Using a Machine-Learning Algorithm to Classify the Formal Developmental Levels of Ugandan Preschool Children's Drawings

Journal:	*Journal of the Association for Information Science and Technology*
Manuscript ID	Draft
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Dent, Valeda; Emory University, Emory Libraries, Museum, & Center for Digital Scholarship Goodman, Geoff; Emory University, School of Medicine, Candler School of Theology Vinh, Tuan; Emory University, Department of Chemistry; Department of Computer Science Nimmakayala, Kiran; Emory University, Center for AI Learning Ottolin, Thomas; Emory University, Center for AI Learning Morejon, Alisha; Emory University, Department of Computer Science Sutherland, Joseph; Emory University, Center for AI Learning; Department of Computer Science
Keywords:	machine learning < artificial intelligence < computer applications < computer operations < (activities and operations), computer vision < artificial intelligence < computer applications < computer operations < (activities and operations), image analysis < image processing < information processing < computer operations < (activities and operations)

Using a Machine-Learning Algorithm to Classify the Formal Developmental Levels of Ugandan Preschool Children's Drawings

Valeda F. Dent, Emory University (corresponding author)
vfdent@emory.edu

Geoff Goodman, Emory University
Ggoodm6@emory.edu

Tuan Vinh, Emory University
tuan.vinh@emory.edu

Kiran Kumar Nimmakayala, Emory University
kiran.kumar.nimmakayala@emory.edu

Thomas Ottolin, Emory University
thomas.holden.ottolin@emory.edu

Alisha Morejon, Emory University
alisha.morejon@emory.edu

Joseph Sutherland, Emory University
joseph.lyons.sutherland@emory.edu

Developing a Machine Learning Model to Categorize Children's Drawings

Abstract

Artificial intelligence (AI) has revolutionized the ability to process and analyze large-scale data sets, surpassing human capacity in both speed and efficiency. This study presents a machine learning model based on the ResNet50 architecture, developed to classify Ugandan preschool children's drawings according to Lowenfeld's (1947) stages of early artistic development: scribble, pre-schematic, and schematic. Utilizing Scale 10 of the culturally sensitive Formal Elements (FE) scale (Tuman, 1998, 1999; Goodman et al., 2022), the model was trained on 757 manually labeled drawings—a subset of an extensive collection nearing 140,000 images. The model achieved an intraclass correlation coefficient (ICC) of 0.85 with human coders, a validation accuracy of 81%, and a test accuracy of 78%. By automating the classification process, this AI model not only reduces the burden on human coders but also enables high-throughput analysis of large data sets, facilitating scalability for similar studies. The findings have significant implications for information science professionals involved in data analysis and knowledge management, offering valuable insights into children's cognitive and artistic development. Future work aims to expand the model's classification capabilities to include additional FE scales and apply it within other cultural contexts.

Keywords*: Machine Learning, ResNet50, Children's Drawings, Artistic Development, Uganda*

Introduction

In 2014, in collaboration with two local village libraries in Uganda, researchers conducted an early childhood school readiness and literacy study (Goodman & Dent, 2019) that included implementing the Storytelling/Story Acting (STSA) protocol (Paley, 1990), a play intervention program designed to facilitate preschool children's school readiness skills.

Children aged 3 to 5 and their caregivers from the villages of Mpigi (n = 61) and Kabubbu (n = 62) participated. Children were randomly assigned to participate in either the STSA play intervention (n = 63) or a story-reading activity (n = 60) for one hour twice weekly for six months. All children were given the opportunity to draw during their participation time, and these drawings form the data set on which this project is based. Goodman, Dent, Tuman, and Lee (2022) present the results of the initial drawing analysis, which included 263 drawings, but these drawings represent only a fraction of the total number of drawings still being generated. As of March 2024, the longitudinal STSA intervention continues in Kabubbu, with additional drawings being generated each week. The current estimate of drawings collected to date is close to 140,000, which presented researchers with a challenge: how is it possible to analyze the growing data set of original drawings when the work to code each drawing is so time and expertise-dependent? During the course of the original study (Goodman et al., 2022), it took a team of two expert coders more than 8 hours each for training and reliability and another 88 hours to code the initial sample of 263 drawings against two different scales, including the Formal Elements (FE) scale used in this study (coding typically took between 20 to 60 minutes per drawing). It became clear that without unlimited resources and time, it would be impossible to code the entire data set. Researchers became interested in applying machine learning to support the continual analysis of the drawings as the only way to understand the relationship between children's artistic development and school readiness and make this unique data set available to other scholars and educators. The researchers hypothesized that the machine learning model would more accurately identify features from FE Scale 10 (Goodman et al., 2022) as compared to human coders. Specifically, ratings coded by the ResNet50 machine learning algorithm could establish an ICC of 0.80 or higher with the developmental level ratings coded by the human coder.

Drawing analysis and scale development

The rationale for studying the drawings of children is most succinctly described by Jensen, Sumanthiran, Kirkorian, Travers, Rosengren, and Rogers (2023):

> "Most importantly, where many assessments seek to isolate and measure distinct, individual aspects of functioning, drawing requires the joint use and coordination of many faculties together: perception, imagery, spatial cognition, planning, conceptual knowledge, and motor control. Drawings thus have the potential to uncover many different and intersecting facets of the developing mind using an engaging task that does not rely heavily on language and that children regularly undertake in everyday life" (p.2).

This project takes place against the backdrop of a more extensive longitudinal study that relies on drawings to provide insight into children's inner worlds (Goodman et al., 2022). Analyzing children's drawings as a way to deepen our understanding of children's emotional states is not new (Farokhi & Hashemi, 2011; Thomas & Silk, 1990; Bat Or, Kourkoutas, Smyrnaki, & Potchebutzky, 2019; Jiggetts, 2021; Stauffer, 2019), but it can be time-consuming and intense work (Merriman & Guerin, 2006). The best-known early work in this area suggests that all children progress through similar artistic developmental stages (Alter-Muri & Vazzano, 2014; Brittain & Lowenfeld, 1987; Feldman, 1987; Gardner, 1980; Lowenfeld, 1947). One motivating factor for the current project is that while there is ample literature on the impact of culture and environment on children's artistic development (Ahman, 2018: Farokhi & Hashemi, 2011; Glewwe, Ross, & Wydick, 2018), the same cannot be said for drawings created by African children. There are distinct differences in terms of how children from across the continent represent themselves, others, and the environment in their drawings (Court, 1989; Kalveston & Odman, 1979; Aronsson & Andersson, 1996), offering challenges to the claim of universal artistic development as proffered by Lowenfeld (1947) and others.

The work of Viktor Lowenfeld influenced the development of the two scales used in this study. One of these scales, the FE scale, was used to train and test the machine learning model.

Researchers have suggested that drawings with highly developed content and formal elements are indicators of school readiness (Gan, Meng, & Xie, 2016; Haidkind, Henno, Kikas, & Peets, 2011), so the use of scales that can measure these elements is an important feature of our study. Lowenfeld's stage theory of artistic development (1947) features six stages that progress from early childhood to late adolescence: The Scribbling Stage, The Preschematic Stage, The Schematic Stage, The Gang Age, The Pseudo-Naturalistic Stage, and finally, Adolescent Art (Lowenfeld, 1947). Each stage is classified into three basic categories: Drawing Characteristics, Space Representation, and Human Figure Representation (Lowenfeld, 1947).

It should be noted that in the current study, the initial investigation to identify appropriate drawing analysis instruments revealed that none were suitable for the Ugandan population being studied. Very few validated instruments have an international focus (Stiles & Gibbons, 2000), and none of these were suitable for a Ugandan preschool sample. To address this, researchers drew from previous work (Tuman, 1998, 1999; Gantt & Anderson, 2009) to create a more culturally sensitive rating scale that captures structural characteristics of drawings without imposing content-specific interpretations, making it suitable for diverse populations.

Machine learning as a tool for drawing analysis in educational and developmental research

Machine learning has made notable contributions in automating the classification of large visual data sets, such as children's drawings, often used as indicators of cognitive and emotional development (Kallitsoglou, Repana, & Shiakou, 2022). Attempts to analyze drawings and paintings using computer-aided approaches are widely documented (Wang, Kandemir, & Li, 2020; Krizhevsky, Sutskever, & Hinton, 2012; Thomas, Powell, Polsley, Ray, & Hammond, 2022). Barton (1997) reminds us of the evolving nature of this work, stating that "the dominant model of drawing research has moved from an emphasis on the finished product to a focus on the active process of interpretation and generation" (Barton, 1997, p.301). Discoveries in computer vision and other computerized tools have enabled more precise identification of elements in drawings (Eitz, Hays, & Alexa, 2012; Li, 2013; Ravindran, 2022; Lee, Kim, & Kim, 2024; Jensen, Sumanthiran, Kirkorian, Travers, Rosengren, & Rogers, 2023). Burton (1997) describes early attempts to harness intelligent computing to understand the dynamics and mechanics of children's drawings. Representation of Spatial Experience (ROSE) and Emergence of Representation (EOR) (Burton, 1995) are two early systems created to emulate children's drawings from a computer-

generated prompt. How these systems continuously learn by repetition to recognize themes and actions to achieve their goals is an important and foundational principle in more recent machine learning models. Beltzung, Pele, Renoult, & Sueur (2023) also address bias reduction that might be present when humans code drawings as an affordance of using machine learning to analyze and categorize drawings. Beltzung et al. (2023) and Martinent (2021) remind us that while such biases may seem small when only identifying elements such as color, the stakes are higher when the goal is to identify cognitive and psychological constructs. Beltzung et al. (2023) also suggest that neural networks (a type of deep learning) hold great promise for drawing analysis. Deep learning can facilitate the segmented labeling of the parts of a drawing or the classification of the entire drawing. Yu, Yang, Liu, Song, Xiang, & Hospedales (2017) demonstrate this approach when describing the Sketch-a-Net prototype, which used a convolutional neural network, or CNN, to create a supervised classification model for drawings using the TU Berlin data set (Eitz et al., 2012). The researchers describe pre-training the model, stroke management, and how they guarded against overfitting (Yu et al., 2017). The experience described by Yu et al. (2017) holds particularly relevant knowledge for the current classification project.

