# Machine Learning in Art Teacher Education: A Comparative Analysis and Student Perceptions

**Botagoz Kystaubayeva[1], Gulmira Mailybaeva[1], Kairat Dzhanabaev[2], Ainur Ansabayeva[1], Elmira Kydyrbekova[3] and Aivar Sakhipov[4]**

[1]Department of Teaching and Upbringing Methods, Zhetysu university named after I. Zhansugurov, Taldykorgan, Kazakhstan
[2]Department of Physical and Basic Military Training, Zhetysu university named after I. Zhansugurov, Taldykorgan, Kazakhstan
[3]School of Arts and Humanities, Astana International University, Kazakhstan
[4]School of Software Engineering, Astana IT University, Kazakhstan

b.kystaubayeva@zu.edu.kz
g.mailybayeva@zu.edu.kz
dzanabaevkajrat072@gmail.com
ainuransabayeva@gmail.com (corresponding author)
e20250707@gmail.com
aivar.sakhipov@astanait.edu.kz

https://doi.org/10.34190/ejel.24.2.4313



**Abstract**: Amid the global push for digital transformation in higher education, there is a critical need for objective, scalable assessment tools in subjective disciplines like visual arts. Modern teacher education increasingly integrates intelligent technologies, yet the application of machine learning (ML) for formative assessment in art education remains underexplored. While ML offers scalable feedback, its capacity to evaluate subjective creativity remains contested. The study aims to examine the technical accuracy of a CNN-based model trained on a local dataset of 300 archived projects, compared to instructor evaluations, and to analyze how future teachers (N = 180) perceive algorithmic feedback in assessment contexts. A mixed-methods design was employed using a highly reliable survey instrument (Cronbach's $\alpha$ = .925) and comparative scoring analysis across four key dimensions: Technique, Composition, Color, and Creativity. Results indicate that the model aligns strongly with human assessments on technical execution ($r = .426$, $p < .001$), and moderate alignment for Composition ($r = .430$, $p < .001$) and weaker alignment for Color ($r = .327$, $p < .001$), while correlations for Creativity were notably weaker ($r = .181$, $p = .015$), indicating persistent limitations in modeling abstract artistic intent. ANOVA results revealed that students' digital literacy significantly predicts their trust in the system ($F = 3.547$, $p = .031$) and willingness to use it ($F = 8.476$, $p < .001$). Furthermore, discrepancy analysis indicated systematic divergence across proficiency levels, with the model exhibiting increasing underestimation for highly proficient students, particularly in cases involving stylistic deviation or non-standard cultural expression. The findings suggest that while the algorithm provides consistent, transparent scoring that enhances assessment literacy, it lacks the sensitivity to evaluate high-level originality due to standardization bias. This study contributes to the field by empirically demonstrating the "accuracy–creativity trade-off" in ML-based art assessment and by validating a hybrid assessment framework that balances algorithmic precision with pedagogical intuition. The study concludes that ML tools should function as "human-in-the-loop" support systems rather than autonomous graders, fostering critical reflection and digital competence in future educators.

**Keywords**: Pre-service primary teachers, Visual arts education, Teaching practices, Artificial intelligence, Machine learning

## 1. Introduction

Modern teacher education is undergoing rapid digital transformation, requiring a reconsideration of both the content of teacher preparation and the methods used to assess students' learning outcomes (Cai, 2025; Bedir and Freedman, 2024). As digital technologies become integral to contemporary pedagogical practice, intelligent systems are increasingly explored as tools for supporting assessment and formative feedback (Buckingham Shum et al., 2023). However, while machine learning (ML) approaches are widely applied in structured academic domains, their use in evaluating students' visual and creative work remains limited (Patterson et al., 2024). This gap is particularly pronounced in pre-service teacher education, where creative assignments are central to professional formation, yet assessment practices continue to rely heavily on subjective instructor judgment (Buckingham Shum et al., 2023; Patterson et al., 2024).

The assessment of visual and creative work is challenging due to the difficulty of formalizing evaluative criteria such as originality, composition, color harmony, and expressive intent. This subjectivity complicates consistency



Cite this article: Kystaubayeva, B. et al. 2026. "Machine Learning in Art Teacher Education: A Comparative Analysis and Student Perceptions", *Electronic Journal of e-Learning*, 24(2), pp 01-16, https://doi.org/10.34190/ejel.24.2.4313

and transparency in evaluation, constrains opportunities for structured student self-reflection, and increases instructors' assessment workload. Recent advances in computer vision and machine learning suggest that algorithmically supported approaches may contribute to formative assessment by providing consistent, rubric-aligned feedback on selected visual features. Within a human-in-the-loop framework, such outputs are designed to support professional pedagogical judgment rather than to replace it (Zheng et al., 2025; Patterson et al., 2024).

In this study, algorithmic assessment refers to the generation of criterion-based formative feedback through machine learning models that analyze predefined visual and semantic features of student work. It does not involve automated summative grading or the delegation of evaluative authority to algorithms. Engagement with such feedback is closely related to digital pedagogical competence, understood here as the ability of future teachers to critically engage with digital tools, reflect on their pedagogical affordances and limitations, and apply them responsibly within educational practice.

The present study explores the potential of algorithmically supported formative assessment in art-oriented teacher education. Specifically, it investigates the use of a machine learning model for evaluating digital art projects produced by pre-service primary school teachers. The scope of the study is deliberately limited to undergraduate teacher education and to the analysis of visual characteristics of digital artworks alongside keyword-based thematic features derived from project descriptions. Other modalities, broader learning analytics, and automated summative grading fall outside the focus of the research.

The study was conducted within the bachelor-level academic program 6B01301 "Pedagogy and Methods of Primary Education," implemented at Zhetysu University named after I. Zhansugurov and Abai Kazakh National Pedagogical University in Kazakhstan. The program is delivered within faculties responsible for teacher education and focuses on pedagogical reflection, diagnostic competence, and the integration of digital educational technologies. The analyzed data consist of final digital art projects completed within a course designed to foster students' creative, digital, and pedagogical competencies. Before outlining the theoretical framework and methodology, the main contributions of this study are summarized as follows:

- The study demonstrates how algorithmically supported formative assessment can be applied to creative tasks in teacher education, clearly distinguishing it from automated summative grading and positioning it as a human-in-the-loop support tool.
- It proposes a hybrid ML-based assessment model that combines convolutional neural networks, keyword-based thematic clustering, and rubric-aligned feedback for evaluating digital art projects.
- It provides empirical evidence from an authentic teacher education context by comparing ML-generated feedback with instructor evaluations and analyzing pre-service teachers' perceptions of and trust in automated feedback.
- It examines how engagement with algorithmic feedback supports assessment literacy, critical reflection, and understanding of algorithmic decision-making, without claiming evidence of competence development.

The study involved volunteer participants from two Kazakhstani universities and was fully integrated into the regular educational process. It did not interfere with students' academic trajectories or grading decisions, involved no sensitive or identifiable data, and was conducted under pedagogically neutral conditions.

Guided by these aims, the following research questions (RQ) were formulated:

*RQ1: To what extent do the assessments generated by the machine learning model correlate with instructor evaluations of students' digital projects based on visual criteria?*

*RQ2: How do prospective primary school teachers perceive automated formative evaluation of visual work, and to what degree are they willing to trust such forms of feedback?*

*RQ3: To what extent does algorithmic assessment contribute to the development of elements of digital pedagogical competence, such as critical reflection, understanding of algorithmic principles, and self-assessment skills?*

 

## 2. Literature Review

### 2.1 Machine Learning in Educational Assessment: Opportunities and Epistemic Limits

Machine learning is increasingly employed in educational assessment to provide scalable and consistent feedback, particularly in domains where evaluation criteria can be precisely formalized (Samuel, 2024; U.S. Department of Education, 2023). In such contexts, ML-based systems show substantial alignment with expert judgment, especially when trained on well-annotated datasets and applied to rule-governed outputs (Kusuma et al., 2022; Misgna et al., 2024).

