Anggraini, A. L. (2025). Computer vision on education: Fostering AI literacy using RBL-STEM with Google Teachable Machine. Jurnal Penelitian & Pengembangan Pendidikan Fisika, 11(2), 197–210. doi:10.21009/1.11205

Ridlo, Z. R., Ningsih, S. P. A., & Anggraini, A. L. (2025). Computational thinking and deep learning on science education framework: A systematic review. Science Education International, 36(4), 480–489. doi:10.33828/sei.v36.i4.11

Rusmin, L., Misrahayu, Y., Pongpalilu, F., Radiansyah, R., & Dwiyanto, D. (2024). Critical thinking and problem-solving skills in the 21st century. Join: Journal of Social Science, 1(5), 144–162. doi: 10.59613/svhy3576

Sari, N. M., Munif, A., & Purwantara, A. (2025). Management of cocoa plant pests and diseases in East Luwu Regency. Indonesian Journal of Agricultural Sciences/Jurnal Ilmu Pertanian Indonesia, 30(1). 10.18343/jipi.30.1.177

Sun, C., Yang, S., & Becker, B. (2024). Debugging in computational thinking: A meta-analysis on the effects of interventions on debugging skills. Journal of Educational Computing Research, 62(4), 867–901. doi:10.1177/07356331241227793

Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. Computers & Education, 148, 103798. doi:10.1016/j.compedu.2019.103798

Thibaut, L., Ceuppens, S., De Loof, H., De Meester, J., Goovaerts, L., Struyf, A., Pauw J. B.D., Dehaene, W., Deprez, J., Cock, M. D., Hellinckx, L., Knipprath, H., Langie, G., Struyven, K., Velde, D. V.D., Petegem, P. V., & Depaepe, F. (2018). Integrated STEM education: A systematic review of instructional practices in secondary education. European Journal of STEM Education, 3(1), 2. doi:10.20897/ejsteme/85525

Thifany, A. J., Santosa, E., & Khumaida, N. (2020). *Faktor-faktor yang memengaruhi produksi dan efektivitas panen pada kakao mulia* [Factors that influence production and harvest effectiveness of noble cocoa]. *Jurnal Agronomi Indonesia (Indonesian Journal of Agronomy), 48*(2), 187–195. doi:10.24831/jai.v48i2.30565

Tupouniua, J. G. (2023). What challenges emerge when students engage with algorithmatizing tasks?. Journal of Pedagogical Research, 7(2), 93–107. doi: 10.33902/JPR.2023 18518

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

Wong, G. K., Jian, S., & Cheung, H. Y. (2024). Engaging children in developing algorithmic thinking and debugging skills in primary schools: A mixed-methods multiple case study. Education and Information Technologies, 29(13), 16205–16254. doi: 10.1007/s10639-024-12448-x

Zengin, E., & Karal, Y. (2024). Developing a game-based test to assess middle school sixth-grade students' algorithmic thinking skills. International Journal of Assessment Tools in Education, 11(1), 88–108. doi: 10.21449/ijate.1327082

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J. B., Yuan, J., & Li, Y. (2021). A review of artificial intelligence in education from 2010 to 2020. Complexity, 2021, 8812542. doi:10.1155/2021/8812542