Variability in outcomes across the different machine-learning approaches is also important. Refaat and Atiya (2009) used a support vector machine kernel model that achieved a 93% accuracy rate in human figure drawing identification. Eitz et al. (2012) used a large data set of 20,000 drawings and found that machine-learning identification was accurate only 56% of the time as compared to human identification. A 2019 study by Monica, Davu, Caroline, and Jagganath (2019) achieved a 77% accuracy rate by using a k-nearest neighbor machine learning model. Lawrie (2022) successfully used a data set of 453 drawings from children in rural India to cross-validate LASSO (Least Absolute Shrinkage and Selection Operator, a machine learning approach) and certain psychological distress indicators, while Thomas et al. (2022) used a data set of over 3,000 drawings to evaluate the precision of machine learning models in distinguishing between the strokes in child and adult drawings. Thomas and his team used five different machine learning classifier models in their study: Random Forest, Decision Tree, Zero-Rule, Naive Bayes, and Support Vector Machines (Thomas et al., 2022). Of these five, Random Forest and Decision Tree proved the most precise, with F1 scores of 0.906 and 0.885, respectively.

Prior studies have validated that CNNs are highly effective in classifying complex image data, especially in tasks that require discerning developmental patterns (Krizhevsky, Sutskever, &

Hinton, 2012; Lecun, Bengio, & Hinton, 2015). The ResNet architecture, introduced by He, Zhang, Ren, & Sun (2016), has proven particularly effective for deep learning tasks, thanks to its residual learning techniques that address vanishing gradients, enabling deeper networks to be trained effectively. ResNet models are especially suited for tasks requiring detailed image classification, such as identifying various stages of artistic development in children's drawings. ResNet has been successfully applied across diverse domains, including medical image analysis (Litjens, Kooi, Bejnordi, Setio, Ciompi, Ghafoorian, & van der Laak, 2017) and education (Wang, Bovik, Sheikh, & Simoncelli, 2004), which require the automated analysis of large data sets.

In developmental psychology, the classification of children's drawings is commonly structured around Lowenfeld's (1947) stages of artistic development—mainly scribble, pre-schematic, and schematic—which outline how children's visual expression evolves. Historically, manual coding has been used to classify these stages. Still, the advent of artificial intelligence (AI) has made it possible to automate this process, leading to more efficient and consistent classifications. AI-based coding facilitates large-scale longitudinal studies, which would otherwise be labor-intensive to conduct manually.

Additionally, AI models have enabled the analysis of large-scale data sets, such as children's drawings, making it easier to perform cross-cultural comparisons and analyze developmental trends across time. The scalability provided by AI ensures the efficiency and consistency of data analysis, which would be unattainable using human coders alone. With large data sets like the 140,000 drawings generated so far by this study, AI's role in automating classification has proven essential in exploring children's artistic and cognitive development.

Method

Participants

Data were collected from two groups of Ugandan villagers: caregivers and children ages 3 to 5 from Mpigi ($n = 61$) and Kabubbu ($n = 62$). Both groups of children were randomly assigned to participate in either the Storytelling/Story-Acting (STSA) play intervention ($n = 63$) or a story-reading activity ($n = 60$) for one hour twice per week for six months (Goodman & Dent, 2019). The play intervention and story-reading activity, however, are not the focus of this study.

All primary caregivers with children ages 3 to 5 consented to their participation and the participation of their children and were enrolled in the study. Caregivers in both conditions tended to be living in economically poor conditions with their spouses or partners in large households and were primary school graduates. Recruitment was conducted through community outreach and word of mouth on a first-come, first-served basis due to the lack of centralized records and logistical challenges in these rural settings. Children were 43.1% male ($n = 53$) and had not yet begun primary school (Goodman & Dent, 2019). The selection criterion is based on measurable improvement, ensuring that the selected children are representative of those showing notable developmental gains within the sample. We aimed to maintain diversity in our subsample by selecting two boys and two girls, two participants from each village, and an equal number from the STSA intervention and the control group. This approach helps to mitigate potential biases and enhances the applicability of our findings.

From the original study of 123 preschool children, we selected the four children who had made the most significant improvement in their school readiness total scores on the Bracken School Readiness Assessment during the six-month study (Table 1).

Table 1

Participants

	Age	Gender	Group
Child 1	64 months	Female	Control: Kabubbu/Zebras
Child 2	50 months	Male	STSA: Kabubbu/Lions
Child 3	48 months	Female	STSA: Mpigi/Pigeons
Child 4	65 months	Male	Control: Mpigi/Eagles

In this manner, we selected two boys, two girls, two STSA participants, two control group participants, two children from Kabubbu, and two children from Mpigi (Goodman & Dent, 2019). However, we acknowledge that additional stratification dimensions—such as socioeconomic status, parental education levels, or health indicators—could provide a more comprehensive assessment of representativeness. Due to resource constraints and the exploratory nature of our study, we were limited to the variables mentioned.

Regarding the representativeness of the selected children compared to the larger population of Ugandan preschool children, we are not able to provide data on the proportion of scribble, pre-schematic, and schematic drawings in the larger population, as this is one of the first studies to collect such data in sub-Saharan Africa. Therefore, our findings contribute novel insights into children's artistic development in this context.

Child Measure

At both Time 1 (January 2014) and Time 2 (August 2014), all children completed three measures designed to assess three indices of school readiness: 1) emergent literacy, 2) oral language, and 3) theory of mind skills. These measures were selected because they are easy to administer and do not depend heavily on expressive vocabulary skills (Goodman & Dent, 2019). We report on only the emergent literacy measure, the Bracken School Readiness Assessment.

Bracken School Readiness Assessment—Third Edition (BSRA-3; Bracken, 2007). The BSRA-3 is a 10-minute, 85-item interview that assesses school readiness skills in five domains: 1) colors, 2) letters, 3) numbers/counting, 4) sizes/comparisons, and 5) shapes. The child is shown

a series of four pictures and instructed to point to the picture corresponding to the correct response. Split-half reliabilities for children ages 3 to 5 ranged from .96 to .97 (Bracken, 2007). The BSRA-3 was also shown to distinguish language-impaired and intellectually disabled children from a normative sample that included a range of ethnicities, US geographical locations, and socioeconomic status (SES; Bracken, 2007).

Each item was assigned 1 point for each correct response. A subtest was discontinued when the child received three consecutive scores of 0. Items were then summed both within and between domains so that total scores could range from 0 to 85. Raw scores rather than norm-referenced standard scores were used in the statistical analyses because we were not comparing this sample to a normative sample in the US. The BSRA-3 was used as a measure of emergent literacy skills (Goodman & Dent, 2019).

Procedure

In January 2014, researchers traveled to the rural Ugandan villages of Mpigi and Kabubbu to collect Time 1 data on the caregivers and children. The librarians at the Mpigi Community Library and Kabubbu Community Library, as well as our project coordinator, recruited participants by word of mouth for a study documenting the impact of the STSA play intervention on caregivers and their children ages 3 to 5 over whom they have primary responsibility. All caregivers and children provided oral or written consent to participate in this Institutional Review Board-approved study after the scope and procedures of the study were carefully explained (Goodman & Dent, 2019).

The project coordinator and his assistant, both trained by one of the researchers, conducted two STSA play-based intervention groups and two control intervention groups at both the Mpigi Community Library and the Kabubbu Community Library. The four Mpigi groups were conducted on Mondays and Wednesdays for a total of 42 sessions, while the four Kabubbu groups were conducted on Tuesdays and Thursdays for a total of 41 sessions. Some group sessions were canceled because of exam and holiday schedules, as well as unscheduled events that interfered with participation (e.g., all children from a village attending a child's funeral) (Goodman & Dent, 2019).

Children in both conditions were given one piece of paper and one crayon to draw anything they wished on both front and back. Only one crayon and one piece of paper were given to each

child during each session to standardize the procedure across conditions and children (Goodman & Dent, 2019).

Two experienced art therapy professors (one of whom developed this coding system) independently coded a random selection of 20 drawings from the larger collection of drawings from this study (N = 7,679 by the end of August 2014). The ICC for the FE assessment instrument (described later) was .93 (Goodman & Dent, 2019). After this successful coding practice, these same two coders independently coded a random selection of 61 drawings from the total number of drawings (n = 263) by the four children featured in this study (23.19% of this smaller sample of drawings). The ICC reliability for the FE scale was .83. One of these coders coded the remaining 202 drawings (Goodman & Dent, 2019).

Formal Elements: A Culturally Sensitive Rating Scale

The structural characteristics of Ugandan children's drawings were examined using the FE scale (see Table 2; Tuman, 1998, 1999a, 1999b; Goodman et al., 2022), which consists of 10 scales designed to capture the more formal aspects of art and design.

Table 2

Scale of Formal Elements

Scale 1	Number of colors
Scale 2	Implied Energy
Scale 3	Space
Scale 4	Composition
Scale 5	Line Quality
Scale 6	Overall Shape Quality
Scale 7	Integration
Scale 8	Details of Objects
Scale 9	Repetition of Schematic Elements
Scale 10	Developmental Level

Without negotiating the content or theme of a drawing, this coding system identifies the underlying structural organization and formal qualities of a two-dimensional drawing. Rating scales 1 through 8 are based on elements of art and design traditionally employed to describe the language of art in Western art criticism. The FE scale also relies on charting the child's developmental level, a schematic progression of artistic development aligned with Lowenfeld's (1947) stages of development. There are 32 different characteristics of design and projected application described by Lowenfeld (1947). Each item belongs to a scale (many scored as yes = 1 or no = 0) that is summed across all 10 scales (see Table 2) to produce a total score. In the original Ugandan study (Goodman et al., 2022), total scores ranged from 8 to 37. The drawings with the two highest FE total scores are displayed in Figures 1 and 2. For comparison, the drawings with the two lowest scaled scores are presented in Figures 3 and 4.