This alignment weakens when assessment relies on pedagogical interpretation rather than formal pattern recognition. ML models optimize statistical regularities, whereas educational judgment in teacher education depends on context, intent, and professional expertise (Hopfenbeck et al., 2023; Samuel, 2024). Accordingly, automated scoring is increasingly positioned as rubric-aligned, human-in-the-loop support rather than a replacement for instructor judgment (Xu et al., 2025).

Explainability improves transparency and learner trust, as interpretable rationales are perceived more positively than opaque scores (Conijn, Kahr and Snijders, 2023; Chai et al., 2024). However, explainability does not grant access to pedagogical intent, meaning-making, or creative reasoning, even when instance-level explanations are provided (Rachha and Seyam, 2023; Saqr and López-Pernas, 2024; Khosravi et al., 2022). Accordingly, recent literature emphasizes mitigation rather than elimination of epistemic limits, through rubric alignment, provisional feedback framing, and explicit preservation of instructor authority (Xu et al., 2025; Conijn, Kahr and Snijders, 2023).

Responsible deployment therefore requires pedagogically grounded and equity-oriented frameworks, as automated assessment can otherwise reinforce bias and narrow evaluative norms (Dringó-Horváth, Rajki and Nagy, 2025; Miao and Cukurova, 2024). Interdisciplinary collaboration between educators, assessment specialists, and data scientists is widely identified as essential for ethical and educational validity (Guo et al., 2024). In teacher education, ML-based assessment thus plays a dual role: supporting learning while shaping future teachers' understanding of assessment practices and their limits.

### 2.2 Automated Assessment in Visual Arts: Technical and Conceptual Limits

Despite advances in educational ML, applications in visual and creative domains remain structurally constrained. Visual artworks frequently encode meaning through symbolism, abstraction, emotional expression, and stylistic deviation, dimensions that resist formal standardization. Convolutional neural networks show moderate to high agreement with expert ratings when evaluation targets formal visual properties such as composition, symmetry, or color distribution (Cropley and Marrone, 2025; Patterson et al., 2024). However, performance declines systematically when evaluation depends on originality, abstraction, or expressive intent (Cropley and Marrone, 2025; Patterson et al., 2024). Recent studies attempt to approximate creativity and emotional expression through indirect proxies, including novelty detection, stylistic divergence from training distributions, or multimodal alignment between images and short textual descriptors (Messer, 2024; Spee et al., 2023). These approaches remain fundamentally indirect and do not constitute pedagogical interpretation, as they lack access to artistic intent, emotional meaning, and culturally situated symbolism (Mazzone and Elgammal, 2019).

Empirical studies show systematic underestimation of works that deviate from dominant visual regularities, particularly those incorporating cultural symbolism or non-dominant aesthetic traditions (Spee et al., 2023; Zhang et al., 2025). Such divergence reflects a semantic gap rather than random noise. Models trained on visual regularities tend to interpret deviation as error because intent, context, and symbolic reference are not represented in visual feature space (Cetinic and She, 2022; Cibotaru, 2025). This limitation is architectural rather than empirical: expanding datasets may reduce variance but cannot resolve the absence of semantic understanding.

Cultural asymmetries further complicate automated assessment. Much of the literature is grounded in Western art education traditions that privilege individualist and modernist aesthetic norms (Crawford and Paglen, 2021). In contrast, Eastern and Central Asian artistic practices often emphasize culturally embedded symbolism, narrative continuity, and collective meaning-making (Bao et al., 2016). When models trained predominantly on Western datasets are applied in such contexts, stylistic deviation may be misclassified as low quality rather than culturally situated expression, a concern particularly relevant to the present study conducted in Kazakhstan (Coeckelbergh, 2023).

Pedagogical usefulness is further constrained by opacity. When automated systems output scores without interpretable justification, learners struggle to relate feedback to artistic intent, undermining trust and formative value (Nazaretsky et al., 2025). Although multimodal approaches combining visual and textual inputs show promise, most remain experimental and are rarely embedded in authentic instructional settings.

Overall, ML systems can reliably quantify visual structure but remain poorly suited to evaluating expressive deviation. Effective use in arts education therefore requires explicit role delimitation and instructor mediation, positioning algorithmic output as analytic support rather than autonomous judgment, particularly where meaning, symbolism, and originality are central (Fong and Schallert, 2023; Grájeda et al., 2024).

Student perceptions play a critical role in the educational legitimacy of AI-based assessment systems. Across higher education contexts, students tend to adopt a pragmatic but cautious stance toward automated evaluation: structured feedback is valued for clarity and consistency, while interpretive authority is consistently reserved for human instructors, especially in creative domains (Holmes, Bialik and Fadel, 2019; Chan and Hu, 2023; Tierney, Peasey and Gould, 2025). Trust and acceptance are closely linked to students' digital and AI-related competence, with higher literacy associated with greater confidence in algorithmic feedback and its responsible use (Jin et al., 2025; Anand and Hu, 2024; Ng et al., 2023; Tenberga and Daniela, 2024). These findings suggest that learner engagement with AI assessment depends less on technical accuracy alone than on transparency, contextual framing, and pedagogical mediation.

### 2.3 Research Gaps and Motivation for the Present Study

Despite growing interest in ML-assisted assessment, several gaps remain. First, most studies on creative ML assessment are conducted in laboratory or experimental settings, limiting pedagogical relevance and transferability to authentic coursework (Bulut et al., 2024; Gunasekara and Saarela, 2025; Küchemann et al., 2025). Second, learner-facing explainability remains underdeveloped, restricting opportunities for reflective engagement with assessment logic (Khosravi et al., 2022; Gunasekara and Saarela, 2025). Third, empirical validation within teacher education is scarce, with few studies combining instructor–model performance comparison, diagnostic failure analysis, and student perception data within the same instructional context (Jankowsky and Schroeders, 2022). Finally, ethical and cultural sensitivity issues remain insufficiently addressed, particularly for non-Western and stylistically diverse student populations (Chinta et al., 2024; Ferrara, 2024; Fu and Weng, 2024).

These gaps motivate the present study, which examines a rubric-aligned ML assessment pipeline for digital art projects within pre-service teacher education. By combining quantitative performance analysis, structured interpretation of failure modes, and student perception data in an authentic classroom context, the study aims to clarify both the capabilities and the structural limits of algorithmic assessment in creative domains.

By integrating technical performance analysis with epistemological and pedagogical perspectives, this review contributes to the emerging theoretical understanding of why ML-based assessment remains structurally constrained in visual arts education.

## 3. Materials and Methods

### 3.1 Research Design

The study employed a mixed-methods design combining quantitative and qualitative components. This approach enabled analysis of both the correlation between instructor-assigned grades and machine-generated scores, and students' perceptions of automated assessment within teacher education. Quantitative analysis was based on a four-criterion rubric, while qualitative data were collected through an anonymous online survey. This combination provided a comprehensive view of the model's performance and its educational implications.

The research was integrated into a regular academic course focused on developing digital visual competence among prospective primary school teachers. Student-created digital artworks, submitted as final assignments, served as the dataset. All instructor evaluations were completed before applying the machine learning model, ensuring that academic outcomes were not influenced and that the results reflect independent analysis. Quantitative data were analyzed using Pearson correlation coefficients, mean absolute error (MAE), and one-way ANOVA, while qualitative responses were examined through descriptive content analysis.

 

### 3.2 Participants and Educational Context

The study involved 180 undergraduate students enrolled in the academic program 6B01301 "Pedagogy and Methods of Primary Education" at Zhetysu University named after I. Zhansugurov and Abai Kazakh National Pedagogical University. All participants were taking a course in digital art, during which they completed their individual final projects using raster and vector graphic editors. These 180 student works formed the core dataset for model testing.

Student artworks were collected after course completion via the institutional learning management system (LMS), following the official recording of all course grades. Participants were undergraduate pre-service teachers in their second to fourth year of study, and the analyzed projects corresponded to the final assessment of the digital art course within the academic semester.

To train the machine learning model, an additional dataset of 300 archived projects from previous cohorts was used. These earlier works had been evaluated using the same four-criterion rubric by four instructors specializing in digital art education. To assess the consistency of human scoring, inter-rater reliability was calculated on a random subset of 60 projects evaluated independently by pairs of instructors. The average Cohen's kappa coefficient across all criteria was 0.72, indicating substantial agreement. All evaluations followed a collaboratively developed rubric. Prior to assessment, calibration sessions were held to ensure consistent interpretation of the criteria and to reduce subjectivity.

All new student projects were independently assessed both by instructors and by the machine learning model using identical evaluation criteria. After the evaluation phase, an anonymous online survey was administered to gather students' perceptions of the system's fairness, clarity, and relevance to their future teaching practice.

### 3.3 Evaluation Rubric and Analysis Structure

Assessment was conducted using a four-criterion rubric covering key aspects of the visual product: composition, color scheme, technical execution, and creativity. Each criterion was rated on a 5-point scale. The rubric was developed in consultation with instructors to ensure standardization and was applied in both human evaluation and machine processing. The structure of the rubric is presented in Table 1.

**Table 1: Evaluation Rubric for Digital Art Projects**

| Criterion | Description |
|---|---|
| Composition | Visual balance and logical arrangement of elements in the image, including spatial organization and alignment |
| Color Scheme | Harmony, contrast, and emotional tone of selected colors, including consistency of color relationships and overall visual coherence |
| Technical Execution | Accuracy and clarity in the use of digital tools and techniques, including line precision, layering, and resolution |
| Creativity | Originality and divergence from typical patterns within the task constraints, reflected in non-standard visual solutions |

The evaluation rubric was grounded in widely accepted formal criteria used in visual arts education, focusing on observable compositional, chromatic, and technical features (e.g., spatial balance, color contrast, and tool accuracy) rather than culturally specific symbolic interpretations. Creativity was operationalized as originality and divergence from typical patterns within the task constraints, avoiding references to stylistic canons or culturally bound aesthetic norms. This design allowed consistent application across student works reflecting diverse artistic traditions and cultural backgrounds.

The model produced outputs on a continuous numerical scale, which were rounded to one decimal place and then converted into rubric scores using predefined intervals. For instance, values from 1.0 to 1.4 corresponded to a score of 1, values from 1.5 to 2.4 to a score of 2, and so forth.

### 3.4 Model Architecture and Training

The machine learning model was developed using a supervised learning approach and trained on labeled data consisting of digital images and instructor-assigned scores. To provide limited contextual grounding, the model

incorporated a keyword-based thematic grouping mechanism. Students selected keywords from a predefined list, allowing each submission to be associated with a thematic cluster of previously evaluated works. This approach offered minimal contextual reference while remaining aligned with rubric-based assessment constraints. The overall task was framed as a multivariate regression problem, aimed at predicting four continuous values corresponding to the rubric criteria.

A shallow convolutional neural network (CNN) architecture was intentionally selected due to the moderate size of the training dataset (300 archived projects). This design balanced feature expressiveness with reduced overfitting risk and improved transparency for educational use. Three convolutional blocks were sufficient to capture mid-level visual features relevant to rubric-based criteria such as composition, color distribution, and technical execution.

A convolutional neural network was implemented to extract visual features from student artworks and map them to rubric-based scores. The architecture comprised three convolutional blocks followed by fully connected layers and an output layer with four neurons corresponding to the rubric criteria. Input images were normalized and resized to 512×512 pixels.

Model training was conducted using the mean squared error (MSE) as the loss function and the Adam optimizer. Training used an 80/20 train–validation split with early stopping based on validation loss, a batch size of 32, and a maximum of 50 epochs. Standard data augmentation techniques were applied to improve generalization. A detailed breakdown of the architecture is provided in Table 2.

**Table 2: CNN Model Architecture**

| Step | Layer Type | Parameters | Output Dimensions |
|---|---|---|---|
| **1** | Conv2D + ReLU | 3×3 kernel, 32 filters | 512×512×32 |
| **2** | MaxPooling | 2×2 window | 256×256×32 |
| **3** | Conv2D + ReLU | 3×3 kernel, 64 filters | 256×256×64 |
| **4** | MaxPooling | 2×2 window | 128×128×64 |
| **5** | Dense | 128 neurons | – |
| **6** | ReLU | – | – |
| **7** | Output Layer | 4 neurons | – |

Figure 1 presents a schematic overview of the assessment pipeline, illustrating image and keyword input, feature extraction, inference, and rubric-aligned feedback generation, with optional visual interpretability overlays.

 

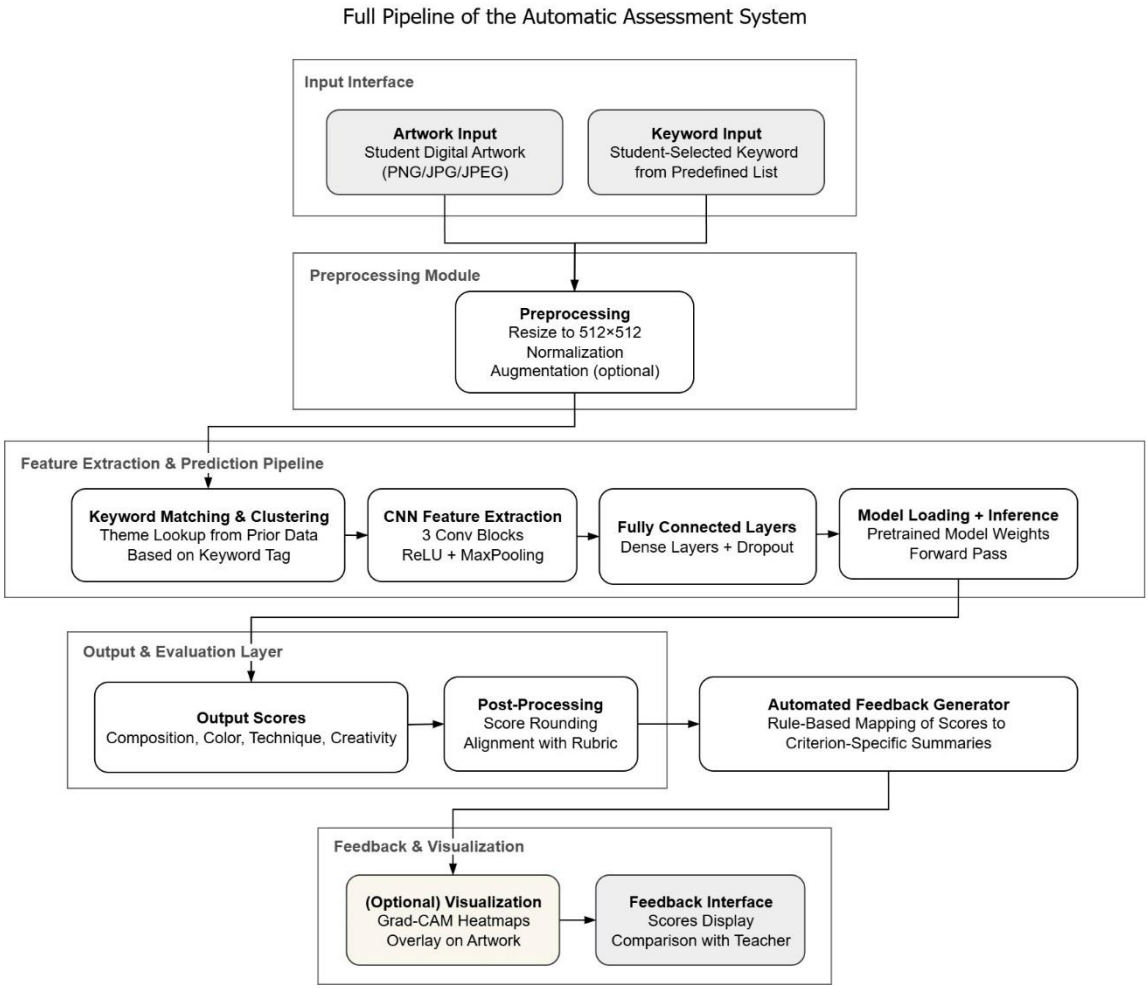


**Figure 1: Automatic Assessment Pipeline for Student Digital Artwork**

### 3.5 Comparative Analysis of Scores

Each of the 180 new student works was independently evaluated by an instructor and by the machine learning model. Discrepancies between the two scoring systems were analyzed using Pearson correlation coefficients, as well as metrics such as mean absolute error (MAE) and the proportion of matches within a one-point margin. Additional attention was given to cases where the difference between the two evaluations was two points or greater. These outliers were examined in detail to identify possible sources of divergence, such as the use of unconventional color palettes, abstract composition, expressive symbolism, or other artistic elements that may not have been well represented in the training data. The analysis revealed key strengths of the model, including high agreement on technical criteria, alongside notable limitations, particularly reduced sensitivity to conceptual originality.