Figure 1

Example of a drawing with high Formal Elements total score (#88)

Figure 2

Example of a drawing with high Formal Elements total score (#155)

Figure 3

Example of a drawing with low Formal Elements total score (#51)

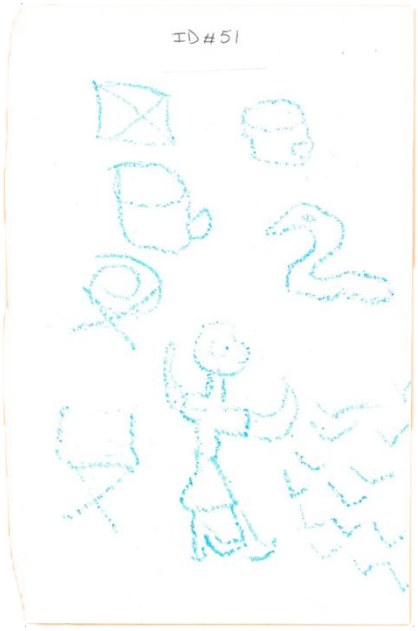
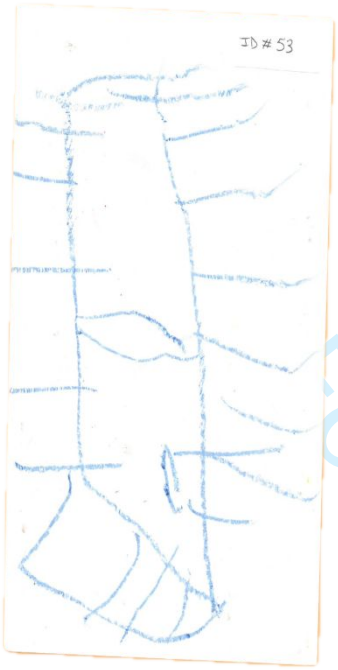


Figure 4

Example of a drawing with low Formal Elements total score (#53)



According to Lowenfeld (1947), the developmental level contains six stages from infancy to adolescence: scribble, pre-schematic, schematic, gang, pseudo-naturalistic, and adolescent art. Our study used the three anchor points most relevant to the population being studied: scribble (coded as 1), pre-schematic (coded as 2), and schematic (coded as 3). Visual perception begins during the scribbling stage when the child discovers that they can control and repeat motions (Lowenfeld, 1957, p. 102). In the pre-schematic and schematic stages, the child increasingly includes themselves and their feelings in drawings or transfers these feelings to someone else (Lowenfeld, 1957, pp. 51-53). We wanted to test whether the developmental level ratings coded by the ResNet50 machine learning algorithm could establish an ICC of 0.80 or higher with the developmental level ratings coded by the human coder. The general procedure of our experiment is portrayed in Figure 5.

Figure 5

Procedure for Data Collection, Labeling, Model Training and Testing

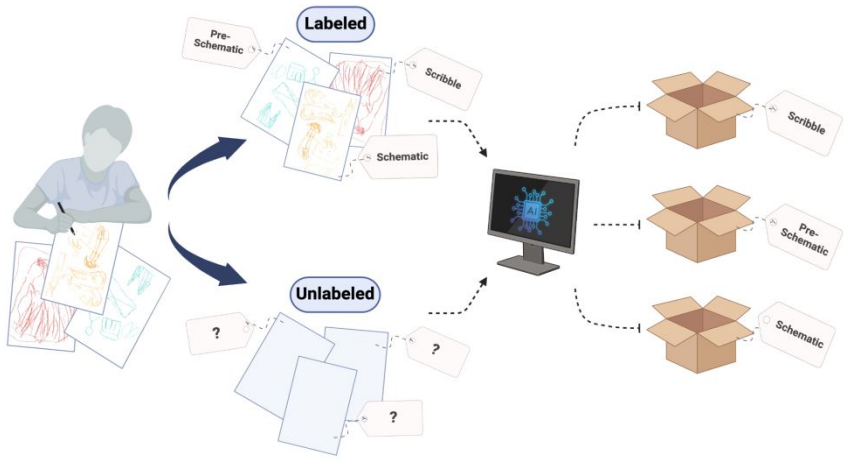
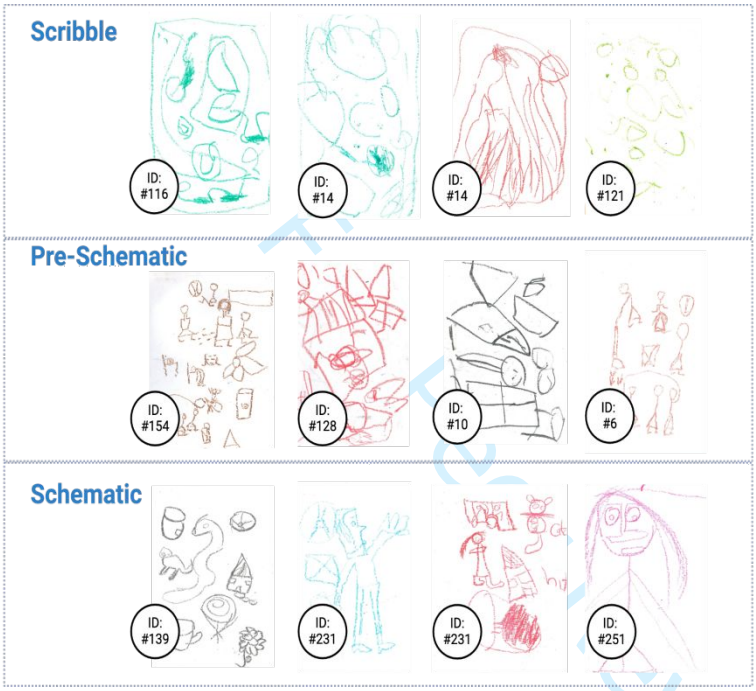


Figure 6 indicates some of the distinguishing features of each scribble, pre-schematic, and schematic category.

Figure 6

Sample of Scribble, Pre-Schematic, and Schematic Drawings



Drawings categorized as scribble are primarily characterized by a high density of lines composed of random, uncontrolled strokes that lack discernible patterns. These drawings exhibit minimal spatial organization, with elements often scattered without a clear focus or intentional placement. The absence of recognizable shapes further aids in classification, as the lines do not form identifiable or coherent objects. Additionally, there is a random use of space with disproportionate elements, signifying a need for balance and proportion. The elements within scribble drawings appear disjointed and lack cohesion, distinguishing them from more structured art forms. These features collectively indicate the earliest stage of artistic expression, where the emphasis is on drawing rather than creating recognizable forms or organized compositions.

In the pre-schematic stage, children's drawings exhibit a moderate level of line complexity with emerging patterns and the beginnings of structured shapes. These drawings show emerging spatial organization, featuring a basic layout and initial attempts at positioning objects

meaningfully. Though not fully detailed or accurate, the introduction of basic shapes that suggest the presence of objects serves as a critical indicator of developmental progression. Additionally, there is a beginning to use space more thoughtfully, even though proportions may still be inconsistent. This reflects a developing understanding of spatial relationships and scaling. Furthermore, the presence of some integration of elements—where parts of the drawing start to work together cohesively—signals the onset of storytelling or representation. These characteristics collectively aid in accurately classifying the drawing within the pre-schematic category, marking a transition from random scribbling to more intentional and organized artistic expression.

Drawings classified as schematic demonstrate a low density of lines characterized by well-defined, purposeful strokes that form recognizable shapes and figures. These drawings exhibit high spatial organization, with elements consistently placed and proportioned within the drawing space. There is a clear and accurate representation of shapes and forms, allowing for easy identification of depicted objects, which marks a significant departure from the randomness observed in earlier stages. The effective use of space with balanced proportions reflects the child's understanding of scale and perspective, showcasing an advanced grasp of spatial relationships. Additionally, there is a high level of integration, where various elements work together harmoniously to convey a complete scene or narrative. This integration signifies advanced cognitive and motor skills, enabling the differentiation of schematic drawings from less organized forms. These features collectively indicate a sophisticated level of artistic development, where the focus shifts from mere mark-making to creating meaningful and coherent compositions.

Data Set

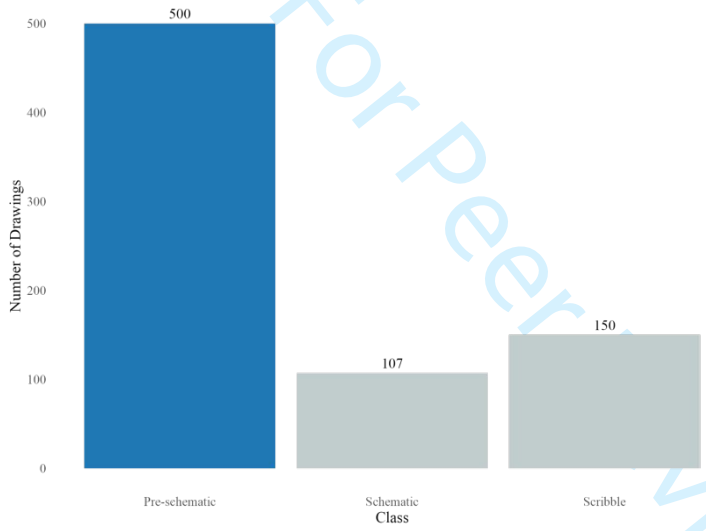
The data set used in this study comprises 757 drawings collected from Ugandan children as part of a longitudinal study investigating the STSA play intervention. Each drawing was manually labeled according to Lowenfeld's (1947) stages of artistic development: scribble, pre-schematic, and schematic. Among these, 500 drawings were classified as pre-schematic, 150 as scribble, and 107 as schematic. A human coder randomly selected and labeled each drawing from all the collected drawings.

Class Distribution and Imbalance

Upon analyzing the distribution of the drawings across these categories, we identified a significant imbalance in the representation of the different stages within the dataset as shown in Figure 7. This bar chart shows the number of drawings across three stages of artistic development: pre-schematic (500 drawings, 66.1%), scribble (150 drawings, 19.8%), and schematic (107 drawings, 14.1%). The chart highlights the significant distribution imbalance, with most drawings falling in the pre-schematic stage.

Figure 7

Class Distribution of Ugandan Preschool Children's Drawings



This imbalance indicates that the pre-schematic category dominates the data set, while the scribble and schematic categories are underrepresented. Such an imbalance can lead to a biased machine learning model that performs well on the majority class but poorly on the minority classes, referred to as overfitting. This is particularly problematic when the goal is to achieve high accuracy across all developmental stages.

Addressing the Imbalance

To mitigate this issue, additional drawings were manually coded to slightly increase the sample size of the scribble and schematic categories. However, the imbalance remained significant. Therefore, we employed the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) to artificially augment the minority classes. SMOTE generates synthetic

examples by interpolating between existing minority class samples, effectively balancing the training data without simply duplicating existing instances.

An imbalanced dataset can result in a poor machine-learning estimator because the model tends to become biased toward the majority class, often neglecting the minority classes. This bias occurs because the learning algorithm is overwhelmed by the majority class examples, leading it to misclassify or ignore minority class instances. Consequently, the model may achieve high overall accuracy by predominantly predicting the majority class but perform poorly in correctly identifying instances from the minority classes. This issue is especially problematic when the goal is to achieve high accuracy across all developmental stages.

By balancing the dataset using SMOTE, we aimed to improve the model's ability to accurately classify drawings from all developmental stages. This approach ensures that the classifier pays equal attention to all classes during training and can generalize well to unseen data from all categories, thereby mitigating the bias toward the pre-schematic category.