### 3.6 Survey Instrument Reliability and Analysis of Student Perceptions

Following the completion of the evaluation process, students participated in an anonymous online survey that included both Likert-scale and open-ended questions. The survey focused on levels of trust in the results, perceptions of objectivity, impact on motivation, and willingness to use similar tools in future teaching practice.

Prior to inferential analysis, the internal consistency of the survey instrument was evaluated. The Digital Literacy Scale (Q1–Q3) was developed specifically for this study context to measure domain-relevant self-efficacy rather than general digital skills. It aggregates perceived competence in three key areas: technical proficiency with editing tools, theoretical understanding of algorithms, and prior practical experience with AI. The scale demonstrated excellent reliability (Cronbach's $\alpha = .912$). Participants were stratified into three levels based on a summative index (Range: 3–15) established according to Likert scale anchor points: Low Literacy ($n = 49$, scores 3–8) reflected average responses below the neutral midpoint; Medium Literacy ($n = 61$, scores 9–11)

represented functional proficiency centering around the neutral value; and High Literacy (n = 70, scores 12–15) corresponded to consistently high confidence ratings averaging 4 or above.

The Student Perceptions Scale (Q4–Q11), capturing trust, perceived objectivity, clarity, emotional response, and willingness to use AI-based assessment tools, also showed high internal consistency (α = .925; ω = .927), confirming the suitability of the instrument for group comparisons and subsequent inferential analyses.

Responses to the open-ended questions were subjected to content analysis, with coding carried out independently by two researchers. The resulting thematic structure was organized into four categories: perceived fairness, emotional response, pedagogical applicability, and critical attitudes toward algorithmic assessment.

### 3.7 Ethical Considerations

This study was part of everyday teaching practice at the university and followed basic institutional ethical principles normally used in low-risk educational studies. Teaching and grading were conducted in the usual way for the course, with instructors responsible for evaluating student work. Only after the course had ended were selected works used to examine the performance of the machine learning model in a post-hoc and purely experimental manner, aimed at analysis and formative insight rather than evaluation. The model was trained and tested using only anonymized student work, with all identifying details removed. Participation in the survey was voluntary and anonymous. As the study had no effect on teaching or student outcomes, formal ethical approval was not required.

## 4. Results and Findings

### 4.1 Comparative Analysis of Instructor and Model Scores

Agreement between instructor and model scores was examined for 180 student digital artworks. The evaluation focused on four rubric criteria: composition, color scheme, technical execution, and creativity. The results are presented in Table 3. Technical Execution showed the strongest alignment between instructor and model scores ($r = .426$, $p < .001$), with the lowest error values (MAE = 0.48; MSE = 0.85) and comparable mean scores (Instructor M = 4.14; Model M = 4.01), indicating stable performance on formally defined visual features.

**Table 3: Agreement Between Instructor and Model Scores by Criterion**

| Criterion | Instructor Mean (SD) | Model Mean (SD) | MAE | MSE | Pearson r |
|---|---|---|---|---|---|
| **Technical Execution** | 4.14 (0.80) | 4.01 (0.90) | 0.48 | 0.85 | .426*** |
| **Creativity** | 3.98 (0.97) | 3.67 (1.17) | 1.03 | 1.97 | .181* |
| **Composition** | 4.03 (0.91) | 3.79 (1.02) | 0.62 | 1.12 | .430*** |
| **Color Scheme** | 3.90 (0.97) | 3.72 (1.06) | 0.74 | 1.41 | .327*** |

Note. * $p < .05$, *** $p < .001$

Creativity exhibited the largest divergence, with the highest absolute error (MAE = 1.03; MSE = 1.97) and a moderate correlation ($r = .181$, $p = .015$). Model scores were consistently lower than instructor ratings (Model M = 3.67 vs. Instructor M = 3.98), indicating systematic underestimation rather than random disagreement. Composition ($r = .430$; MAE = 0.62) and Color Scheme ($r = .327$; MAE = 0.74) showed intermediate agreement, with the model assigning consistently lower mean scores than instructors, suggesting conservative estimation for partially interpretive visual features.

Analysis by digital literacy level revealed systematic divergence patterns. While the model followed overall performance trends, it showed a ceiling effect for highly proficient students. In the High literacy group, the instructor mean for Creativity (M = 4.27) exceeded the model mean (M = 3.89) by over one point, whereas the gap was smaller in the Low literacy group (Instructor M = 3.16; Model M = 3.43). This indicates that the model tends to underestimate highly original work, particularly among highly proficient students, interpreting stylistic deviation as error rather than creative intent.

Overall, model performance differed systematically across rubric dimensions. Criteria grounded in formally defined visual properties showed stronger alignment, whereas creativity and expressive intent exhibited higher absolute error and consistent downward bias. This pattern, reported in prior work on visual arts and automated scoring (Cetinic and She, 2022; Cropley and Marrone, 2025; Misgna et al., 2024; Patterson et al., 2024), indicates

 

that correlation alone may mask systematic score compression. Taken together, the results delineate a functional boundary: the model supports structured assessment but remains limited in evaluating originality.

### 4.2 Score Distributions and Divergence Analysis

To further examine agreement patterns, the distributions of instructor and model scores were visualized separately for each rubric criterion. Figure 2 presents overlaid histograms comparing the score distributions of the instructor and the model across all four criteria.

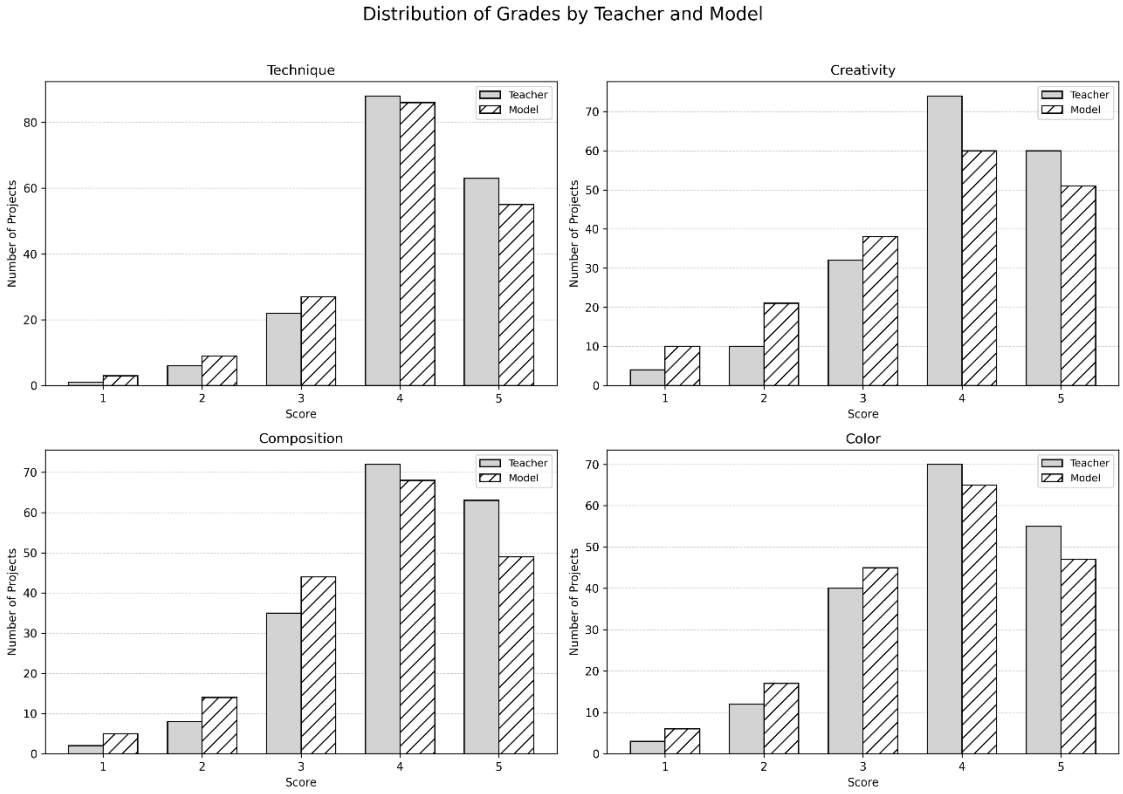


**Figure 2: Comparative Post-Rounded Distribution of Instructor and Model Scores by Criterion**

Across all criteria, model score distributions were shifted downward relative to instructor ratings, most prominently for Creativity. Smaller but consistent shifts were also present for Composition and Color Scheme, whereas distributions for Technical Execution showed closer alignment.