Image Processing

The collected drawings were manually cropped and resized to 224×224 pixels to ensure consistency during training. Since each drawing was created using a single crayon color, the images were converted to grayscale to simplify the data without losing essential features. Converting to grayscale reduces computational complexity and allows the model to focus on structural and textural elements crucial for classification. Additionally, data augmentation techniques such as rotation, zooming, and flipping were applied to enhance the model's ability to generalize and prevent overfitting (Shorten & Khoshgoftaar, 2019).

Model Architecture and Training

The model used in this study is a customized adaptation of the ResNet50 architecture, a deep convolutional neural network known for its residual learning capabilities (He et al., 2016). While ResNet50 is pre-trained on the ImageNet data set (Deng et al., 2009) and designed for 1,000-class classification tasks, we modified the architecture to suit our specific three-class classification problem.

Customizations

Several customizations were implemented to tailor the ResNet50 architecture for our specific classification task. The original fully connected (FC) layer, designed to output 1,000 classes, was replaced with a new FC layer that outputs three classes corresponding to the scribble, pre-schematic, and schematic stages of children's drawings. This modification refocused the model on our specific developmental categories. To prevent overfitting, a dropout layer with a probability of 0.5 was added after the new FC layer. Dropout regularization randomly disables neurons during training, encouraging the network to learn more robust features (Srivastava et al., 2014). Additionally, label smoothing was applied with a smoothing factor of 0.1. Instead of assigning a probability of 1.0 to the true class, label smoothing assigns 0.9 to the true class and distributes the remaining 0.1 among the other classes. This technique prevents the model from becoming overconfident and improves generalization (Szegedy et al., 2016). The final output of the network passes through a log-softmax activation function, converting the logits into log probabilities compatible with the label smoothing loss function.

Training Details

The model was trained using the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.0001 to update the model's parameters effectively. A StepLR scheduler was employed to decrease the learning rate by a factor of 0.1 every seven epochs, aiding in more efficient convergence. Training was conducted using a GPU when available to expedite the process. The training process ran for a maximum of 50 epochs, with early stopping based on validation performance to prevent overfitting. During training, input images were processed through the modified ResNet50 architecture, with dropout regularization applied and the loss computed using the label smoothing loss function. The optimizer and scheduler adjusted the learning rates and updated the model's weights to minimize the loss function.

Rationale for Model Adaptation

The rationale for adapting the model centers on three key aspects. Customization for the specific task is achieved by modifying the output layer, which tailors the model to our three-class classification problem and enhances its performance beyond that of a generic model. Overfitting

prevention is addressed by adding a dropout layer, which is crucial given the relatively small size of our dataset. Additionally, improved generalization is ensured through label smoothing, which helps the model generalize better to unseen data by preventing it from becoming overly confident in its predictions. Together, these adaptations optimize the model's performance and reliability for our specific classification task.

Data Augmentation and Regularization

Data augmentation techniques were applied during training to further enhance the model's ability to generalize. These included random rotations, translations, and flips of the input images. Such transformations expose the model to various scenarios, improving its robustness. The overall process and pipeline of how we trained the model is in Table 3.

Table 3

Customizations for the Base ResNet50 Model and Training Process

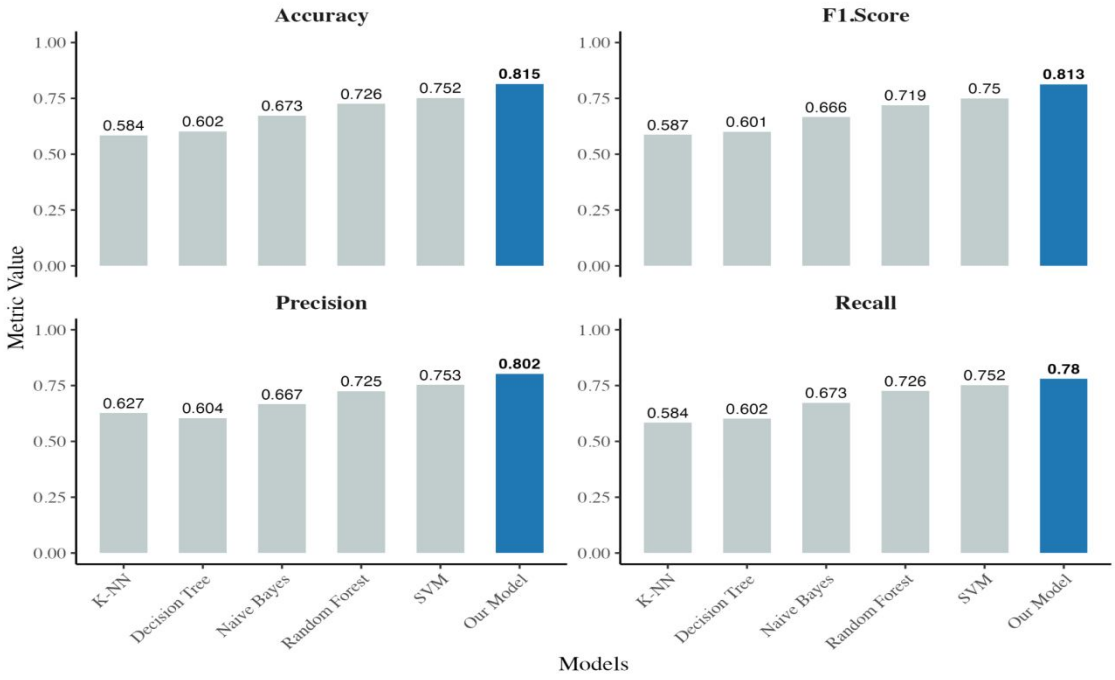
Initialize: Load dataset of 757 labeled drawings
Class Distribution and Imbalance:
  Analyze class distribution:
    Pre-schematic: 66.1%
    Scribble: 19.8%
    Schematic: 14.1%
Addressing the Imbalance:
for each minority class in {Scribble, Schematic} **do**
    Convert images to NumPy arrays
    Reshape images into one-dimensional vectors
    Apply SMOTE to generate synthetic samples
end for
Image Processing:
for each image in dataset do **do**
    Crop and resize to 224×224 pixels
    Convert to grayscale
    Apply data augmentation:
      • Random rotations
      • Zooming
      • Flipping
end for
Model Architecture:
Initialize ResNet50 with pre-trained ImageNet weights
Replace the original fully connected (FC) layer:
  Original FC: Outputs 1000 classes
  New FC: Outputs 3 classes ({Scribble, Pre-schematic, Schematic})
Add Dropout layer with probability $p = 0.5$
Apply Log-Softmax activation function
Define Loss Function:
Initialize Label Smoothing Loss with smoothing factor $\alpha = 0.1$
Initialize Optimizer and Scheduler:
Move model to GPU if available, else CPU
Define Adam optimizer with learning rate $\eta = 0.0001$
Define StepLR scheduler with step size $s = 7$ epochs and $\gamma = 0.1$
Training Loop:
for epoch = 1 to 50 do **do**
    for each batch in training data do **do**
      Move input images and labels to device
      $preds \leftarrow$ model($input$)
      $loss \leftarrow$ LabelSmoothingLoss(preds, targets)
      $loss$.backward()
      $optimizer$.step()
      $optimizer$.zero_grad()
    end for
    $scheduler$.step()
end for
Validate:
  Evaluate on validation dataset
  if validation performance does not improve **then**
    Break training loop (Early Stopping)
  end if
Model Evaluation:
Evaluate model on test dataset:
  Compute Precision, Recall, F1-score, Accuracy
Compute Intraclass Correlation Coefficient (ICC) with human coders
Error Analysis:
  Analyze misclassifications between Pre-schematic and Schematic stages
  Analyze misclassifications of detailed Scribbles as Pre-schematic

Several metrics were used to evaluate the model's performance, including precision, recall, F1-score, and accuracy. Additionally, an ICC was calculated to assess the agreement between the

AI model's classifications and the manual coding performed by human experts (Koo & Li, 2016; Shrout & Fleiss, 1979). The model achieved an ICC of 0.85 (95% CI: 0.82–0.88), indicating strong agreement with human coders. The validation accuracy was 81% (Figure 8), and the test accuracy was 78% (Figure 9). The bar graphs compare various machine learning models, including K-Nearest Neighbors (K-NN), Decision Tree, Naive Bayes, Random Forest, Support Vector Machine (SVM), and our custom ResNet50-based model, across metrics such as accuracy, precision, recall, and F1-score. Our custom ResNet50-based model consistently outperforms the other models across all metrics.

Figure 8

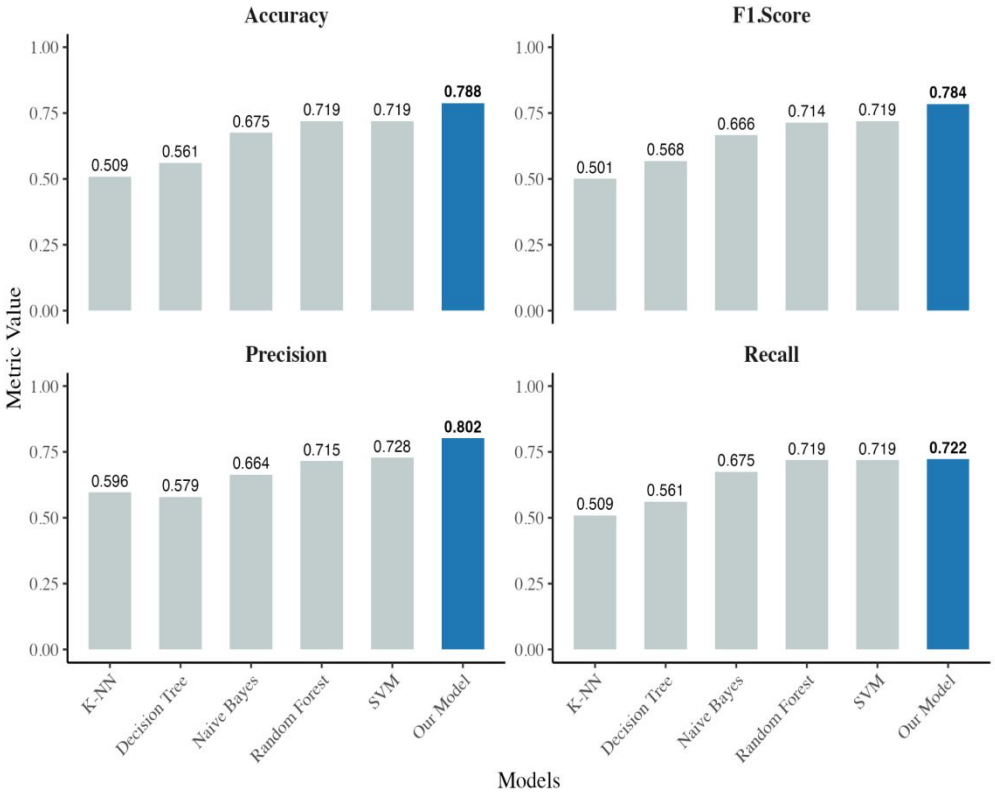
Performance Comparison of Machine Learning Models on Validation Data



Developing a Machine Learning Model to Categorize Children's Drawings

Figure 9

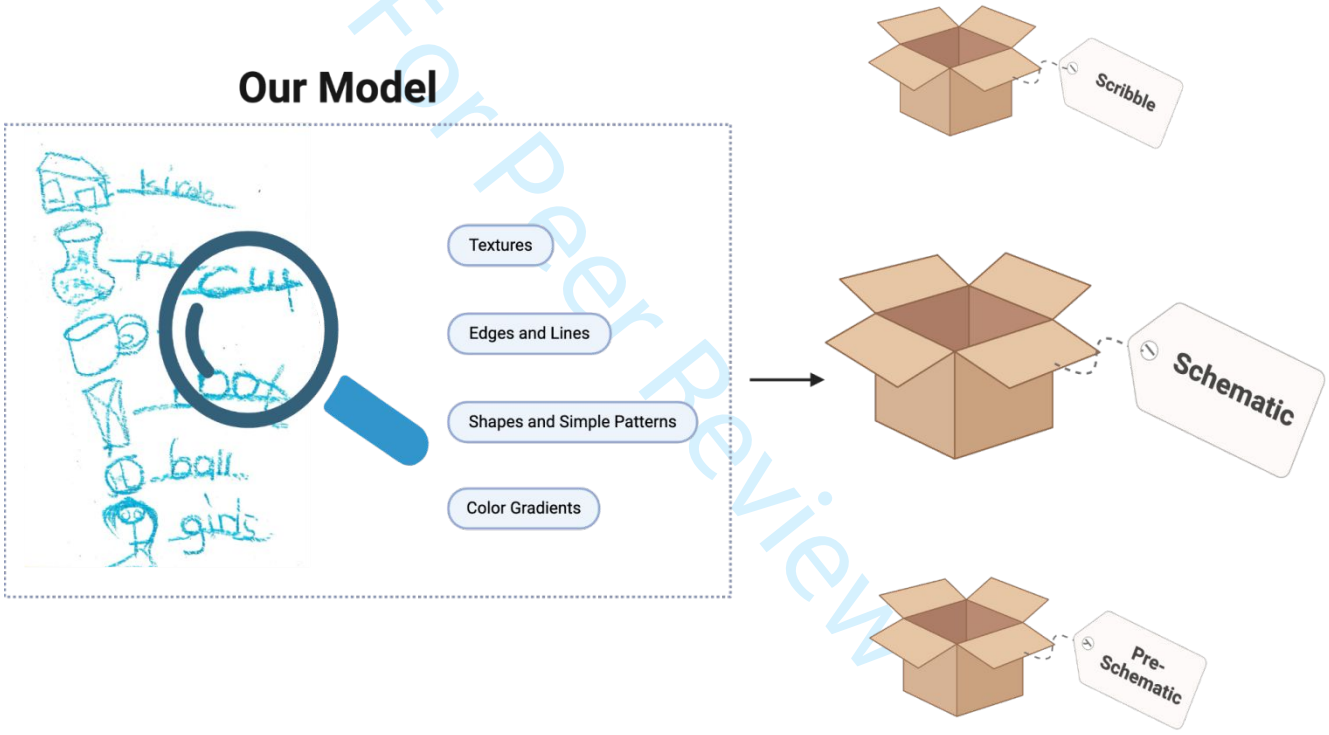
Performance Comparison of Machine Learning Models on Test Data



These results suggest that the customized AI model is capable of replicating human coding performance effectively, making it suitable for large-scale, high-throughput analysis. Our model achieved an intraclass correlation coefficient (ICC) of 0.85 with human coders, a validation accuracy of 81%, and a test accuracy of 78%. These metrics indicate strong agreement and reliable performance across all developmental stages. Figure 10 indicates some of the features our model considers when classifying the images into scribble, pre-schematic, and schematic.

Figure 10

Image Classification Features for Consideration



Error Analysis

An analysis of misclassifications revealed that most errors occurred between the pre-schematic and schematic stages, likely due to overlapping features in these developmental levels. The model occasionally misclassified detailed scribbles as pre-schematic drawings, suggesting a need for finer-grained classification criteria. This result is detailed in Figure 11.

Figure 11
Error Rate Across Different Predicted and Actual Labels in Validation and Test Set

Validation Confusion Matrix

Actual	Predicted: Scribble	Pre-Schematic	Schematic
Scribble	86%	14%	0.0%
Pre-Schematic	6.3%	86%	7.7%
Schematic	5.0%	31%	64%

*values rounded to 2 significant figures

Test Confusion Matrix

Actual	Predicted: Scribble	Pre-Schematic	Schematic
Scribble	70%	30%	0.0%
Pre-Schematic	6.3%	89%	4.7%
Schematic	0.0%	40%	60%

*values rounded to 2 significant figures

Discussion

The primary goal of this study was to determine whether the developmental level ratings assigned by the ResNet50 machine learning algorithm could achieve an ICC of 0.80 or higher when compared to ratings by human coders. The findings demonstrate that the algorithm meets and exceeds this threshold, achieving an ICC of 0.85 against the test set. This result underscores three key advantages of utilizing a machine-learning approach in analyzing children's drawings:

1. **Scalability and Efficiency**: Machine learning algorithms can process and evaluate vast amounts of data rapidly and consistently, far surpassing human capabilities. This is particularly significant given the extensive data set of approximately 140,000 drawings, which would be impractical to analyze manually.
2. **Enhanced Reproducibility and Objectivity**: Traditional methods in psychology and comparative cognition often rely on researchers' subjective interpretations, which can introduce bias and limit reproducibility (Belzung et al., 2023). By contrast, machine learning models apply consistent criteria across all data points, reducing subjectivity. As Lee et al. (2024) note, minimizing inherent subjectivity is crucial for obtaining reliable results.
3. **Longevity and Accessibility:** Once the model has been trained to consistently achieve accurate and reliable results on any specific scale, the coding process related to this scale becomes significantly easier. Access to this model could be provisioned to researchers globally. While an internet connection to process these images may prevent specific locations or devices from soundly running the model, this method would likely be more accessible than relying on the small set of researchers properly trained to code. Additionally, a valid model would allow all future researchers to establish reliability across other research populations, regardless of the location and capabilities of the previously cited researchers.

The selection and prioritization of scales was a major challenge of this research. Unlike other projective drawing tests where children are given a prompt, the children in the study were able to draw whatever they wanted. Each drawing contained a variety of objects, scenes, themes, and styles, as well as a range of formal choices, like the use of space, orientation of the drawings,

and cleanliness of the features. The researchers discussed and explored two other FE scales before settling on scale 10 since the developmental levels are perhaps the most fundamental and universal of all the formal scale elements. While there were differences between the Ugandan sample and the results reported in American samples (Goodman et al., 2022), overall, the children in the Ugandan sample seemed to demonstrate the same developmental drawing progress as described by Lowenfeld (1947) and Cox (1993). Training a machine learning model to accurately categorize more nuanced elements and features such as "Implied Energy" (scale 2) and to decipher whether line quality is "Sketchy" or "Expressive" (scale 3) were also discussed but not pursued.

ResNet50 provides a robust and scalable solution, thus contributing to more objective and reproducible research outcomes in the field. The successful application of the ResNet50 model in classifying children's drawings underscores the potential of AI in automating visual data analysis within information science. This approach facilitates the efficient processing of large-scale visual data, enabling researchers and professionals to extract meaningful insights without the constraints of manual coding.

Limitations

The study's limitations include the relatively small data set size and its cultural specificity to Ugandan children, which may affect the generalizability of the model. While there are a large number of drawings, the team determined there was too large of a variance across drawings to attempt an unsupervised learning approach, where the model would assign scales to drawings that no human had coded yet. Training the model to consistently and accurately code drawings will require a much greater human coding effort for scale 10 and additional scales of interest. Ultimately, while artificial intelligence seems to have a clear potential to save time and expertise in the coding process, human evaluation is still integral for any savings to be realized. For certain scales, it is fair to assume that the effort needed to expand the ResNet50 model's capacity to evaluate a new scale accurately may be roughly equivalent to the effort that could be spent solely on continued coding by humans.

Additionally, the reliance on manually labeled data introduces potential biases that could influence the model's performance. While the coders achieved a satisfactory interrater reliability

metric before coding the actual data set began, there was no step to confirm that the coder's interpretation was an accurate depiction of what the child meant to draw. At the time of writing this article, even though the STSA intervention is still active and the children in Kabubbu are still generating drawings, the research team is no longer regularly evaluating the children, which eliminates the possibility of another means of validating both human- and AI-coded interpretations. In the case of scale 10, the team does not have access to other health or educational assessments that could ratify the longitudinal growth of the four sample participants' emotional development.

Lastly, while the research team had deep expertise in their respective areas they had little to no cross-disciplinary expertise. This gap may have impacted the ability of the research team to have truly interdisciplinary discussions about the project. Ideally, the computer scientists who designed the machine learning model would have first been trained on how to identify the developmental stages of scale 10 manually, and the researchers who led the Uganda study would have been trained on the basics of machine learning. Time and resources made this untenable, and in the end, the researchers were able to address these knowledge gaps by having extended and thorough planning meetings, studying the literature on the respective disciplinary areas (machine learning and children's artistic development), and leveraging additional expertise to code additional drawings for the study manually.

Future Work & Conclusion

The uniqueness of this study is predicated upon the fact that the scale used to train the model was formulated by subject experts specifically for the population being studied.

In terms of future directions, the researchers discussed the expansion of the study to include one of the content scales, which was also developed for the original study (Goodman et al., 2022). The content scale "consists of 10 scales designed to capture the culture, socially informed behaviors, and attitudes that could influence the artistic production of children living in a rural Ugandan community" (Goodman et al., 2022, p. 4). The content scale features categories such as "Daily Experience" and "Domestic Life" that were difficult for even human coders to classify, rendering them inherently more challenging for an ML model to recognize. On the other hand,

content scales such as "Number of Objects/Figures" and "Letters, Numbers, Words, Sentences" would be simpler to categorize and entail a more straightforward machine learning training process.