Analysis of cases with large discrepancies (≥ 2 points) revealed recurring patterns rather than isolated errors. According to the data in Figure 3, the most frequent sources of divergence were abstract composition (38%), unconventional color use (27%), symbolic or metaphorical content (22%), and mixed stylistic approaches (13%). These categories represent a stable set of divergence types across the dataset.

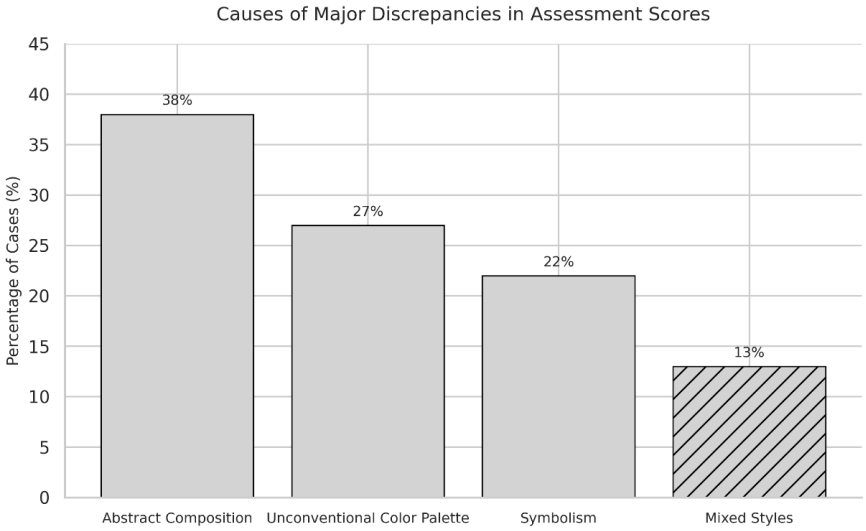


**Figure 3: Sources of Major Score Discrepancies**

Distributional patterns confirm that divergence follows recurring visual characteristics rather than random error, with the greatest dispersion observed for originality and non-standard visual expression.

### 4.3 Qualitative Evaluation and Sample-Wide Agreement

Qualitative inspection of selected cases revealed recurring patterns of disagreement. Divergences were most frequent in works employing abstract composition, expressive color contrasts, or minimalistic strategies, where instructors interpreted deviation as intentional, while the model treated it as deficiency.

Figure 4 presents average score deviations between model predictions and instructor evaluations for the full dataset (N = 180). The largest negative deviation is observed for the Creativity criterion (mean post-rounded deviation = –0.31), indicating a systematic tendency of the model to assign lower creativity scores relative to instructor judgments. Smaller but consistent negative deviations are also evident for Composition, Color, and Technique, suggesting a generally conservative scoring pattern across evaluation dimensions.

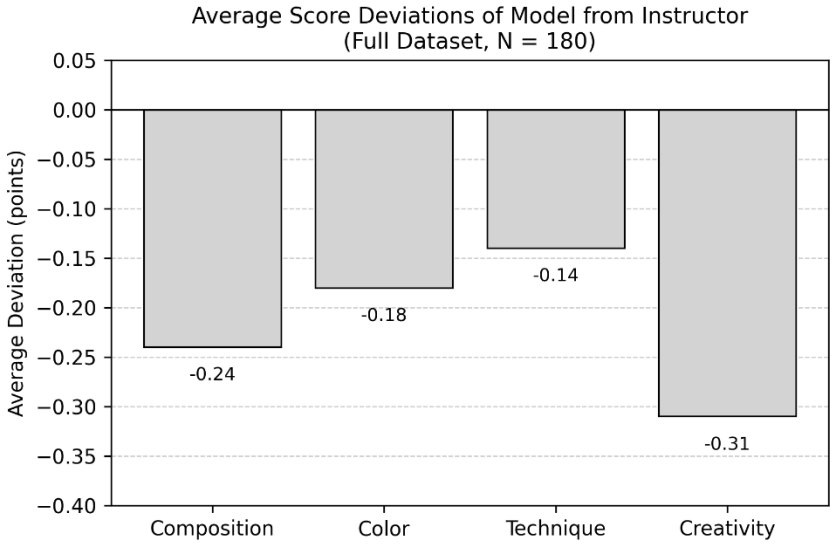


**Figure 4: Average Post-Rounded Deviations Between Model and Instructor Evaluations**

Combined qualitative and aggregate analyses show that agreement is highest for technically explicit criteria, while divergence increases for expressive and symbolic dimensions, supporting the use of the system as analytic support rather than autonomous evaluation.

 

### 4.4 Student Perceptions of the Technology

One component of the study focused on how students perceived the use of automated assessment during the learning process. An anonymous online survey was completed by 180 participants. The survey included Likert-scale items measuring trust in the model, perceived clarity and fairness of feedback, emotional response, and willingness to use similar systems in future teaching practice. Table 4 summarizes the distribution of respondents who agreed or strongly agreed with each statement.

**Table 4: Distribution of Student Responses to Scaled Statements**

| Statement | Agree or strongly agree |
|---|---|
| The model evaluates student work more objectively than a human assessor. | 114 (63.3%) |
| The model's scores are clear and understandable. | 129 (71.7%) |
| I would consider using such a system in my future teaching practice. | 105 (58.3%) |
| The absence of emotional nuance reduces the quality of feedback. | 120 (66.7%) |
| The model helps clarify assessment criteria. | 144 (80.0%) |

Survey responses indicate that students viewed the system as transparent and helpful for understanding assessment criteria, while clearly recognizing its limitations in conveying emotional nuance. Open-ended responses (n = 48) emphasized improved clarity of criteria, conditional trust, limited emotional sensitivity, and cautious classroom applicability.

To examine whether digital literacy influenced student perceptions, respondents were stratified into Low (n=49), Medium (n = 61), and High (n = 70) levels based on their self-reported confidence in basic image editing and their understanding of algorithmic principles. Table 5 presents the descriptive statistics (Mean and Standard Deviation on a 5-point scale) and the results of a one-way ANOVA for each perception variable. The analysis confirmed statistically significant differences across all three constructs. Students with higher digital literacy reported significantly greater trust in the model, a stronger willingness to use the tool, and a clearer understanding of its logic. Post-hoc comparisons indicated that the "High" literacy group consistently rated the system more favorably than the "Low" literacy group, suggesting that technical competence is a key predictor of AI acceptance.

**Table 5: Influence of Digital Literacy on Student Perceptions (Mean Scores and ANOVA Results)**

| Perception Variable | Mean (SD) | | | ANOVA Results |
|---|---|---|---|---|
| | Low Literacy (n = 49) | Medium Literacy (n = 61) | High Literacy (n = 70) | |
| **Trust in the Model** | 3.37 (1.38) | 3.69 (1.37) | 4.00 (1.12) | $F(2, 177) = 3.55$, $p = .031$ |
| **Willingness to Use** | 2.76 (1.56) | 3.54 (1.41) | 3.81 (1.28) | $F(2, 177) = 8.48$, $p < .001$ |
| **Understanding of Logic** | 3.16 (1.38) | 3.77 (1.35) | 4.27 (0.96) | $F(2, 177) = 11.92$, $p < .001$ |

Note. Scores are based on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). SD = Standard Deviation. Differences between groups are statistically significant at $p < .05$.