Future research will also focus on expanding the model to classify additional formal element scales and apply the methodology to drawings from different cultural contexts. There are several well-known large data sets of children's drawings from around the world, including the Terezin Ghetto drawings (The Jewish Museum, 2024), a collection of 4387 drawings by Jewish children during WWII. This multicultural data set, which might require a unique set of culturally informed content and formal element scales, is an example of the kind of data set that might be coded by a machine learning algorithm, thus creating new vistas of understanding into the experiences of children living in Terezin during WWII.

Yuan, Huang, Ma, and Yan (2020) suggest that one of the main challenges related to evaluating children's drawings using AI technology is that a single drawing can have multiple embedded meanings and reflect multiple emotional states of the child. That is a challenge that, according to the authors, makes it difficult for AI to work effectively (Yuan et al., 2020). Jensen et al. (2023) found the most reliable approach to identifying drawing characteristics was to pair human coding and machine learning techniques. This project faced a number of challenges but ultimately succeeded in demonstrating the efficacy of using a machine learning model based on the ResNet50 architecture to classify Ugandan preschool children's drawings into developmental stages. By achieving strong agreement with human coders, the model provides a scalable solution for analyzing large visual data sets. The automation of this classification process holds significant promise for advancing research in cognitive and artistic development, particularly in cross-cultural contexts. It should be noted that this field of work, although still quite compact, continues to uncover novel insights into how to best capture meaning from children's drawings using AI (Long, Fan, Huey, Chai, & Frank, 2024; Yim, Lee, Kim, & Yu, 2021; Ali, Abd-Alrazaq, Shah, Alajlani, Alam, & Househ, 2022). Continued research can hopefully establish models that are valid and reliable in coding and reliable across cultural contexts. Once such a model exists, it can be leveraged to support the study of how children experience the world, no matter their background or culture. Collaborations with trained coders, educators, and psychologists are planned to enhance the model's applicability and explore its integration into educational and rural library settings.

Conflict of Interest Statement

The authors declare that they have no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data and code used in this study are available upon request and are in line with JASIST's policies on open science and replication.

Acknowledgments

We extend our heartfelt gratitude to the children, caregivers, and local libraries in Mpigi and Kabubbu for their invaluable participation and contributions to this study. Our sincere appreciation goes to Alisha Morejon for her indispensable assistance with the research background of this computational study. We are also deeply grateful to Kaisi Xing for her creation and aesthetic evaluation of the figures and to Holly Rosen, who provided coding for additional drawings to support the training of the model. Additionally, we thank the Center for AI Learning for introducing this research team, organizing the project, and providing support and essential resources, all of which were instrumental in the successful completion of this project. Figures 5, 6, 10, and 11 are created with biorender.com.

References

Ahman, E. (2018). Cultural influences on children's artistic development. *Journal of Cross-Cultural Psychology*, 49(5), 731–748.

Ali, W., Abd-Alrazaq, A., Shah, U., Alajlani, M., Alam, T., & Househ, M. (2022). The use of artificial intelligence in screening and diagnosing autism spectrum disorder: A scoping review. *Journal of Medical Internet Research*, 24(1), e30087.

Alter-Muri, S., & Vazzano, S. (2014). The effects of culture on children's drawing styles. *Art Therapy: Journal of the American Art Therapy Association*, 31(1), 12–18.

Aronsson, K., & Andersson, S. (1996). Social scaling in children's drawings of classroom life: A cross-cultural study. *British Journal of Developmental Psychology*, 14(3), 301–314.

Bat Or, M., Kourkoutas, E., Smyrnaki, M., & Potchebutzky, D. (2019). Children's drawings and narratives: Using arts to access their emotional world. *Frontiers in Psychology*, 10, 1114.

Barton, G. (1997). Drawing on the tablet: A new look at drawing development in the digital age. In M. Cox (Ed.), *Visual order: The nature and development of pictorial representation* (pp. 301–317). Cambridge University Press.

Belzung, C., Pele, M., Renoult, J. P., & Sueur, C. (2023). Subjective biases in behavioral sciences: How artificial intelligence can help. *Behavioral Sciences*, 13(2), 123.

Bracken, B. A. (2007). *Bracken School Readiness Assessment (3rd ed.): Examiner's Manual*. Pearson Publishing Company.

Brittain, W. L., & Lowenfeld, V. (1987). *Creative and Mental Growth* (8th ed.). Prentice Hall.

Burton, J. M. (1995). Emergence of representation in children's drawings. *Visual Arts Research*, 21(1), 90–105.

Burton, J. M. (1997). The unintended consequences of intelligent computing in art education. *Studies in Art Education*, 38(3), 170–186.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Court, E. (1989). Art from childhood: African children's art development. *African Arts*, 22(4), 62–85.

Cox, M. V. (1993). *Children's Drawings of the Human Figure*. Lawrence Erlbaum Associates.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE.

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Transactions on Graphics*, 31(4), 44.

Farokhi, M., & Hashemi, M. (2011). The analysis of children's drawings: Social, emotional, physical, and psychological aspects. *Procedia - Social and Behavioral Sciences*, 30, 2219–2224.

Feldman, D. H. (1987). Stages in artistic development. In A. H. Housen (Ed.), *Stages of Artistic Development* (pp. 93–111). J. Paul Getty Trust.

Gan, Y., Meng, W., & Xie, Q. (2016). Children's drawing development as an indicator of school readiness. *Early Child Development and Care*, 186(9), 1488–1500.

Gardner, H. (1980). *Artful Scribbles: The Significance of Children's Drawings*. Basic Books.

Gantt, L., & Anderson, F. E. (2009). The Formal Elements Art Therapy Scale: A measurement system for global variables in art. *Art Therapy: Journal of the American Art Therapy Association*, 26(3), 124–129.

Glewwe, P., Ross, P., & Wydick, B. (2018). Developing hope among impoverished children: Using child self-portraits to measure poverty program impacts. *Journal of Human Resources*, 53(2), 330–355.

Goodman, G., & Dent, V. F. (2019). Impact of a play-based intervention on school readiness skills of rural Ugandan preschool children. *International Journal of Play Therapy*, 28(3), 151–161.

Goodman, G., Dent, V., Tuman, D., & Lee, S. Y. (2022). Drawings from a play-based intervention: Windows to the soul of rural Ugandan preschool children's artistic development. *The Arts in Psychotherapy*, 77, 101876.

Haidkind, P., Henno, I., Kikas, E., & Peets, K. (2011). A longitudinal study of the development of drawing skills in Estonian preschoolers. *Early Child Development and Care*, 181(3), 371–385.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Jensen, H., Sumanthiran, M., Kirkorian, H., Travers, B., Rosengren, K., & Rogers, T. (2023). Integrating human and machine intelligence in children's drawing analysis. *Developmental Psychology*, 59(4), 1–15.

Jewish Museum (2024). Children's drawings from the Terezin ghetto. Retrieved from https://www.jewishmuseum.cz/en/collection-research/collections-funds/visual-arts/children-s-drawings-from-the-terezin-ghetto/.

Jiggetts, J. (2021). Exploring emotional expression through children's drawings in therapeutic settings. *Child Art Therapy Journal*, 38(2), 85–97.

Kalvesten, E., & Ödman, P. (1979). Children's drawings in different cultures: A comparative study between Swedish and Tanzanian children. *Child Development*, 50(4), 1129–1136.

Kallitsoglou, A., Repana, V., & Shiakou, M. (2022). Machine learning approaches to analyzing children's drawings: A systematic review. *Computers in Human Behavior*, 128, 107127.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Lawrie, J. (2022). Machine learning analysis of children's drawings: Indicators of psychological distress in rural India. *International Journal of Psychology*, 57(2), 215–228.

Lee, S., Kim, H., & Park, J. (2024). Reducing subjectivity in psychological assessments using machine learning. *Journal of Psychology and Technology*, 15(1), 10–25.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

Li, Y. (2013). Sketch recognition for large classes using geometry and context. *ACM Transactions on Graphics*, 32(6), 1–10.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., & van der Laak, J. A. W. M. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.

Long, B., Fan, Q., Huey, D., Chai, Y., & Frank, M. C. (2024). Using AI to understand children's visual concepts: A study of drawing classification. *Cognitive Science*, 48(1), e13012.

Lowenfeld, V. (1947). *Creative and Mental Growth*. Macmillan.

Lowenfeld, V. (1957). *Creative and Mental Growth* (3rd ed.). Macmillan.

Martinent, G. (2021). Biases in human coding of children's drawings: Implications for machine learning applications. *Journal of Child Psychology*, 42(3), 455–468.

Merriman, M., & Guerin, P. (2006). Using children's drawings as evaluative tools: Guidelines for maximizing benefits. *Journal of Social Service Research*, 33(2), 13–27.

Monica, S., Davu, S., Caroline, J., & Jagganath, V. (2019). Classification of children's drawings using k-nearest neighbor algorithm. *International Journal of Computer Applications*, 182(29), 21–26.

Paley, V. G. (1990). *The Boy Who Would Be a Helicopter*. Harvard University Press.

Ravindran, R. (2022). Deep learning techniques for classification of children's drawings. *International Journal of Machine Learning and Computing*, 12(5), 350–356.

Refaat, M., & Atiya, A. F. (2009). Machine learning techniques for children's drawing classification. *Pattern Recognition Letters*, 30(7), 1107–1115.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958.

Stauffer, S. (2019). Children's drawings as measures of emotional status: A meta-analysis. *Art Therapy*, 36(4), 191–200.

Stiles, J., & Gibbons, J. (2000). Drawing development in children: Cross-cultural comparisons. *Child Development*, 71(4), 855–866.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826).

Thomas, G. V., & Silk, A. M. J. (1990). *An Introduction to the Psychology of Children's Drawings*. New York University Press.

Thomas, R., Powell, L., Polsley, S., Ray, N., & Hammond, T. (2022). Distinguishing child and adult drawing strokes using machine learning. *Journal of Experimental Child Psychology*, 217, 105319.

The Jewish Museum. (2021). Terezin children's drawings collection. Retrieved from Jewish Museum Website.

Tuman, D. M. (1998). Gender difference in form and content: The relationship between preferred subject and the formal characteristics of children's drawing. Unpublished doctoral dissertation, Teachers College Columbia University, New York.

Tuman, D. M. (1999a). Gender style as form and content: An examination of gender stereotypes in the subject preference of children's drawing. *Studies in Art Education*, 41(1), 40–60.

Tuman, D. M. (1999b). Sing a song of sixpence: An examination of sex difference in the subject preference of children's drawings. *Visual Arts Research*, 25(1), 51–62.