These results indicate a monotonic pattern: as digital literacy increases, students report higher confidence in the fairness and usefulness of algorithmic assessment, as well as greater comprehension of how model outputs are generated. At the same time, even among highly literate students, reservations regarding emotional nuance and creative interpretation remain present, underscoring the perceived need for continued instructor involvement. Overall, students perceived the system as a supportive assessment aid rather than a replacement for teacher judgment, valuing its structure while assigning responsibility for creative interpretation to instructors.

## 5. Discussion

### 5.1 Differential Model Performance Across Rubric Dimensions

The results demonstrate systematic differences in model performance across rubric dimensions. Alignment with instructor ratings is strongest for technically defined criteria such as Technical Execution and remains relatively high for Composition, where evaluation relies on formally operationalized visual features. Creativity follows a distinct pattern. Despite low correlations, absolute error is substantially higher and model scores are

consistently lower than instructor ratings, indicating systematic underestimation rather than random disagreement. Reduced variance further suggests conservative model behavior when originality departs from familiar visual patterns. This boundary reflects the system's design: convolutional feature extraction and keyword aggregation support reliable assessment of technical and structural properties but provide limited access to expressive intent. Across metrics and samples, the findings support the use of the model as an analytic aid for structured criteria, with instructor judgment remaining central for creative evaluation.

### 5.2 Systematic Divergence, Failure Modes, and the Role of Instructor Judgment

Analysis of large score discrepancies shows that divergence between model and instructor evaluations follows stable failure modes linked to specific elements of the assessment pipeline rather than random error. Three mechanisms are consistently observed. First, a semantic gap emerges in keyword-based clustering: categorical prototypes represent an "average" visual solution, causing symbolic or metaphorical works to be penalized when their visual features diverge from expected patterns. Second, the shallow CNN architecture introduces a formalism bias. While effective for detecting mid-level features such as edge clarity and structural balance, it favors conventional compositional regularities and interprets abstract or non-standard structures as technical deficiencies rather than intentional rule-breaking. Third, regression-based scoring applies a distributional penalty to unconventional color schemes, treating palettes outside dominant training distributions as noise rather than expressive choice. Together, these mechanisms function as a normalization filter that compresses score variance and systematically pulls highly original work toward the mean.

These limitations define clear boundaries of autonomous algorithmic judgment. Agreement with instructor ratings is highest where assessment depends on visible and repeatable cues, such as technical execution, and declines when evaluation relies on sensitivity, emotional nuance, or symbolic meaning. Creativity therefore shows the largest absolute deviation and a consistent downward bias, even when correlations remain low. This pattern reflects the system's representational scope: the model processes formal visual features and recurring statistical patterns but has no access to intention, cultural reference, or affect. As a result, sensitivity, originality, and creativity are preserved through use design rather than algorithmic inference. Model outputs function as provisional indicators that structure attention around explicit criteria, while instructors retain interpretive authority by contextualizing feedback through analysis of intent, symbolism, and expressive trajectory. The system thus operates as analytic support for formal assessment, stabilizing technical evaluation while preserving human judgment over creative and contextual meaning

### 5.3 Pedagogical and Methodological Implications

The findings have several implications for teacher education practice. Algorithmic assessment can increase transparency and consistency in formative evaluation, particularly in domains characterized by subjective judgment. At the same time, such systems should function as complementary tools rather than substitutes for human evaluators (Li and Botelho, 2024). The results confirm that ML-based assessment is most effective for technical and structural criteria, while rubric-aligned feedback and keyword-based grouping can support awareness of creative and symbolic aspects without claiming interpretive authority.

Engaging with algorithmic feedback supports students' understanding of assessment criteria and facilitates reflective comparison between human and machine judgment, contributing to assessment literacy and digital awareness (Lawasi, Rohman and Shoreamanis, 2024). Rather than prescribing evaluation outcomes, the system encourages reflection on how formal criteria are operationalized and where algorithmic judgment reaches its limits. The study also identifies methodological directions for improvement, including expanded training datasets, clearer visualization of scoring logic, and the integration of explanatory comments, reinforcing the role of the system as a reflective aid rather than a prescriptive evaluator. Similar principles have been noted in related applications of ML for academic integrity and monitoring in digital learning environments (Sakhipov, Omirzak and Fedenko, 2025).

To support transparency, a dedicated interface was developed to visualize key stages of the assessment pipeline. The interface (Figure 5) presents score distributions, keyword tagging, and editable instructor inputs, illustrating how automated output and human judgment can be combined in a proof-of-concept configuration.

 

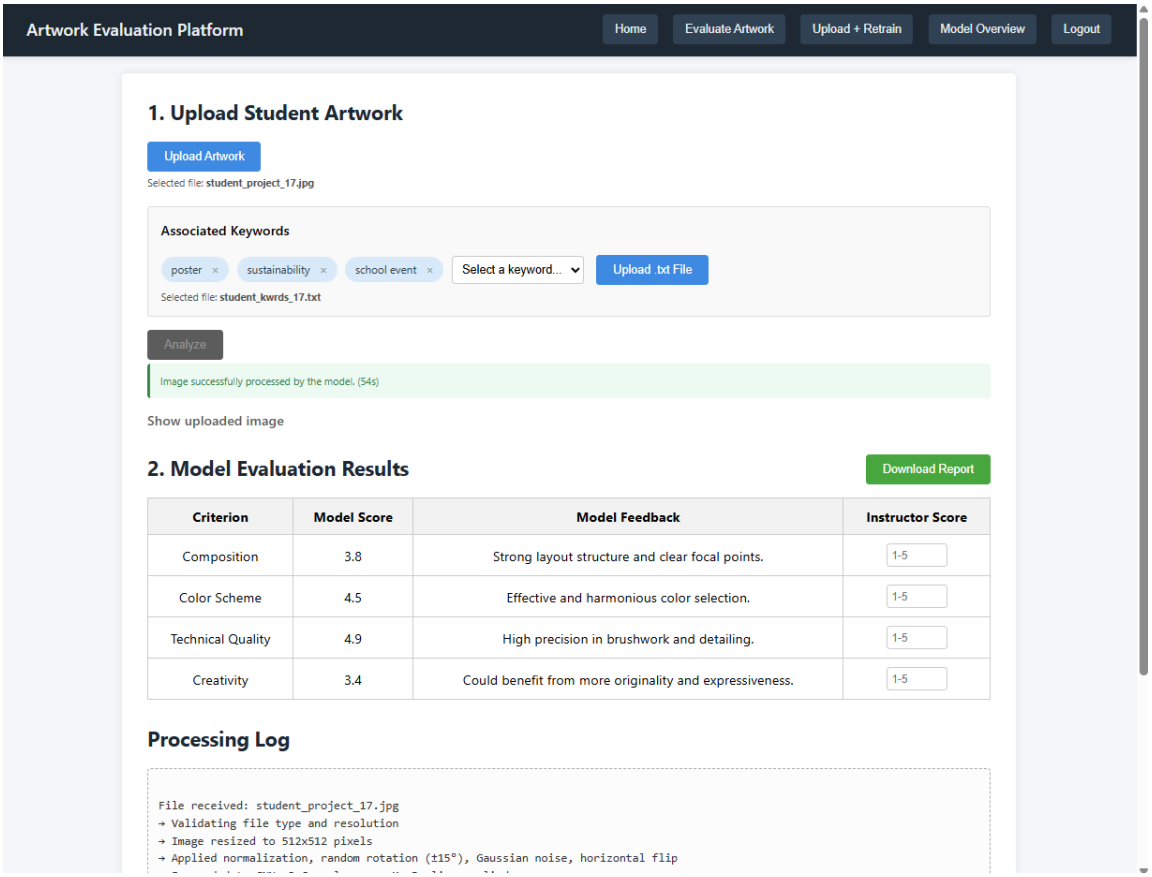


**Figure 5: Interface of the Automated Formative Assessment System**

### 5.4 Responses to the Research Questions

Regarding RQ1, the analysis showed relatively strong alignment between instructor evaluations and model scores for structurally defined criteria such as technical execution and composition, with low error rates and most scores falling within one point of instructor judgments. Performance declined for creative criteria involving originality and symbolic intent, indicating that while the model reliably assesses measurable visual features, interpretive evaluation remains dependent on instructor judgment.

For RQ2, survey results indicate a generally positive but cautious student stance toward automated feedback. Respondents valued clarity and consistency, while expressing concerns about limited emotional and contextual sensitivity. Trust and willingness to use the system varied significantly by digital literacy level, with higher literacy associated with greater confidence in the system's fairness and relevance. These differences are interpreted as associative rather than causal.

Regarding RQ3, the study does not provide evidence of learning gains or competence development. Instead, it documents increased awareness of assessment criteria and greater engagement in reflective comparison between human and algorithmic evaluations. In this study, algorithmic assessment thus contributes to reflective orientation and assessment literacy, rather than directly to pedagogical skill acquisition.

### 5.5 Study Limitations and Validity Considerations

This study, while informative, has several limitations. The dataset of 180 current and 300 archived projects supports comparative analysis but does not capture the full diversity of visual expression, particularly forms where meaning is conveyed through culturally specific, symbolic, or minimal visual strategies. The research was conducted within a single pre-service teacher education program in Kazakhstan and focused on one course type, which constrains generalization to other educational settings. In contrast to much automated assessment research that targets narrowly formalized criteria, and to creative ML studies typically conducted in laboratory or benchmark settings, this study examines model behavior in an authentic classroom context; this increases ecological validity but also introduces contextual variability that cannot be fully controlled. The model processes visual features and keyword-based thematic labels but lacks semantic and contextual understanding, resulting

in reduced sensitivity to conceptually rich but visually sparse works; this limitation is architectural and cannot be resolved through dataset expansion alone. Survey findings rely on self-reported perceptions and are subject to response bias, while variables such as prior AI experience were not independently controlled. The cross-sectional design precludes claims about learning gains or competence development, and the study does not provide longitudinal or behavioral evidence of instructional impact.

### 5.6 Future Directions for Model Development and Research

The findings point to several directions for further work. While a multimodal framework that combines visual features with keyword grouping is already implemented, future research should examine richer forms of semantic integration and interpretability support. Expanding the training corpus may reduce variance in technical scoring but is unlikely to resolve limitations related to symbolic interpretation. Additional studies should test the model in other visually oriented disciplines and educational contexts, with adapted rubrics and comparative instructor baselines. Further research is also needed to evaluate how different feedback representations, including visual vignettes and process-level explanations, influence understanding of algorithmic judgment and support reflective engagement with assessment criteria.

## 6. Conclusion

This study examined a rubric-aligned machine learning pipeline for supporting the assessment of digital art projects in pre-service teacher education. Quantitative comparison with instructor scores addressed RQ1 by showing strong alignment for structured criteria such as technical execution and composition ($r \approx .43$; MAE ≈ 0.5), alongside systematic underestimation for creativity (MAE > 1.0), indicating a stable accuracy–creativity trade-off rather than random error. RQ2 was addressed through survey and inferential analysis, which showed generally positive but cautious perceptions of automated feedback, with trust and willingness to use the system varying significantly by digital literacy level and concerns directed primarily at creative scoring rather than technical criteria. RQ3 was addressed by documenting increased awareness of assessment criteria and reflective comparison between human and algorithmic judgments, while explicitly not providing evidence of competence development or learning gains. Across all research questions, the findings reinforce that the system functions as rubric-aligned triage and analytic support, augmenting but not replacing instructor judgment. Safe and responsible use depends on preserving instructor authority over interpretation, contextual meaning, and creative intent, while treating algorithmic output as provisional rather than evaluative. Future work should focus on multimodal extensions that incorporate semantic input, longitudinal designs that test instructional impact, and classroom-based deployments that examine how such systems can be integrated into formative feedback practices without constraining creativity.

**AI Statement**: During the preparation of this manuscript, the authors used Grammarly to enhance readability and language.

**Ethics Statement**: Ethical approval was not required, as the study involved no intervention, personal data, or impact on academic outcomes.

## References

Anand, B. and Hu, Y. (2024). Play testing and reflective learning AI tool for creative media courses. In: *Proceedings of the 16th International Conference on Computer Supported Education – Volume 1: CSEDU*, pp.146–158. https://doi.org/10.5220/0012633600003693

Bao, Y., Yang, T., Lin, X., Fang, Y., Wang, Y., Pöppel, E. and Lei, Q. (2016). Aesthetic preferences for Eastern and Western traditional visual art: identity matters. *Frontiers in Psychology, 7*, 1596. https://doi.org/10.3389/fpsyg.2016.01596

Bedir, S. and Freedman, K. (2024). Integrating digital technologies and AI in art education: pedagogical competencies and the evolution of digital visual culture. *Participatory Educational Research, 11*, pp.57–79. https://doi.org/10.17275/per.24.94.11.6

Buckingham Shum, S., Lim, L.-A., Boud, D., Bearman, M. and Dawson, P. (2023). A comparative analysis of the skilled use of automated feedback tools through the lens of teacher feedback literacy. *International Journal of Educational Technology in Higher Education, 20(1)*, 40. https://doi.org/10.1186/s41239-023-00410-9

Bulut, O., Beiting-Parrish, M., Casabianca, J.M., Slater, S.C., Jiao, H., Song, D., Ormerod, C., Fabiyi, D.G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S.N., Wongvorachan, T., Liu, J.X., Tan, B. and Morilova, P. (2024). The rise of artificial intelligence in educational measurement: opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation*, *5*(3), 3. https://doi.org/10.59863/MIQL7785

Cai, Y. (2025). Integrating machine learning in art education: research framework and theoretical analysis approach. *International Journal of Advanced Research and Interdisciplinary Scientific Endeavours*, *2*, pp.537–542. https://doi.org/10.61359/11.2206-2515

 

Cetinic, E. and She, J. (2022). Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *18*(2), 66. https://doi.org/10.1145/3475799

Chai, F., Ma, J., Wang, Y., Zhu, J. and Han, T. (2024). Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. *Frontiers in Psychology*, *15*, 1221177. https://doi.org/10.3389/fpsyg.2024.1221177

Chan, C.K.Y. and Hu, W. (2023). Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*, 43. https://doi.org/10.1186/s41239-023-00411-8

Chinta, S.V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Le Quy, T. and Zhang, W. (2024). *FairAIED: Navigating fairness, bias, and ethics in educational AI applications*. arXiv. Available at: https://arxiv.org/abs/2407.18745 [Accessed 10 December 2025].

Cibotaru, V. (2025). Is there computational creativity?. *AI & Society*. https://doi.org/10.1007/s00146-025-02708-w

Coeckelbergh, M. (2023). Narrative responsibility and artificial intelligence. *AI & Society*, *38*, pp.2437-2450. https://doi.org/10.1007/s00146-021-01375-x

Conijn, R., Kahr, P. and Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics*, *10*(1), pp.37–53. https://doi.org/10.18608/jla.2023.7801

Crawford, K. and Paglen, T. (2021). Excavating AI: the politics of images in machine learning training sets. *AI & Society*, *36*, pp.1105–1116. https://doi.org/10.1007/s00146-021-01162-8

Cropley, D.H. and Marrone, R.L. (2025). Automated scoring of figural creativity using a convolutional neural network. *Psychology of Aesthetics, Creativity, and the Arts*, *19*(1), pp.77–86. https://doi.org/10.1037/aca0000510

Dringó-Horváth, I., Rajki, Z. and Nagy, J.T. (2025). University teachers' digital competence and AI literacy: moderating role of gender, age, experience, and discipline. *Education Sciences*, *15*(7), 868. https://doi.org/10.3390/educsci15070868

Ferrara, E. (2024). Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. *Sci*, *6*(1), 3. https://doi.org/10.3390/sci6010003