Wang, L., Kandemir, M., & Li, Y. (2020). Machine learning for analyzing children's drawings: A review. *ACM Computing Surveys*, 53(6), 1–35.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612.

Yim, J., Lee, D., Kim, S., & Yu, H. (2021). Understanding children's drawing development using deep learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17), 15290–15297.

Yu, Q., Yang, Y., Liu, F., Song, Y.-Z., Xiang, T., & Hospedales, T. M. (2017). Sketch-a-net: A deep neural network that beats humans. *International Journal of Computer Vision*, 122(3), 411–425.

Yuan, L., Huang, X., Ma, Z., & Yan, J. (2020). Challenges in AI evaluation of children's drawings. *Frontiers in Psychology*, 11, 215.

Developing a Machine Learning Model to Categorize Children's Drawings

Example of a drawing with high Formal Elements total score (#88)

54x84mm (144 x 144 DPI)

Example of a drawing with high Formal Elements total score (#155)

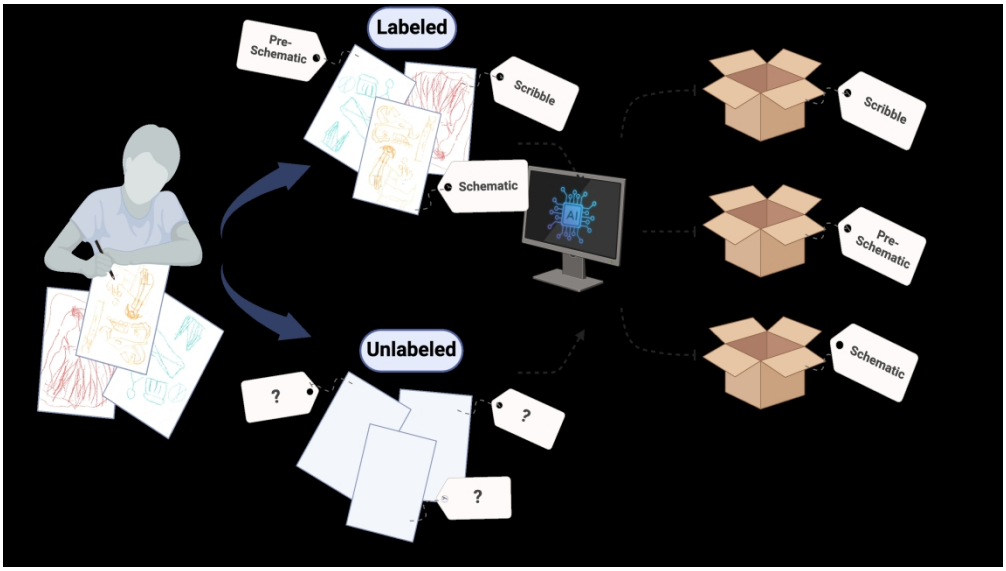
59x89mm (144 x 144 DPI)

Example of a drawing with low Formal Elements total score (#51)

69x96mm (144 x 144 DPI)

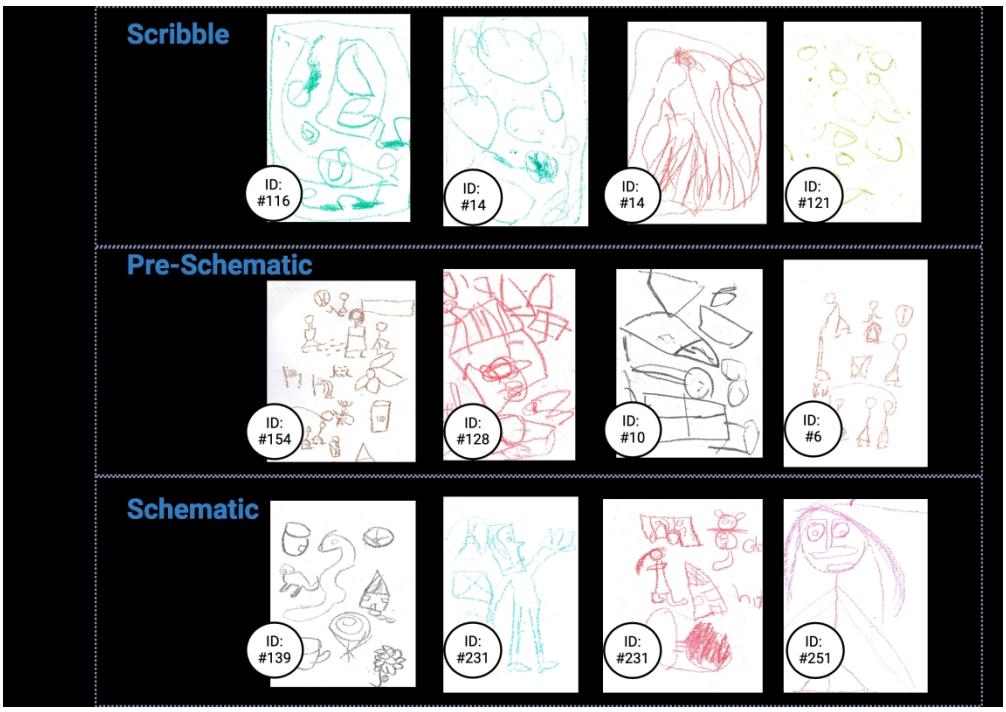
Example of a drawing with low Formal Elements total score (#53)

54x95mm (144 x 144 DPI)



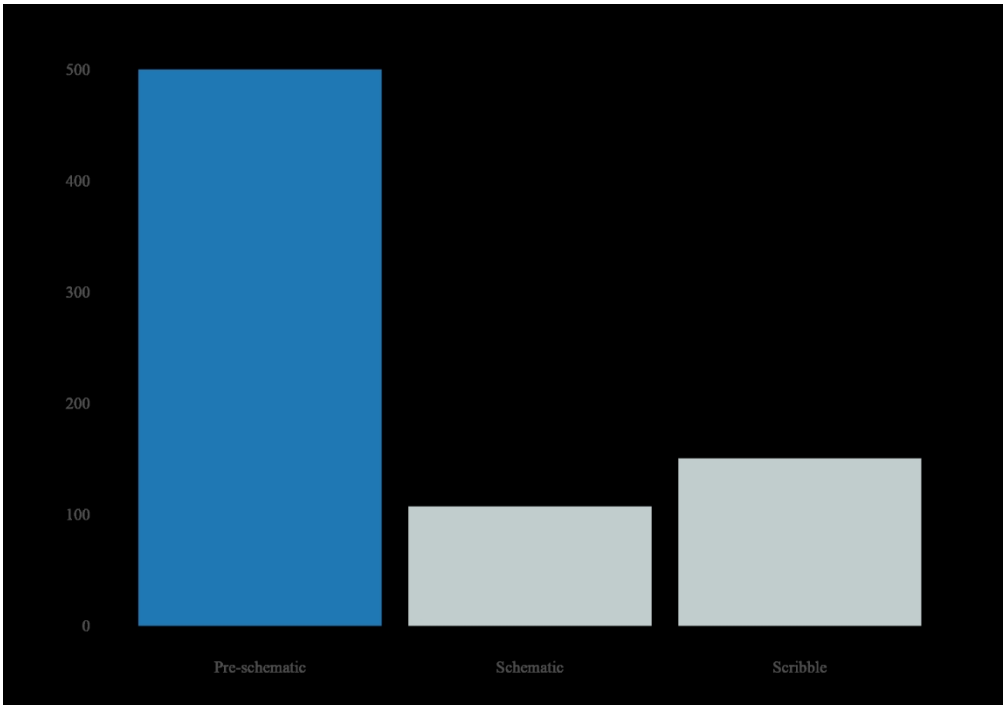
Procedure for Data Collection, Labeling, Model Training and Testing

645x363mm (118 x 118 DPI)



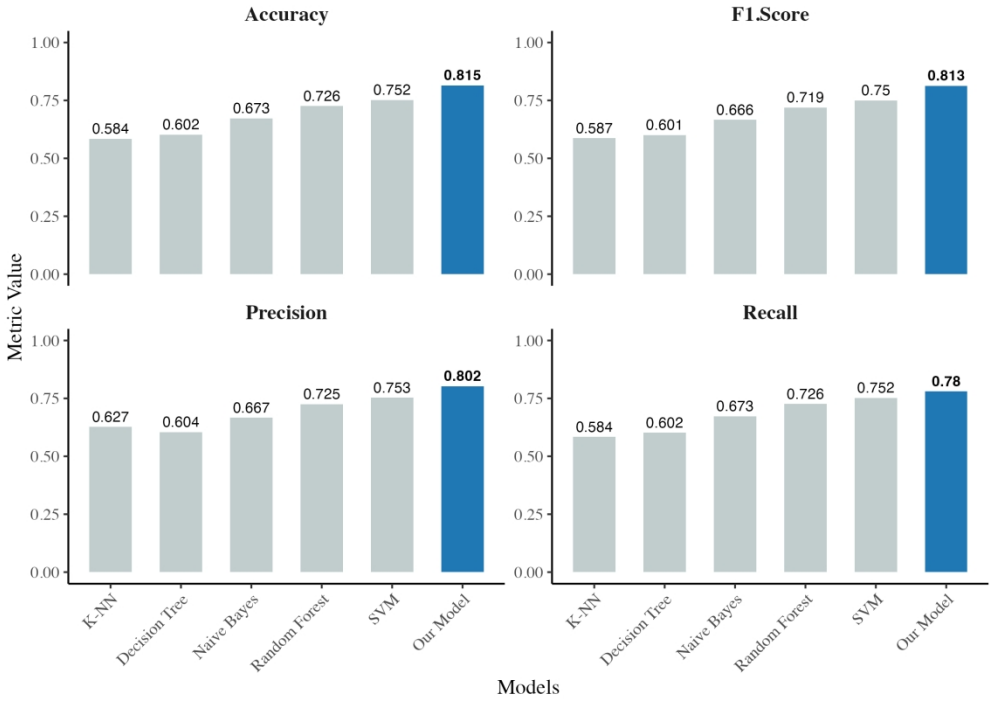
Sample of Scribble, Pre-Schematic, and Schematic Drawings

645x452mm (118 x 118 DPI)