Fong, C. and Schallert, D. (2023). "Feedback to the future": advancing motivational and emotional perspectives in feedback research. *Educational Psychologist*, *58*(1), pp.1–16. https://doi.org/10.1080/00461520.2022.2134135

Fu, Y. and Weng, Z. (2024). Navigating the ethical terrain of AI in education: a systematic review on framing responsible human-centered AI practices. *Computers and Education: Artificial Intelligence*, *7*, 100306. https://doi.org/10.1016/j.caeai.2024.100306

Grájeda, A., Córdova, P., Córdova, J.P., Laguna-Tapia, A., Burgos, J., Rodríguez, L. and Sanjinés, A. (2024). Embracing artificial intelligence in the arts classroom: understanding student perceptions and emotional reactions to AI tools. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2378271

Gunasekara, S. and Saarela, M. (2025). Explainable AI in education: techniques and qualitative assessment. *Applied Sciences*, *15*(3), 1239. https://doi.org/10.3390/app15031239

Guo, S., Latif, E., Zhou, Y., Huang, X. and Zhai, X. (2024). *Using generative AI and multi-agents to provide automatic feedback*. arXiv. Available at: https://arxiv.org/abs/2411.07407 [Accessed 10 December 2025].

Holmes, W., Bialik, M. and Fadel, C. (2019). *Artificial intelligence in education: promise and implications for teaching and learning*. Boston: Center for Curriculum Redesign.

Hopfenbeck, T.N., Zhang, Z., Sun, S.Z., Robertson, P. and McGrane, J.A. (2023). Challenges and opportunities for classroom-based formative assessment and AI: a perspective article. *Frontiers in Education*, *8*, 1270700. https://doi.org/10.3389/feduc.2023.1270700

Jankowsky, K. and Schroeders, U. (2022). Validation and generalizability of machine learning prediction models on attrition in longitudinal studies. *International Journal of Behavioral Development*, *46*(2), pp.169–176. https://doi.org/10.1177/01650254221075034

Jin, F.J.-Y., Maheshi, B., Lai, W., Li, Y., Gasevic, D., Chen, G., Charwat, N., Chan, P.W.K., Martinez-Maldonado, R., Gašević, D. and Tsai, Y.-S. (2025). Students' perceptions of generative AI-powered learning analytics in the feedback process: a feedback literacy perspective. *Journal of Learning Analytics*, *12*(1), pp.152–168. https://doi.org/10.18608/jla.2025.8609

Khosravi, H., Buckingham Shum, S., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S. and Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kusuma, J., Halim, K., Pranoto, E., Kanigoro, B. and Irwansyah, E. (2022). Automated essay scoring using machine learning. In: *Proceedings of the 2022 International Conference on Cybernetics and Intelligent Systems (ICORIS)*, pp.1–5. https://doi.org/10.1109/ICORIS56080.2022.10031338

Küchemann, S., Avila, K.E., Dinc, Y., Hortmann, C., Revenga, N., Ruf, V., Stausberg, N., Steinert, S., Fischer, F., Fischer, M., Kasneci, E., Kasneci, G., Kuhr, T., Kutyniok, G., Malone, S., Sailer, M., Schmidt, A., Stadler, M., Weller, J. and Kuhn, J. (2025). On opportunities and challenges of large multimodal foundation models in education. *npj Science of Learning*, *10*(1), 11. https://doi.org/10.1038/s41539-025-00301-w

Lawasi, M., Rohman, V. and Shoreamanis, M. (2024). The use of AI in improving students' critical thinking skills. *Proceedings Series on Social Sciences & Humanities*, *18*, pp.366–370. https://doi.org/10.30595/pssh.v18i.1279

Li, H. and Botelho, A.F. (2024). Developing explainable AI systems to support feedback for students. In: *Proceedings of the 17th International Conference on Educational Data Mining*, pp.998–1002. https://doi.org/10.5281/zenodo.12730029

Mazzone, M. and Elgammal, A. (2019). Art, creativity, and the potential of artificial intelligence. *Arts*, *8*(1), 26. https://doi.org/10.3390/arts8010026

Messer, U. (2024). Co-creating art with generative artificial intelligence: Implications for artworks and artists. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100056. https://doi.org/10.1016/j.chbah.2024.100056

Miao, F. and Cukurova, M. (2024). *AI competency framework for teachers*. Paris: UNESCO. https://doi.org/10.54675/ZJTE2084

Misgna, H., On, B.-W., Lee, I. and Choi, G.S. (2024). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, *58*(2), 36. https://doi.org/10.1007/s10462-024-11017-5

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J. and Käser, T. (2025). The critical role of trust in adopting AI-powered educational technology for learning: an instrument for measuring student perceptions. *Computers and Education: Artificial Intelligence*, *8*, 100368. https://doi.org/10.1016/j.caeai.2025.100368

Ng, D.T.K., Leung, J.K.L., Su, J., Ng, R.C.W. and Chu, S.K.W. (2023). Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world. *Educational Technology Research and Development*, *71*(1), pp.137–161. https://doi.org/10.1007/s11423-023-10203-6

Patterson, J.D., Barbot, B., Lloyd-Cox, J. and Beaty, R.E. (2024). AuDrA: An automated drawing assessment platform for evaluating creativity. *Behavior Research Methods*, *56*(4), pp.3619–3636. https://doi.org/10.3758/s13428-023-02258-3

Rachha, A. and Seyam, M. (2023). Explainable AI in education: current trends, challenges, and opportunities. In: *2023 IEEE SoutheastCon*, pp.232–239. https://doi.org/10.1109/SoutheastCon51012.2023.10115140

Sakhipov, A., Omirzak, I. and Fedenko, A. (2025). Beyond face recognition: a multi-layered approach to academic integrity in online exams. *Electronic Journal of e-Learning*, *23*(1), pp.81–95. https://doi.org/10.34190/ejel.23.1.3896

Samuel, K. (2024). The role of artificial intelligence in educational assessment. *Eurasian Experiment Journal of Scientific and Applied Research*, *5*, pp.44–48.

Saqr, M. and López-Pernas, S. (2024). Why explainable AI may not be enough: predictions and mispredictions in decision making in education. *Smart Learning Environments*, *11*(1), 52. https://doi.org/10.1186/s40561-024-00343-4

Spee, B.T.M., Mikuni, J., Leder, H., Scharnowski, F., Pelowski, M. and Steyrl, D. (2023). Machine learning revealed symbolism, emotionality, and imaginativeness as primary predictors of creativity evaluations of western art paintings. *Scientific Reports*, *13*(1), 12966. https://doi.org/10.1038/s41598-023-39865-1

Tenberga, I. and Daniela, L. (2024). Artificial intelligence literacy competencies for teachers through self-assessment tools. *Sustainability*, *16*(23), 10386. https://doi.org/10.3390/su162310386

Tierney, A., Peasey, P. and Gould, J. (2025). Student perceptions on the impact of AI on their teaching and learning experiences in higher education. *Research and Practice in Technology Enhanced Learning*, *20*, 005. https://doi.org/10.58459/rptel.2025.20005

U.S. Department of Education, Office of Educational Technology (2023). *Artificial intelligence and the future of teaching and learning: insights and recommendations*. Washington, DC.

Xu, W., Kassim, M.S.S., Hoo, W.L., Yang, W. and Xu, T. (2025). Explainable AI for education: Enhancing essay scoring via rubric-aligned chain-of-thought prompting. *International Journal of Modern Physics C, 37*(6), 2542013. https://doi.org/10.1142/S0129183125420136

Zhang, R., Lee, H., Wu, R., Yang, W. and Pan, Y. (2025). Is the impact of artificial intelligence generation tools on improving art education positive or negative? Perspectives of professors and students. *Interactive Learning Environments*, *33*(10), 5944–5975. https://doi.org/10.1080/10494820.2025.2490173

Zheng, C., Yu, Z., Jiang, Y., Zhang, M., Lu, X., Jin, J. and Gao, L. (2025). ArtMentor: AI-assisted evaluation of artworks to explore multimodal large language models capabilities. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 659. https://doi.org/10.1145/3706598.3713274

 