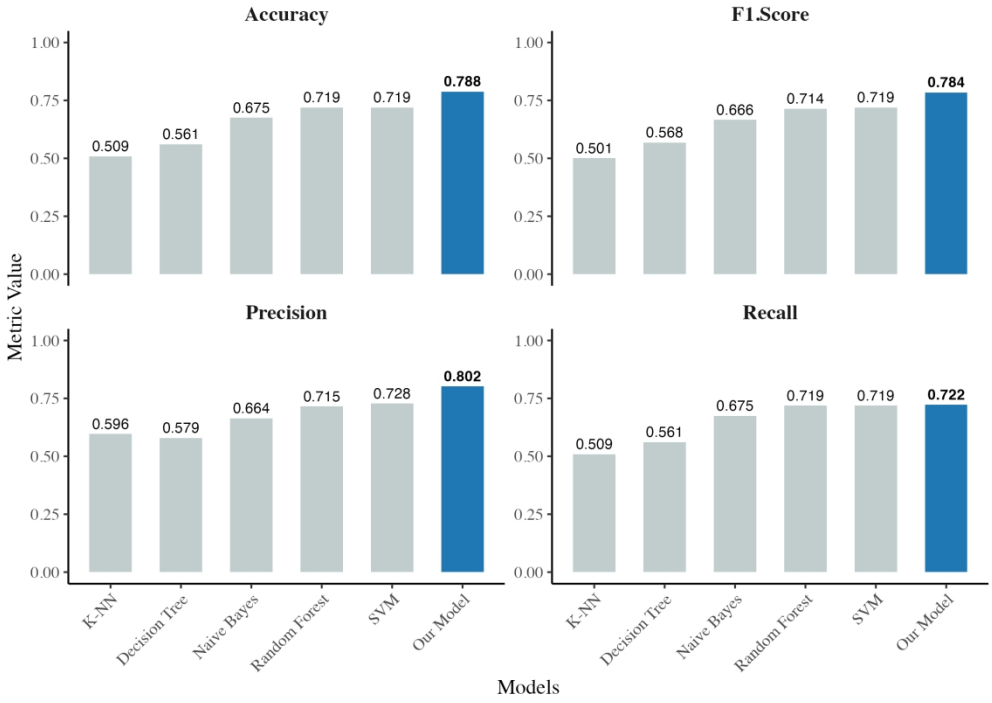
Class Distribution of Ugandan Preschool Children's Drawings

254x177mm (300 x 300 DPI)



Performance Comparison of Machine Learning Models on Validation Data

203x152mm (300 x 300 DPI)



Performance Comparison of Machine Learning Models on Test Data

203x152mm (300 x 300 DPI)

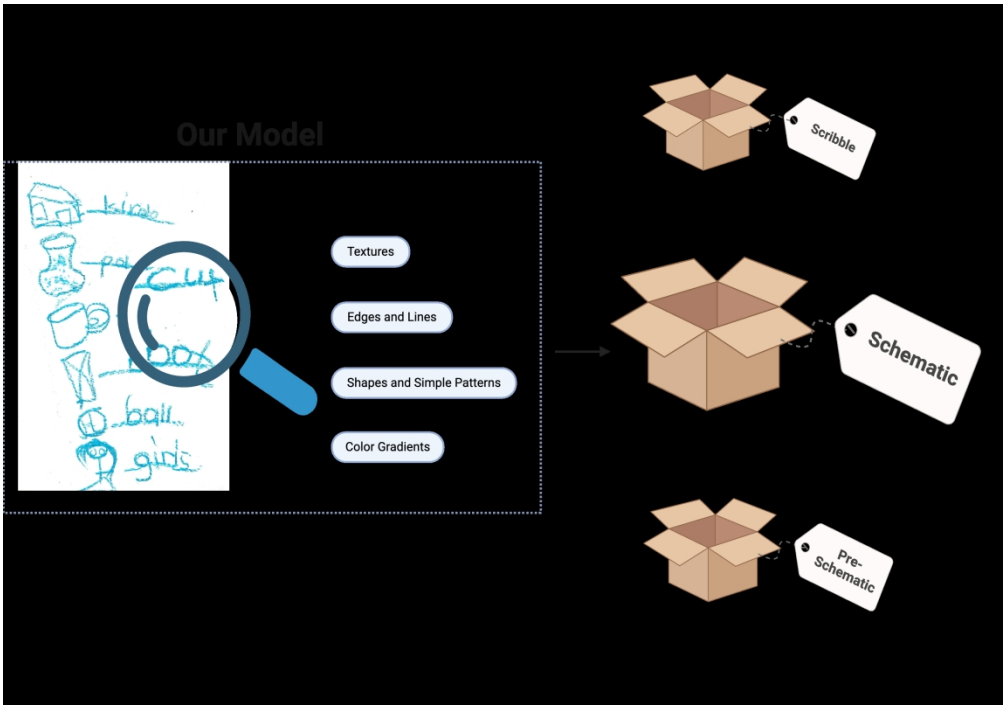
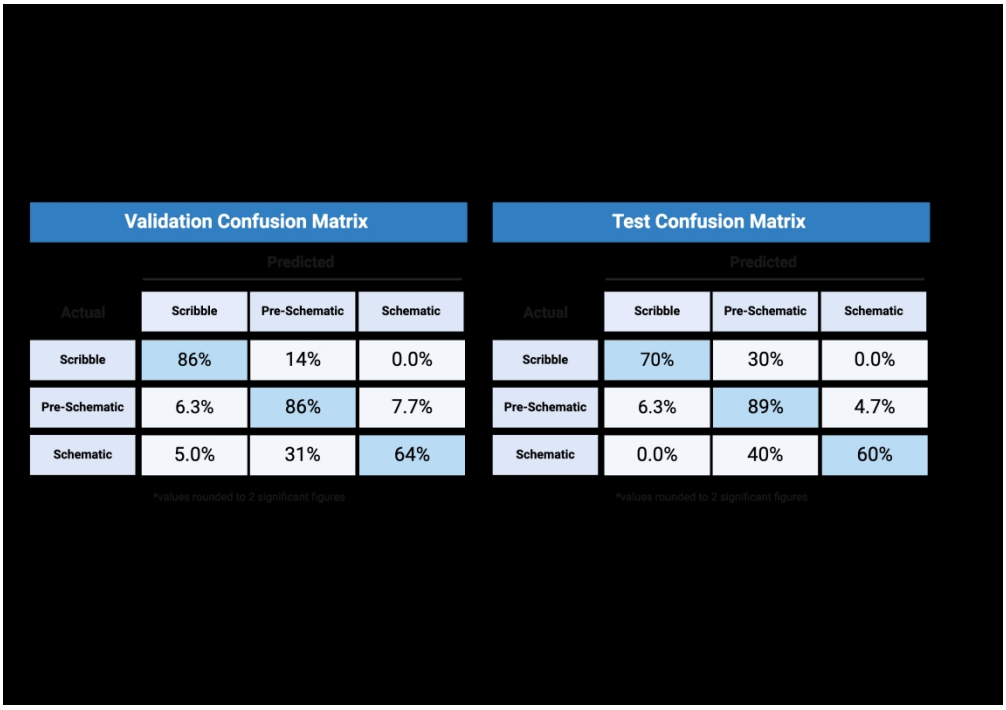


Image Classification Features for Consideration

645x452mm (118 x 118 DPI)



Validation Confusion Matrix

Actual \ Predicted	Scribble	Pre-Schematic	Schematic
Scribble	86%	14%	0.0%
Pre-Schematic	6.3%	86%	7.7%
Schematic	5.0%	31%	64%

*values rounded to 2 significant figures

Test Confusion Matrix

Actual \ Predicted	Scribble	Pre-Schematic	Schematic
Scribble	70%	30%	0.0%
Pre-Schematic	6.3%	89%	4.7%
Schematic	0.0%	40%	60%

*values rounded to 2 significant figures

Error Rate Across Different Predicted and Actual Labels in Validation and Test Set

645x452mm (118 x 118 DPI)

Table 1

Participants

	Age	Gender	Group
Child 1	64 months	Female	Control: Kabubbu/Zebras
Child 2	50 months	Male	STSA: Kabubbu/Lions
Child 3	48 months	Female	STSA: Mpigi/Pigeons
Child 4	65 months	Male	Control: Mpigi/Eagles

303x166mm (144 x 144 DPI)

Table 2

Scale of Formal Elements

Scale 1	Number of colors
Scale 2	Implied Energy
Scale 3	Space
Scale 4	Composition
Scale 5	Line Quality
Scale 6	Overall Shape Quality
Scale 7	Integration
Scale 8	Details of Objects
Scale 9	Repetition of Schematic Elements
Scale 10	Developmental Level

172x180mm (144 x 144 DPI)

Initialize: Load dataset of 757 labeled drawings
Class Distribution and Imbalance:
  Analyze class distribution:
    Pre-schematic: 66.1%
    Scribble: 19.8%
    Schematic: 14.1%
Addressing the Imbalance:
for each minority class in {Scribble, Schematic} **do**
  Convert images to NumPy arrays
  Reshape images into one-dimensional vectors
  Apply SMOTE to generate synthetic samples
end for
Image Processing:
for each image in dataset do **do**
  Crop and resize to 224×224 pixels
  Convert to grayscale
  Apply data augmentation:
    • Random rotations
    • Zooming
    • Flipping
end for
Model Architecture:
Initialize ResNet50 with pre-trained ImageNet weights
Replace the original fully connected (FC) layer:
  Original FC: Outputs 1000 classes
  New FC: Outputs 3 classes ({Scribble, Pre-schematic, Schematic})
Add Dropout layer with probability $p = 0.5$
Apply Log-Softmax activation function
Define Loss Function:
Initialize Label Smoothing Loss with smoothing factor $\alpha = 0.1$
Initialize Optimizer and Scheduler:
Move model to GPU if available, else CPU
Define Adam optimizer with learning rate $\eta = 0.0001$
Define StepLR scheduler with step size $s = 7$ epochs and $\gamma = 0.1$
Training Loop:
for epoch = 1 to 50 do **do**
  for each batch in training data do **do**
    Move input images and labels to device
    preds ← model(*input*)
    loss ← LabelSmoothingLoss(preds, targets)
    loss.backward()
    optimizer.step()
    optimizer.zero_grad()
  end for
  scheduler.step()
end for
Validate:
  Evaluate on validation dataset
  if validation performance does not improve **then**
    Break training loop (Early Stopping)
  end if
Model Evaluation:
Evaluate model on test dataset:
  Compute Precision, Recall, F1-score, Accuracy
Compute Intraclass Correlation Coefficient (ICC) with human coders
Error Analysis:
  Analyze misclassifications between Pre-schematic and Schematic stages
  Analyze misclassifications of detailed Scribbles as Pre-schematic

128x255mm (144 x 144 DPI)